ORTHOGONAL LEARNING FOR CONDITIONAL GEODESIC TREATMENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating conditional average treatment effects (CATEs) is a fundamental task in causal inference, and orthogonal learners reduce sensitivity to nuisance estimation. In many scientific applications, however, outcomes are naturally represented on Riemannian manifolds, where Euclidean differences do not define geometry-respecting treatment effects. Existing work develops Fréchet regression for manifold-valued outcomes and geodesic causal effects, but leaves open how to perform Neyman-orthogonal pointwise estimation and inference for covariate-dependent geodesic effects from observational data. To address this, we define the conditional geodesic treatment effect through treatment-specific conditional Fréchet means and develop the Geodesic Orthogonal learner (GO-learner) for its estimation. We derive a joint error expansion in which errors in estimating the two means are propagated through the logarithm map, propensity and conditional regression errors enter through products, and the resulting bounds yield convergence guarantees and asymptotically valid pointwise confidence sets. Experimental results demonstrate that (i) GO-learner has lower average MSE among the three evaluated estimators with sample size increasing; (ii) the predicted bias reduces when either nuisance class is correctly specified; and (iii) inference-specific bandwidths attain near-nominal pointwise coverage at the largest sample sizes.

1 INTRODUCTION

Heterogeneous treatment effects are commonly characterized by comparing conditional mean potential outcomes under treatment and control. For scalar or vector outcomes, this gives the conditional average treatment effect (CATE), for which a large literature has developed flexible estimators from observational data (Athey & Imbens, 2016; Wager & Athey, 2018; Künzel et al., 2019; Nie & Wager, 2021; Kennedy, 2023). In many scientific applications, however, outcomes have structures that are not well described by ordinary scalar or vector representations. Examples include symmetric positive-definite covariance and diffusion-tensor matrices, directional observations on spheres, and landmark shape data (Pennec et al., 2006; Dryden et al., 2009; Bhattacharya & Patrangenaru, 2003). These outcomes are naturally modeled using Riemannian geometry, where the Euclidean notions of an average and a difference need not respect the geometry of the outcome space.

Treatment effect analysis for manifold-valued outcomes therefore requires definitions that respect the geometry of the outcome space. The Fréchet mean extends the ordinary mean to a metric space, while on a Riemannian manifold the logarithm map represents the displacement from one point to another in a tangent space (Fréchet, 1948; Pennec et al., 2006; Petersen & Müller, 2019). Existing work studies distribution-valued potential outcomes, population-level geodesic treatment effects, absolute treatment effects in general metric spaces, random object outcomes under continuous treatments, cutoff-specific geodesic effects, conditional kernel treatment effects, and topological treatment effects (Lin et al., 2023; Kurisu et al., 2024; Shin et al., 2024; Bhattacharjee et al., 2025; Kurisu et al., 2025; Näf et al., 2026; Kim & Lee, 2026). Existing methods, however, do not address Neyman-orthogonal pointwise estimation and inference for treatment effects that vary with pretreatment covariates when outcomes lie on a Riemannian manifold. We formalize this target through the tangent representation of the geodesic connecting the two treatment-specific conditional Fréchet means, which we call the *conditional geodesic treatment effect*:

$$\tau(x) = \mathrm{Log}_{\mu_0(x)} \mu_1(x) \in T_{\mu_0(x)}\mathcal{M}, \qquad \mu_a(x) = \operatorname*{arg\,min}_{m \in \mathcal{M}} \mathbb{E}\left[d^2(Y(a), m) \mid X = x\right]. \quad (1)$$

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.

Here $\mu_a(x)$ is the treatment-specific conditional Fréchet mean and $\tau(x)$ is the tangent displacement from the control mean to the treated mean. When $\mathcal{M}$ is Euclidean, the conditional Fréchet means reduce to conditional expectations and the logarithm map reduces to subtraction, so $\tau(x)$ is exactly the ordinary CATE.

In Euclidean causal inference, Neyman-orthogonal moment functions reduce first-order sensitivity to nuisance estimation. A moment function is Neyman orthogonal when the first derivative of its population moment with respect to nuisance perturbations vanishes at the true nuisance functions, so propensity and outcome regression errors enter only through higher-order terms. Combined with localized forest or polynomial methods, orthogonal moments support pointwise estimation and inference for heterogeneous treatment effects (Chernozhukov et al., 2018; Athey et al., 2019; Friedberg et al., 2021; Oprescu et al., 2019; Nie & Wager, 2021; Kennedy, 2023; Kennedy et al., 2024; Foster & Syrgkanis, 2023). For the conditional geodesic treatment effect $\tau(x)$, however, the two conditional Fréchet means are defined implicitly, their estimation errors lie in different tangent spaces, and $\mu_0(x)$ also determines the base point of the logarithm map. The first-order error of $\tau(x)$ therefore requires the two endpoint errors to be transported to a common tangent space and propagated jointly through the logarithm map. This leads to the central question of this paper:

> *How can we develop Neyman-orthogonal estimation and pointwise inference for conditional geodesic treatment effects from observational data?*

To address this question, we propose the **Geodesic Orthogonal learner (GO-learner)** for pointwise estimation of the conditional geodesic treatment effect. GO-learner constructs Neyman-orthogonal moment functions for the two treatment-specific conditional Fréchet means and jointly linearizes the logarithm map to propagate their estimation errors to $\tau(x)$ in a common tangent space. The resulting expansion separates stochastic variation, nuisance estimation, localization bias, and second-order terms from endpoint estimation and the logarithm map. It also identifies the smoothness of the linearized treatment contrast as the quantity governing the leading localization bias and shows when separate arm-wise Fréchet regression incurs an additional bias because the two treatment arms have different local covariate distributions. The expansion yields convergence guarantees and supports a consistent covariance estimator and asymptotically valid pointwise confidence sets.

Our **contributions** are threefold: (i) We develop GO-learner for pointwise estimation of conditional geodesic treatment effects from observational data. (ii) We derive a joint error expansion yielding convergence guarantees and asymptotically valid pointwise confidence sets, and show that separate arm-wise Fréchet smoothing can incur an additional bias term when the two treatment arms induce different local covariate distributions. (iii) In the primary matrix-outcome benchmark, GO-learner has lower average MSE among the three evaluated estimators with sample size increasing. The nuisance-misspecification results show the bias reduction predicted when either nuisance class is correctly specified, and pointwise coverage approaches the nominal level under inference-specific bandwidths.

2 Related work

Fréchet regression and manifold inference. Classical manifold statistics established large-sample theory for intrinsic and extrinsic means (Bhattacharya & Patrangenaru, 2003). Regression methods for manifold-valued and random object responses include geodesic regression (Fletcher, 2013), intrinsic local polynomial regression for symmetric positive-definite matrices (Yuan et al., 2012), and global and local Fréchet regression for general metric-space-valued responses (Petersen & Müller, 2019). Random forest weighting gives locally adaptive Fréchet regression (Qiu et al., 2024), and Kurisu & Otsu (2026) develop empirical-likelihood inference for Fréchet means and local Fréchet regression. These methods estimate a single conditional center. Our target couples two treatment-specific conditional Fréchet means through a logarithm map and adjusts for observational treatment assignment.

Causal inference with non-Euclidean outcomes. Lin et al. (2023) define causal effects for distribution-valued outcomes in Wasserstein space and derive doubly robust estimators and asymptotic theory. Kurisu et al. (2024) introduce the geodesic average treatment effect for general geodesic metric spaces and develop doubly robust estimation, convergence theory, and inference for this

population-level target. They also discuss recovering treatment-specific conditional Fréchet means by minimizing doubly robust estimates of conditional Fréchet risks. Shin et al. (2024) study absolute average and median treatment effects in general metric spaces. Geodesic causal ideas have also been developed for regression discontinuity designs and difference-in-differences (Kurisu et al., 2025; Zhou et al., 2026). Other recent approaches study continuous treatments with random object outcomes through Hilbert-space embeddings (Bhattacharjee et al., 2025), topological causal effects based on persistence summaries (Kim & Lee, 2026), and conditional kernel treatment effects with consistency and asymptotic normality (Näf et al., 2026). Our target is the covariate-indexed tangent displacement between treatment-specific conditional Fréchet means, with pointwise orthogonal estimation and inference for this tangent-valued estimand.

Orthogonal estimation and inference for heterogeneous treatment effects. For Euclidean outcomes, causal forests, meta-learners, residualized learners, and local moment methods estimate heterogeneous effects (Wager & Athey, 2018; Künzel et al., 2019; Athey et al., 2019; Nie & Wager, 2021). Double/debiased machine learning and orthogonal statistical learning formalize first-order robustness to nuisance estimation (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023). Oprescu et al. (2019) combine orthogonal moments with forest localization and establish convergence rates and asymptotic normality. Kennedy (2023) develops doubly robust pointwise CATE estimators whose error depends on products of nuisance errors, and Kennedy et al. (2024) derive minimax rates and local polynomial estimators for smooth CATEs. These results motivate localized orthogonal moments for conditional causal targets. Our analysis treats two implicitly defined conditional Fréchet means and propagates their joint first-order errors through the logarithm map.

3 Problem Setup

Let $\{(X_i, A_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample from the distribution of (X, A, Y), where $X \in [0,1]^p$ denotes pretreatment covariates, $A \in \{0,1\}$ is the treatment, and $Y \in \mathcal{M}$ is the observed outcome. The outcome space $\mathcal{M}$ is a known, complete Riemannian manifold. Let $d_{\mathcal{M}} = \dim(\mathcal{M})$, and let $d(\cdot,\cdot)$ denote its geodesic distance. For $a \in \{0,1\}$, let $Y(a)$ denote the potential outcome under treatment a. We focus on a fixed interior point $x_0 \in (0,1)^p$ and define

$$Q_a(m,x) = \tfrac{1}{2}\mathbb{E}\big\{d^2(Y(a),m) \mid X = x\big\}, \qquad \mu_a(x) = \operatorname*{arg\,min}_{m\in\mathcal{M}} Q_a(m,x). \tag{2}$$

Our target is the conditional geodesic treatment effect $\tau(x) = \mathrm{Log}_{\mu_0(x)} \mu_1(x) \in T_{\mu_0(x)}\mathcal{M}$, with $\mu_0(x)$ and $\mu_1(x)$ the two endpoints of its defining geodesic.

Assumption 1 (Causal identification)**.** *Assume consistency and no interference, so that $Y = Y(A)$, and conditional exchangeability $Y(a) \perp\!\!\!\perp A \mid X$ for $a \in \{0,1\}$. For some fixed $\epsilon \in (0, 1/2)$, the true propensity score $e_{0,a}(x) = \mathbb{P}(A = a \mid X = x)$ lies in $[\epsilon, 1-\epsilon]$ in a neighborhood $\mathcal{N}_0$ of x_0. The density of X is bounded above and away from zero on $\mathcal{N}_0$.*

Assumption 2 (Regular geometry)**.** *For each $a \in \{0,1\}$ and $x \in \mathcal{N}_0$, $Q_a(\cdot, x)$ has a unique global minimizer in a compact set $\mathcal{K}_0$. There exists a compact geodesically convex set $\mathcal{K}$, with $\mathcal{K}_0 \subset \operatorname{int}\mathcal{K}$, contained in a strongly convex normal neighborhood $\mathcal{U}$. The conditional risks are jointly continuous on $\mathcal{U} \times \mathcal{N}_0$. The support of $Y(a) \mid X = x$ is contained in a common compact set on which $(m,y) \mapsto \mathrm{Log}_m y$ is smooth, with its first two covariant derivatives in m uniformly bounded. At the conditional Fréchet means, the negative derivative of the conditional mean log map has eigenvalues in $[\lambda, \Lambda]$ for fixed $0 < \lambda \le \Lambda < \infty$.*

These conditions ensure that the conditional Fréchet means and the local geometric expansions used below are well defined. They hold, for example, for well-conditioned positive-definite matrices under the affine-invariant metric and for spherical outcomes supported in a sufficiently small common ball (Pennec et al., 2006; Afsari, 2011). The Riemannian metric is part of the estimand.

Let ∇ denote the Levi-Civita covariant derivative. For $a \in \{0,1\}$, define

$$g_{0,a}(m,x) = \mathbb{E}(\mathrm{Log}_m Y \mid A = a, X = x), \tag{3}$$

$$\mathcal{H}_m(y) = -\nabla_m \mathrm{Log}_m y, \qquad H_{0,a}(m,x) = \mathbb{E}\{\mathcal{H}_m(Y) \mid A = a, X = x\}. \tag{4}$$

Here $\mathcal{H}_m(y)$ is the Hessian of $d^2(m,y)/2$. Parallel transport $\mathsf{P}_{m\to m'}$ is taken along the unique geodesic joining m and m' within $\mathcal{U}$.

Proposition 1 (Identification). *Under Assumptions 1 and 2, $\mu_a(x_0)$ and $\tau(x_0)$ are identified. Moreover, $g_{0,a}\{\mu_a(x_0), x_0\} = 0$, and the derivative of $g_{0,a}$ with respect to m at $\{\mu_a(x_0), x_0\}$ is nonsingular.*

The proof is deferred to Appendix A. Thus $\tau(x_0)$ is determined by the observed conditional outcome distributions under the causal and geometric assumptions.

4 Orthogonal estimation

4.1 Orthogonal mean equations and sample-split reference points

For candidate functions $e_a \in [\epsilon, 1-\epsilon]$, $g_a(m,x) \in T_m\mathcal{M}$, and self-adjoint operators $h_a(m,x) : T_m\mathcal{M} \to T_m\mathcal{M}$, set

$$\psi_a(O, m, e, g) = g_a(m, X) + \frac{\mathbf{1}(A=a)}{e_a(X)}\{\mathrm{Log}_m Y - g_a(m, X)\}, \tag{5}$$

$$\chi_a(O, m, e, h) = h_a(m, X) + \frac{\mathbf{1}(A=a)}{e_a(X)}\{\mathcal{H}_m(Y) - h_a(m, X)\}. \tag{6}$$

We estimate e_1 and set $e_0 = 1 - e_1$. The candidate point m specifies the response being regressed: the conditional mean of $\mathrm{Log}_m Y$ generally differs from $\mathrm{Log}_m \mu_a(x)$.

Proposition 2 (Conditional moment errors). *For every fixed candidate point and covariate value,*

$$\mathbb{E}(\psi_a \mid X = x) - g_{0,a}(m, x) = \frac{e_a(x) - e_{0,a}(x)}{e_a(x)}\{g_a(m, x) - g_{0,a}(m, x)\}, \tag{7}$$

$$\mathbb{E}(\chi_a \mid X = x) - H_{0,a}(m, x) = \frac{e_a(x) - e_{0,a}(x)}{e_a(x)}\{h_a(m, x) - H_{0,a}(m, x)\}. \tag{8}$$

The conditional moment maps are Neyman orthogonal with respect to the nuisance functions: their first derivatives with respect to nuisance perturbations vanish at the true functions.

Appendix A gives the proof of Proposition 2. Equations (7) and (8) show that each conditional moment is exact when either its propensity or regression nuisance is exact, and otherwise the two nuisance errors enter multiplicatively.

We split the sample into four disjoint blocks. Let $n_b = |I_b| \asymp n$ for $b \in \{\mathrm{ref}, \mathrm{reg}, \mathrm{cal}, \mathrm{final}\}$. The reference block I_{ref} estimates treatment-specific reference points $\widetilde{m}_a$, the regression block I_{reg} estimates the propensity and conditional vector and Hessian regressions, the calibration block I_{cal} selects final localization tuning and calibrates the Hessian operators, and the final block I_{final} evaluates the orthogonal moments. Conditional on final covariates, the final treatments and outcomes are independent of all auxiliary fits and tuning, as required by the theory. Arm-wise Fréchet regression on I_{ref} first yields $\widetilde{m}_a$. Conditional on these reference points, $\mathrm{Log}_{\widetilde{m}_a} Y$ and $\mathcal{H}_{\widetilde{m}_a}(Y)$ lie in fixed tangent spaces, allowing standard vector and operator regressions on I_{reg}, denoted by $\widehat{g}_a$ and $\widehat{h}_a$. Nuisance estimation is needed only at the realized reference points. We do not require the fitted nuisance functions to be smooth in the candidate point m. Candidate-point expansions use the smooth population log-moment map implied by Assumption 2. Because $n_{\mathrm{final}} \asymp n$, rates below are written in terms of the total sample size n.

4.2 Local weights and the one-step update

Let $\alpha_i(x_0) \geq 0$, $i \in I_{\mathrm{final}}$, sum to one. They can be normalized box-kernel weights or weights from a partition trained on auxiliary observations. Given degree ℓ and bandwidth h_n, let $r_\ell(u)$ contain all monomials of total degree at most ℓ, with the constant first. For a nonnegative integer multi-index $\nu = (\nu_1, \ldots, \nu_p)$, write $u^\nu = \prod_j u_j^{\nu_j}$ and $|\nu| = \sum_j \nu_j$. With $\mathbf{e}_1$ denoting the first coordinate vector, define

$$\Gamma_n = \sum_i \alpha_i r_\ell((X_i - x_0)/h_n) r_\ell((X_i - x_0)/h_n)^\top,$$

$$\omega_i = \alpha_i \mathbf{e}_1^\top \Gamma_n^{-1} r_\ell((X_i - x_0)/h_n), \qquad q_n^2 = \sum_i \omega_i^2. \tag{9}$$

Algorithm 1 Pointwise estimation and inference at x_0

Require: Observations, metric and manifold operations, four disjoint blocks, degrees, bandwidth grids, spectral tolerances, and numerical tolerances.

1: Fit arm-specific Fréchet reference points on I_{ref}.
2: On I_{reg}, fit clipped $\widehat{e}_1$, set $\widehat{e}_0 = 1 - \widehat{e}_1$, and fit vector/operator responses at each reference point.
3: Use auxiliary data to choose estimation tuning and optional partitions. Calibrate the two Hessians on I_{cal}.
4: Using only final covariates, form $\alpha_i, \Gamma_n, \omega_i, q_n$.
5: Evaluate the final arm moments and compute the safeguarded update (10).
6: Form the covariance (22) and, for inference tuning, the confidence set (24). Return the estimate and the requested inference output.

These signed weights reproduce polynomials through degree ℓ. An analogous local-polynomial average of (6) on I_{cal} estimates $H_{0,a}(\widetilde{m}_a, x_0)$. Symmetrize and clip its eigenvalues to $[\lambda/2, 2\Lambda]$, giving $\widehat{H}_a$. The numerical thresholds are fixed before final-sample evaluation.

For $\widehat{\psi}_{i,a} = \psi_a(O_i, \widetilde{m}_a, \widehat{e}, \widehat{g})$, compute

$$\widehat{\mu}_a = \mathrm{Exp}_{\widetilde{m}_a}\left(\widehat{H}_a^{-1}\sum_i \omega_i \widehat{\psi}_{i,a}\right), \qquad \widehat{\tau} = \mathrm{Log}_{\widehat{\mu}_0}\widehat{\mu}_1. \tag{10}$$

Endpoint updates outside $\mathcal{K}$ return their reference points. A singular weight Gram matrix returns both reference points. Appendix G specifies these safeguards, the reference points, calibration weights, regressions, auxiliary validation loss, and an optional query-specific honest forest. Algorithm 1 summarizes the fitted objects.

Proposition 3 (Euclidean reduction)**.** *On $\mathcal{M} = \mathbb{R}^{d_{\mathcal{M}}}$, $g_{0,a}(m, x) = \mathbb{E}(Y \mid A = a, X = x) - m$, $H_{0,a} = I_{d_{\mathcal{M}}}$, and the logarithm-map differential maps (u_0, u_1) to $u_1 - u_0$. With $\widehat{g}_a(m, x) = \widehat{\mu}_a^{\mathrm{reg}}(x) - m$ and $\widehat{H}_a = I_{d_{\mathcal{M}}}$, the untruncated estimator is exactly the local-polynomial weighted difference of the two augmented inverse-probability weighted conditional-mean pseudo-outcomes.*

Appendix A also gives the proof of Proposition 3. Thus GO-learner reduces to the standard local augmented inverse-probability weighted CATE estimator when the outcome space is Euclidean.

5 Error expansion and inference

5.1 Joint linearization of the treatment contrast

Write $m_a = \mu_a(x_0)$ and $H_a = H_{0,a}(m_a, x_0)$. Define

$$\Phi(m_0', m_1') = \mathsf{P}_{m_0' \to m_0}\,\mathrm{Log}_{m_0'}\, m_1', \qquad \mathcal{J}(\cdot) = \mathrm{D}\Phi\big|_{(m_0, m_1)}(\cdot). \tag{11}$$

This differential maps $T_{m_0}\mathcal{M} \times T_{m_1}\mathcal{M}$ to $T_{m_0}\mathcal{M}$. Along the geodesic joining the endpoints, it equals $\nabla_t J(0)$, where the Jacobi field J describes the first-order change of the joining geodesic under the prescribed endpoint perturbations. Appendix A proves this identity and the required second-order bounds. Put

$$b_a(x') = g_{0,a}(m_a, x'), \qquad \vartheta_{x_0}(x') = \mathcal{J}\{H_0^{-1}b_0(x'), H_1^{-1}b_1(x')\}. \tag{12}$$

The function ϑ_{x_0} determines the leading localization bias.

Assumption 3 (Local smoothness)**.** *The functions b_a, ϑ_{x_0}, and $x' \mapsto H_{0,a}(m_a, x')$ belong locally to Hölder classes of orders $s_\mu, s_\tau, s_H > 0$, with fixed radius. A function is Hölder of order s here if it has a degree-$\lceil s \rceil - 1$ Taylor approximation with remainder bounded by $C\,\|x' - x_0\|^s$. The corresponding derivative bounds hold in a neighborhood. Choose $\ell \geq \lceil \max(s_\mu, s_\tau) \rceil - 1$.*

This assumption concerns ϑ_{x_0} as defined in (12). Its relationship to the transported function $\tau(x')$ includes both endpoint-quadratic and Hessian-variation terms (Proposition 8). The finite-support positive-definite model in Proposition 17 makes ϑ_{x_0} smoother than both b_0 and b_1 without a commutativity assumption.

Assumption 4 (Honest local weights). *Let $\mathcal{F}_n$ contain the auxiliary samples, partition randomization, and final covariates, but no final treatments or outcomes. On an event $\mathcal{W}_n \in \mathcal{F}_n$, the weights are $\mathcal{F}_n$-measurable and satisfy*

$$\omega_i = 0 \text{ if } \|X_i - x_0\|_\infty > h_n, \quad \sum_i \omega_i = 1, \quad \sum_i \omega_i (X_i - x_0)^\nu = 0 \ (1 \le |\nu| \le \ell),$$
$$\sum_i |\omega_i| \le C_w, \quad \max_i |\omega_i| \le C_w/(n_{\text{final}} h_n^p), \quad c_w/(n_{\text{final}} h_n^p) \le q_n^2 \le C_w/(n_{\text{final}} h_n^p). \tag{13}$$

Lemma 9 in Appendix B verifies these conditions for both implementations and bounds $\mathbb{P}(\mathcal{W}_n^c)$. Conditional on $\mathcal{F}_n$, the final pairs (A_i, Y_i) are independent. Define

$$\xi_{i,a} = \mathsf{P}_{\widetilde{m}_a \to m_a} \left[\widehat{\psi}_{i,a} - \mathbb{E}\{\widehat{\psi}_{i,a} \mid \mathcal{F}_n\} \right],$$
$$v_{i,n} = \mathcal{J}(H_0^{-1}\xi_{i,0}, H_1^{-1}\xi_{i,1}), \qquad B_{\tau,n} = \sum_i \omega_i \vartheta_{x_0}(X_i). \tag{14}$$

These centered moment residuals may depend on n and the auxiliary data.

5.2 A CONDITIONAL MEAN-SQUARE BOUND

All suprema below are taken over $\|x' - x_0\|_\infty \le h_n$. Define

$$r_{a,n} = d(\widetilde{m}_a, m_a), \qquad D_{a,n} = \|\widehat{e}_a - e_{0,a}\|_\infty \, \|\widehat{g}_a(\widetilde{m}_a, \cdot) - g_{0,a}(\widetilde{m}_a, \cdot)\|_\infty,$$
$$E_{a,n}^H = \left\| \mathsf{P}_{\widetilde{m}_a \to m_a} \widehat{H}_a \mathsf{P}_{m_a \to \widetilde{m}_a} - H_a \right\|_{\text{op}}, \qquad \rho_{a,n} = \left\| \sum_i \omega_i H_{0,a}(m_a, X_i) - H_a \right\|_{\text{op}}. \tag{15}$$

The vector suprema are taken with respect to the tangent-space norm. We collect these first-stage errors in

$$\mathcal{R}_n = \sum_{a=0}^{1} \left\{ D_{a,n} + r_{a,n}^2 + (E_{a,n}^H + \rho_{a,n}) r_{a,n} + E_{a,n}^H (q_n + h_n^{s_\mu}) \right\}. \tag{16}$$

These quantities are used only to state the theoretical error bounds and are not required by the algorithm. For the analysis, let $\mathcal{A}_n \in \mathcal{F}_n$ denote the event on which $\mathcal{W}_n$ holds and

$$h_n \le h_0, \qquad \max_a r_{a,n} \le r_0, \qquad \max_a D_{a,n} \le d_0, \qquad \max_a E_{a,n}^H \le \lambda/2,$$

for sufficiently small fixed constants $h_0, r_0, d_0 > 0$. Throughout the following results, C denotes a finite constant independent of n.

Theorem 4 (Joint expansion and mean-square error). *Under Assumptions 1 through 4, with uniformly norm-bounded nuisance fits, on $\mathcal{A}_n$,*

$$\mathsf{P}_{\widehat{\mu}_0 \to m_0} \widehat{\tau} - \tau(x_0) = B_{\tau,n} + \sum_i \omega_i v_{i,n} + R_{\tau,n}, \tag{17}$$

where

$$\left\{ \mathbb{E}(\|R_{\tau,n}\|^2 \mid \mathcal{F}_n) \right\}^{1/2} \le C\{\mathcal{R}_n + q_n^2 + h_n^{2s_\mu}\}. \tag{18}$$

In particular, the conditional mean squared transported tangent-space error is at most

$$C\{q_n^2 + h_n^{2s_\tau} + h_n^{4s_\mu} + \mathcal{R}_n^2\}. \tag{19}$$

The unconditional risk is bounded by the expectation of (19) on $\mathcal{A}_n$, plus $C\mathbb{P}(\mathcal{A}_n^c)$.

Appendix C gives the complete proof. Equation (17) separates localization bias, centered sampling variation, and the remainder generated by first-stage and geometric errors.

Corollary 5 (Convergence upper bound). *Suppose $\mathbb{P}(\mathcal{A}_n^c) \to 0$ and $\mathcal{R}_n = o_p(q_n + h_n^{s_\tau} + h_n^{2s_\mu})$. Then*

$$\|\mathsf{P}_{\widehat{\mu}_0 \to m_0} \widehat{\tau} - \tau(x_0)\| = O_p\{(nh_n^p)^{-1/2} + h_n^{s_\tau} + h_n^{2s_\mu}\}.$$

Writing $s = \min(s_\tau, 2s_\mu)$, the choice $h_n \asymp n^{-1/(2s+p)}$ gives $O_p(n^{-s/(2s+p)})$, when the first-stage condition holds at that bandwidth. If also $\mathbb{E}[\mathbf{1}_{\mathcal{A}_n} \mathcal{R}_n^2] = O(n^{-2s/(2s+p)})$ and $\mathbb{P}(\mathcal{A}_n^c) = O(n^{-2s/(2s+p)})$, the squared risk is $O(n^{-2s/(2s+p)})$.

Appendix C gives the proof of Corollary 5. The rate contains the local variance term q_n, the localization bias $h_n^{s_\tau}$, the second-order endpoint term $h_n^{2s_\mu}$, and the first-stage remainder $\mathcal{R}_n$. The risk conclusion follows from an explicit moment bound. Proposition 16 in Appendix F verifies the required moments and failure probabilities for concrete first stages.

Comparison with separate arm-wise Fréchet regression. The $h_n^{s_\tau}$ term in Corollary 5 reflects smoothness of ϑ_{x_0} and can also benefit a plug-in estimator. To isolate the difference, write $\mathcal{J}(u_0, u_1) = J_0 u_0 + J_1 u_1$ and $f_a(x) = J_a H_a^{-1} b_a(x)$, so that $\vartheta_{x_0} = f_0 + f_1$. Let $\mathcal{L}_{a,h}$ denote the population degree-ℓ local-polynomial intercept operator under the arm-specific design $X \mid A = a$, and let $m_{a,h}^F$ be the corresponding population local Fréchet root. Proposition 12 proves

$$\Phi(m_{0,h}^F, m_{1,h}^F) - \tau(x_0) = \mathcal{L}_{0,h}\vartheta_{x_0} + (\mathcal{L}_{1,h} - \mathcal{L}_{0,h})f_1 + R_F(h), \tag{20}$$

where $\|R_F(h)\| \le C\{h^{2s_\mu} + h^{s_\mu + \bar{s}_H}\}$ and $\bar{s}_H = \min(s_H, \ell + 1)$. If the propensity is locally Hölder of order $s_e > 0$, define $\bar{s}_e = \min(s_e, 1)$. Then $\|(\mathcal{L}_{1,h} - \mathcal{L}_{0,h})f_1\| \le Ch^{s_\mu + \bar{s}_e}$. The first localization term in (20) is shared with GO-learner. The second comes from the difference between the two arm-specific smoothing operators. GO-learner uses the same pooled final-covariate weights for both endpoints, so this term is absent from its leading localization bias. Corollary 13 gives the corresponding empirical rate and states when $h^{s_\mu + \bar{s}_e}$ is no larger than the other leading bias terms. Proposition 18 gives a noncommuting positive-definite construction in which this additional term changes the leading bias-variance rate calculation. Appendix D gives the proofs.

5.3 Covariance estimation and confidence sets

For tangent vectors u, v in a common tangent space, let $u \otimes v$ denote the rank-one operator $w \mapsto \langle v, w\rangle u$. For a linear map L, let L^* denote its adjoint. Define the conditional covariance of the leading term by

$$\Sigma_n = q_n^{-2} \sum_i \omega_i^2 \mathbb{E}(v_{i,n} \otimes v_{i,n} \mid \mathcal{F}_n). \tag{21}$$

For computation, define $\widetilde{\Phi}(z_0, z_1) = \mathsf{P}_{z_0 \to \widetilde{m}_0} \operatorname{Log}_{z_0} z_1$ and $\widetilde{\mathcal{J}} = \mathrm{D}\widetilde{\Phi}|_{(\widetilde{m}_0, \widetilde{m}_1)}$, which maps $T_{\widetilde{m}_0}\mathcal{M} \times T_{\widetilde{m}_1}\mathcal{M}$ to $T_{\widetilde{m}_0}\mathcal{M}$. Set

$$U_i = \widetilde{\mathcal{J}}(\widehat{H}_0^{-1}\widehat{\psi}_{i,0}, \widehat{H}_1^{-1}\widehat{\psi}_{i,1}), \qquad \beta_i = \omega_i^2/q_n^2, \qquad \overline{U} = \sum_i \beta_i U_i,$$

$$\widetilde{\Sigma}_n = \sum_i \beta_i (U_i - \overline{U}) \otimes (U_i - \overline{U}), \qquad \widehat{\Sigma}_n = \mathsf{P}_{\widetilde{m}_0 \to \widehat{\mu}_0} \widetilde{\Sigma}_n \mathsf{P}_{\widehat{\mu}_0 \to \widetilde{m}_0}. \tag{22}$$

The covariance is computed from the joint arm moments at fixed reference points and then transported to the estimated base point $\widehat{\mu}_0$.

Theorem 6 (Studentized pointwise inference)**.** *Assume Theorem 4,* $\mathbb{P}(\mathcal{A}_n^c) \to 0$, $h_n \to 0$, $q_n \to_p 0$, $\max_i |\omega_i|/q_n \to_p 0$, *and*

$$\max_a (r_{a,n} + E_{a,n}^H + D_{a,n}) \to_p 0, \qquad h_n^{s_\tau} + h_n^{2s_\mu} + \mathcal{R}_n = o_p(q_n). \tag{23}$$

Suppose, with probability tending to one, that $\lambda_{\min}(\Sigma_n) \ge c > 0$. *Then*

$$\left\| \mathsf{P}_{\widehat{\mu}_0 \to m_0} \widehat{\Sigma}_n \mathsf{P}_{m_0 \to \widehat{\mu}_0} - \Sigma_n \right\|_{\mathrm{op}} \to_p 0.$$

For $\alpha \in (0, 1)$,

$$\mathcal{C}_{n,\alpha} = \left\{ t \in T_{\widehat{\mu}_0}\mathcal{M} : q_n^{-2} \left\langle \widehat{\tau} - t, \widehat{\Sigma}_n^{-1}(\widehat{\tau} - t) \right\rangle \le \chi^2_{d_\mathcal{M}, 1-\alpha} \right\}, \tag{24}$$

where $\chi^2_{d_\mathcal{M}, 1-\alpha}$ *is the* $1 - \alpha$ *quantile of a chi-squared distribution with* $d_\mathcal{M}$ *degrees of freedom. Moreover,*

$$\mathbb{P}\{\mathsf{P}_{m_0 \to \widehat{\mu}_0} \tau(x_0) \in \mathcal{C}_{n,\alpha}\} \to 1 - \alpha.$$

Appendix E gives the proof and covariance-consistency argument. The confidence set covers the true effect after parallel transport from $T_{m_0}\mathcal{M}$ to $T_{\widehat{\mu}_0}\mathcal{M}$. Condition (23) requires localization and first-stage remainders to be smaller than q_n, the scale of the leading stochastic term. Bandwidth selection is specified in Appendix G.3.

6 EXPERIMENTS

This section verifies the following three questions tied to our main theoretical results: **(1)** How accurately does GO-learner estimate the same target as Fréchet regression baselines? **(2)** Does inference-specific bandwidth selection improve pointwise coverage as sample size increases? **(3)** Do the augmented moments show the bias reduction predicted under nuisance misspecification?

6.1 EXPERIMENTAL DESIGN

Setup. We generate observational data with 2×2 positive-definite matrix outcomes under the affine-invariant metric. Covariates are uniform on $[0,1]^2$. The two conditional Fréchet means vary nonlinearly with covariates and generally do not commute. Outcomes are generated as

$$Y(a) = \mathrm{Exp}_{\mu_a(X)}\left\{\sigma \mathsf{P}_{I \to \mu_a(X)} \varepsilon_a\right\}, \qquad \sigma = 0.3,$$

where I is the identity matrix and the three tangent coordinates of ε_a are independent uniform variables on $[-\sqrt{3}, \sqrt{3}]$. Treatment has propensity $e_{0,1}(x) = 1/2 + (1/5)\sin\{\pi(x_1 + x_2)\}$, and only $Y = Y(A)$ is observed. We use $n = 1000 \times 2^j$ for $j = 0, \ldots, 6$. A separate finite-support matrix model gives exact control over nuisance misspecification. Appendix H.1 specifies both generators.

Methods and evaluation. We use the local-linear kernel implementation of GO-learner. The baselines are arm-wise local-linear Fréchet regression and RF-weighted local-constant Fréchet regression (Qiu et al., 2024), followed by the logarithm map between their fitted means. Within each repetition, GO-learner divides the sample equally across the four blocks, while the baselines fit on all n observations. Hyperparameter grids and validation rules are given in Appendix H.2. The augmentation ablation removes the corrections from both orthogonal moments with the remaining GO construction unchanged.

Point-estimation performance is measured by the mean squared transported tangent-space error (MSE), whose repetition-level loss is

$$\left\| \mathsf{P}_{\widehat{\mu}_0(x) \to \mu_0(x)} \widehat{\tau}(x) - \tau(x) \right\|^2_{\mu_0(x)}.$$

We average this loss over the three fixed queries $(1/2, 1/2)$, $(1/4, 1/2)$, and $(3/4, 1/2)$ within each of 100 independent repetitions. Pointwise inference uses 500 repetitions at $x_0 = (1/2, 1/2)$ and nominal 95% vector ellipsoids from Theorem 6. The practical curves use validation-selected first stages and compare the estimation bandwidth h_{est} with the prespecified inference bandwidth $h_{\mathrm{inf}} = n^{-1/5}$ while keeping those first stages fixed. Appendix H.4 also reports the explicit first-stage schedule from Proposition 16. Ground truth is used only for evaluation.

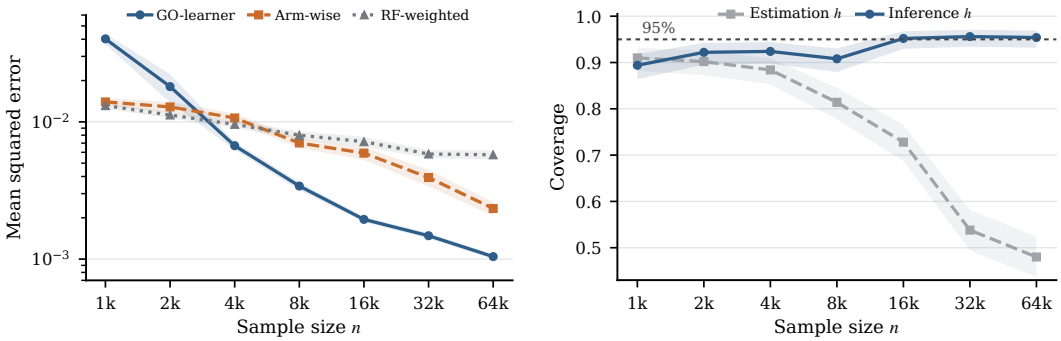
6.2 ESTIMATION AND POINTWISE INFERENCE

Estimation accuracy. Figure 1(a) shows a clear crossover in the primary benchmark. The two baselines have lower average MSE at $n = 1000$ and 2000, while GO-learner has the lowest average MSE for every $n \geq 4000$. At $n = 64000$, the average MSE is 1.04×10^{-3} for GO-learner, compared with 2.34×10^{-3} for arm-wise local-linear Fréchet regression and 5.75×10^{-3} for RF-weighted local-constant Fréchet regression, corresponding to reductions of 55% and 82% for these implementations. For every $n \geq 4000$, both baseline-minus-GO paired mean differences exceed three paired MCSEs. Appendix H.3 reports the query-wise results and paired MCSEs.

Pointwise inference. Figure 1(b) separates estimation tuning from inference tuning. Reusing h_{est} leaves localization bias large relative to the standard error, and coverage is 48.0% at $n = 64000$. Changing only the final bandwidth to h_{inf} gives coverage of 95.2%, 95.6%, and 95.4% at the three largest sample sizes. At the larger sample sizes, the smaller bandwidth substantially reduces squared Monte Carlo bias at the cost of wider confidence sets, as detailed in Appendix H.4. Coverage remains below nominal at the smaller tested sample sizes.

6.3 NUISANCE ROBUSTNESS

To assess Proposition 2, we fit the propensity and the vector and Hessian regressions (g, h) using correctly specified quadratic classes or misspecified intercept-only classes. "Correct" means the true



(a) Estimation accuracy.

(b) Pointwise inference.

Figure 1: **Main estimation and inference results.** (a) Mean squared transported tangent-space error averaged over three queries, with ± 1 MCSE over 100 repetitions. (b) Coverage of nominal 95% vector confidence sets, with 95% Wilson intervals over 500 repetitions. The inference curves share fitted first stages and differ only in the final bandwidth.

Table 1: **Nuisance misspecification and augmentation ablation.** Results at $n = 64000$ over 100 paired repetitions. MSE is reported as mean (MCSE). MSE and squared Monte Carlo bias are reported in units of 10^{-4} and 10^{-6}, respectively.

		GO-learner		Without augmentation	
e	(g, h)	MSE	MC Bias^2	MSE	MC Bias^2
Correct	Correct	1.32 (0.10)	0.185	0.32 (0.03)	1.205
Correct	Misspecified	1.33 (0.10)	0.169	1.35 (0.04)	125.193
Misspecified	Correct	0.98 (0.08)	0.249	0.32 (0.03)	1.205
Misspecified	Misspecified	1.34 (0.10)	37.630	1.35 (0.04)	125.193

nuisance function belongs to the fitted class, with coefficients estimated from I_{reg}. Table 1 reports all four combinations at the largest sample size and a no-augmentation ablation with the remaining construction fixed.

At $n = 64000$, joint misspecification gives squared Monte Carlo bias 3.76×10^{-5}, while the three cases with at least one correctly specified nuisance class are at most 2.49×10^{-7} and below the Monte Carlo variation expected from averaging 100 repetitions, as quantified in Appendix H.5. With a correct propensity and misspecified regressions, removing augmentation raises squared Monte Carlo bias from 1.69×10^{-7} to 1.25×10^{-4}. With correctly specified regressions, the no-augmentation ablation has lower MSE, consistent with the additional variance introduced by the residual correction. These patterns agree with Proposition 2 and illustrate the finite-sample variance cost of augmentation. Appendix H.5 reports the full sample-size paths and bias calculation.

7 Conclusion

In this paper, we develop GO-learner for pointwise estimation and inference of the conditional geodesic treatment effect from observational data. Neyman-orthogonal moments estimate treatment-specific conditional Fréchet means, while a joint logarithm-map expansion propagates their errors to the effect in a common tangent space. This yields convergence guarantees and asymptotically valid pointwise confidence sets, with the Euclidean case reducing to augmented inverse-propensity CATE estimation. In the primary matrix benchmark, GO-learner has the lowest average MSE among the three evaluated estimators for large sample size, and the robustness and inference experiments agree with our main theoretical findings.

Limitations and future work. The current theory is pointwise at an interior covariate value and assumes known geometry. An interesting next step is to extend the analysis to settings where the geometry is estimated from data, which requires accounting for the resulting error in the conditional Fréchet means and logarithm map.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript for better clarity and presentation. The authors derived the mathematical proofs, conducted experiments, and analyzed the results.

REFERENCES

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

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

Satarupa Bhattacharjee, Bing Li, Xiao Wu, and Lingzhou Xue. Doubly robust estimation of causal effects for random object outcomes with continuous treatments. *arXiv preprint arXiv:2506.22754*, 2025.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009. doi: 10.1214/09-AOAS249.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*, volume 66. Chapman and Hall, 1996.

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.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51 (3):879–908, 2023.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10(4):215–310, 1948.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517, 2021.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Edward H. Kennedy, Sivaraman Balakrishnan, James M. Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024. doi: 10.1214/24-AOS2369.

Kwangho Kim and Hajin Lee. Topological causal effects. *arXiv preprint arXiv:2603.02289*, 2026.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Daisuke Kurisu and Taisuke Otsu. Empirical likelihood for manifolds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):91–119, 2026. doi: 10.1093/jrsssb/qkaf043.

Daisuke Kurisu, Yidong Zhou, Taisuke Otsu, and Hans-Georg Müller. Geodesic causal inference. *arXiv preprint arXiv:2406.19604*, 2024.

Daisuke Kurisu, Yidong Zhou, Taisuke Otsu, and Hans-Georg Müller. Regression discontinuity designs for functional data and random objects in geodesic spaces. *arXiv preprint arXiv:2506.18136*, 2025.

Zhenhua Lin, Dehan Kong, and Linbo Wang. Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):378–398, 2023.

Jeffrey Näf, Junhyung Park, and Herbert Susmann. Causal-drf: Conditional kernel treatment effect estimation using distributional random forest. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1000–1008. PMLR, 2026.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pp. 4932–4941. PMLR, 2019.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

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

Rui Qiu, Zhou Yu, and Ruoqing Zhu. Random forest weighted local fréchet regression with random objects. *Journal of Machine Learning Research*, 25(107):1–69, 2024.

Ha-Young Shin, Kyusoon Kim, Kwonsang Lee, and Hee-Seok Oh. Absolute average and median treatment effects as causal estimands on metric spaces. *arXiv preprint arXiv:2407.03726*, 2024.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Ying Yuan, Hongtu Zhu, Weili Lin, and James Stephen Marron. Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(4):697–719, 2012.

Yidong Zhou, Daisuke Kurisu, Taisuke Otsu, and Hans-Georg Müller. Geodesic difference-in-differences. *Biometrika*, pp. asag054, 2026. doi: 10.1093/biomet/asag054.

A GEOMETRY, IDENTIFICATION, AND THE CONTRAST

The arguments use finite-dimensional tangent spaces with their Riemannian norms. An operator between two such spaces is represented in orthonormal frames only for calculation. Changing either frame does not change the norm bounds. Constants below depend on the compact sets, derivative bounds, dimensions, overlap constant, and smoothness radius, and not on sample size. The letter C can denote different finite constants in different inequalities.

A.1 IDENTIFICATION AND MOMENT IDENTITIES

Proof of Proposition 1. Consistency and conditional independence imply, for every bounded measurable f,

$$\mathbb{E}\{f(Y(a)) \mid X\} = \mathbb{E}\{f(Y) \mid A = a, X\} \quad \text{almost surely.}$$

Positivity ensures that both observed conditional distributions are determined for almost every covariate value. In particular, this identifies $Q_a(m, x)$ almost everywhere in x, for each m. Choose a countable dense subset of $\mathcal{U}$. The intersection of the corresponding full-measure covariate sets still has full measure. Continuity in m extends the risk identity from this subset to every $m \in \mathcal{U}$, simultaneously on that set.

Suppose two jointly continuous risk versions on $\mathcal{U} \times \mathcal{N}_0$ agree almost everywhere in x. If they differ at some (m, x_0), continuity supplies a covariate ball on which their difference at m is bounded away from zero. The density lower bound assigns positive probability to this ball, contradicting almost-everywhere equality. The versions therefore agree at x_0 for every $m \in \mathcal{U}$. Since the unique global minimizer is assumed to belong to $\mathcal{K}_0 \subset \mathcal{U}$, minimizing this identified restriction identifies the same $\mu_a(x_0)$. This argument requires the stated location of the global minimizer in addition to local convexity.

On the prescribed outcome support, the first variation of squared distance gives $\nabla_m d^2(m, y)/2 = -\operatorname{Log}_m y$. The compact support and derivative envelopes permit differentiation under the conditional integral, by dominated convergence in a local chart. Consequently,

$$\nabla_m Q_a(m, x) = -g_{0,a}(m, x), \qquad \nabla_m^2 Q_a(m, x) = H_{0,a}(m, x).$$

Equality of the identified differentiable risk functions implies equality of these derivatives. The minimizer is interior, so its gradient is zero. The Hessian eigenvalue lower bound gives nonsingularity. Smoothness of the joining logarithm then identifies $\tau(x_0)$. □

Proof of Proposition 2. For fixed m, x, every vector in the first equation lies in $T_m\mathcal{M}$. Conditional expectation gives

$$\mathbb{E}(\psi_a \mid X = x) = g_a(m, x) + \frac{e_{0,a}(x)}{e_a(x)}\{g_{0,a}(m, x) - g_a(m, x)\}.$$

Subtracting $g_{0,a}(m, x)$ proves (7). Replacing the vector response by the self-adjoint operator response $\mathcal{H}_m(Y)$ gives

$$\mathbb{E}(\chi_a \mid X = x) = h_a(m, x) + \frac{e_{0,a}(x)}{e_a(x)}\{H_{0,a}(m, x) - h_a(m, x)\},$$

which proves (8). For a perturbation $e_t = e_0 + t\dot{e}$, $g_t = g_0 + t\dot{g}$ in the clipped parameter domain, the first error is $t^2\dot{e}\,\dot{g}/(e_0 + t\dot{e})$. Its derivative at zero is zero. The same calculation holds for h_t. This is the asserted Neyman orthogonality. If either factor in a product is exactly zero, the corresponding conditional moment is exactly correct. These algebraic statements do not supply rates for the remaining estimated objects. □

A.2 TRANSPORTED TAYLOR EXPANSIONS

Lemma 7 (Three local geometric expansions)**.** *Uniformly for $m \in \mathcal{K}_0$, sufficiently small $u, t \in T_m\mathcal{M}$, and $x' \in \mathcal{N}_0$, put $m_u = \operatorname{Exp}_m u$. Then*

$$\mathsf{P}_{m_u \to m} g_{0,a}(m_u, x') = g_{0,a}(m, x') - H_{0,a}(m, x')u + R_g, \quad \|R_g\| \le C\|u\|^2, \tag{25}$$

$$\operatorname{Log}_m \operatorname{Exp}_{m_u}(\mathsf{P}_{m\to m_u} t) = u + t + R_E, \qquad \|R_E\| \le C(\|u\| + \|t\|)^2. \quad (26)$$

For $u_a \in T_{m_a}\mathcal{M}$ sufficiently small,

$$\Phi(\operatorname{Exp}_{m_0} u_0, \operatorname{Exp}_{m_1} u_1) = \tau(x_0) + \mathcal{J}(u_0, u_1) + R_\Phi, \quad \|R_\Phi\| \le C(\|u_0\| + \|u_1\|)^2. \quad (27)$$

The differential satisfies $\mathcal{J}(u_0, u_1) = \nabla_t J(0)$, where J solves

$$\nabla_t^2 J + R(J, \dot\gamma)\dot\gamma = 0, \qquad J(0) = u_0, \quad J(1) = u_1,$$

along $\gamma(t) = \operatorname{Exp}_{m_0}(t \operatorname{Log}_{m_0} m_1)$, and R denotes Riemannian curvature with the sign convention in this equation.

Proof. Along $c(s) = \operatorname{Exp}_m(su)$, define $F(s) = \mathsf{P}_{c(s)\to m} g_{0,a}(c(s), x')$. Differentiating a vector field after transport along its own path gives

$$F'(s) = \mathsf{P}_{c(s)\to m} \nabla_{\dot c(s)} g_{0,a}(c(s), x').$$

Because c is geodesic, its velocity is parallel. Differentiating again and using the uniform second covariant derivative bound gives $\|F''(s)\| \le C\|u\|^2$. The integral Taylor identity

$$F(1) = F(0) + F'(0) + \int_0^1 (1-s)F''(s)\, ds$$

proves (25), since $F'(0) = -H_{0,a}(m, x')u$. Differentiation of the true conditional expectation is legitimate by the compact envelopes, and no derivative of a fitted nuisance appears.

For (26), define

$$F_m(u, t) = \operatorname{Log}_m \operatorname{Exp}_{\operatorname{Exp}_m u}(\mathsf{P}_{m\to \operatorname{Exp}_m u} t).$$

The exponential, logarithm, and transport maps are smooth on a sufficiently small common neighborhood of the zero section over $\mathcal{K}_0$. Compactness supplies a uniform radius and a bound on their second derivatives. We have $F_m(u, 0) = u$ and $F_m(0, t) = t$, so $F_m(0, 0) = 0$ and $\mathrm{D}F_m(0, 0)[u, t] = u + t$. Applying the same integral identity to $s \mapsto F_m(su, st)$ bounds the remainder by $C(\|u\| + \|t\|)^2$. This is a Taylor expansion of the coordinate map itself, not an identification of coordinate differentiation with covariant differentiation away from its origin.

Similarly, the map $G(u_0, u_1) = \Phi(\operatorname{Exp}_{m_0} u_0, \operatorname{Exp}_{m_1} u_1)$ is smooth near zero. Compactness of the endpoint sets and absence of conjugate points along the unique normal-neighborhood geodesics give a uniform second derivative bound. Its value and derivative at zero are $\tau(x_0)$ and $\mathcal{J}$. The integral identity proves (27).

For the differential, let $m_a(s) = \operatorname{Exp}_{m_a}(su_a)$, and let γ_s join $m_0(s)$ to $m_1(s)$ in unit time. Smoothness of the logarithm makes this a smooth geodesic variation. Differentiating $\nabla_t \partial_t \gamma_s = 0$ covariantly in s, and commuting the two covariant derivatives, gives the displayed Jacobi equation for $J = \partial_s \gamma_s|_{s=0}$. Its boundary values follow by differentiating the two endpoints. The initial velocity is $\operatorname{Log}_{m_0(s)} m_1(s)$. Differentiation after transport along $m_0(s)$ to m_0 gives its covariant derivative at zero, which, by the torsion-free property, equals $\nabla_t J(0)$. This is exactly the derivative defining $\mathcal{J}$. The smooth normal-neighborhood exponential has invertible derivative, so the boundary-value problem has a unique solution: the difference of two solutions would be a Jacobi field vanishing at both endpoints and hence a kernel vector for that derivative. □

Proposition 8 (Relation to the varying treatment contrast)**.** *For x' such that $u_a(x') = \operatorname{Log}_{m_a} \mu_a(x')$ is sufficiently small,*

$$\begin{aligned}&\left\|\mathsf{P}_{\mu_0(x')\to m_0}\tau(x') - \tau(x_0) - \vartheta_{x_0}(x')\right\| \\ &\quad\le C\sum_{a=0}^{1}\left[\|u_a(x')\|^2 + \|H_{0,a}(m_a, x') - H_a\|_{\mathrm{op}}\,\|u_a(x')\|\right].\end{aligned}$$

Proof. The true mean equation at $\mu_a(x')$ is zero. Applying (25) with $m = m_a$ and $u = u_a(x')$ therefore gives

$$0 = b_a(x') - H_{0,a}(m_a, x')u_a(x') + R_a, \qquad \|R_a\| \le C\|u_a(x')\|^2.$$

Multiplication by H_a^{-1} yields

$$H_a^{-1}b_a(x') - u_a(x') = H_a^{-1}\{H_{0,a}(m_a, x') - H_a\}u_a(x') - H_a^{-1}R_a.$$

The operator norm of H_a^{-1} is at most $1/\lambda$. Substitute these two identities into (12), and compare with (27) at $(u_0(x'), u_1(x'))$. Boundedness of $\mathcal{J}$ gives the asserted inequality. In particular, smoothness of the transported effect alone does not establish the smoothness of ϑ_{x_0}. The Hessian-variation products also matter. □

A.3 Exact Euclidean reduction

Proof of Proposition 3. In Euclidean space, $\mathrm{Log}_m y = y - m$, parallel transport is the identity, $\mathrm{Exp}_m u = m+u$, and $\mathcal{H}_m(y) = I_{d_\mathcal{M}}$. Thus $H_{0,a} = I_{d_\mathcal{M}}$, $\Phi(m_0', m_1') = m_1' - m_0'$, and $\mathcal{J}(u_0, u_1) = u_1 - u_0$. Write $\widehat{r}_a(x) = \widehat{\mu}_a^{\mathrm{reg}}(x)$. Substitution in (5) gives

$$\widehat{\psi}_{i,a} = \widehat{r}_a(X_i) - \widetilde{m}_a + \frac{\mathbf{1}(A_i = a)}{\widehat{e}_a(X_i)}\{Y_i - \widehat{r}_a(X_i)\}.$$

Since $\sum_i \omega_i = 1$, adding $\widetilde{m}_a$ cancels the reference-point term exactly. Subtracting the two updated means gives the asserted augmented pseudo-outcome difference. Both geometric Taylor remainders are identically zero. The reduction concerns the untruncated estimator. A compact-set safeguard is inactive when that estimator is admissible. □

B Local weights and conditional moments

B.1 Weight construction

The kernel implementation uses $\alpha_i = \mathbf{1}(\|X_i - x_0\|_\infty \le h)/N_h$, where N_h is the number of final observations in that cube. The forest implementation averages normalized leaf weights over T_n auxiliary-trained trees. The query leaf $L_t(x_0)$ of every tree is an axis-aligned rectangle containing x_0, with every side length in $[h/3, h]$. Its final count is N_t, and

$$\alpha_i = \frac{1}{T_n}\sum_{t=1}^{T_n}\frac{\mathbf{1}(X_i \in L_t(x_0))}{N_t}.$$

The following result is uniform over all auxiliary-trained partitions with this geometry. It makes no use of their split objective.

Lemma 9 (Local-polynomial weight bounds)**.** *Suppose the density bounds in Assumption 1 hold on the kernel cube or all query leaves. Conditional on the auxiliary data and a bandwidth h, the kernel or forest construction satisfies Assumption 4, except on an event of probability at most*

$$C(T_n + 1)\exp(-cn_{\mathrm{final}}h^p), \tag{28}$$

where $T_n = 1$ for the kernel. Constants depend on fixed p, ℓ and the density bounds. In particular, if T_n is polynomially bounded and $n_{\mathrm{final}}h^p/\log n \to \infty$, the failure probability vanishes.

Proof. Conditioning fixes the query rectangles before final covariates are sampled. Every leaf probability is between constant multiples of h^p. A binomial Chernoff bound therefore puts every N_t between $c_1 n_{\mathrm{final}}h^p$ and $C_1 n_{\mathrm{final}}h^p$, except with the probability order in (28). The same bound controls the final count in the containing cube $\|X - x_0\|_\infty \le h$.

To bound Gram matrices, rescale a leaf by $u = (X - x_0)/h$. The rescaled rectangle contains zero, has all sides in $[1/3, 1]$, and is contained in $[-1, 1]^p$. For a unit coefficient vector v, the integral of $\{v^\top r_\ell(u)\}^2$ over any such rectangle is strictly positive. Otherwise a nonzero polynomial would vanish on its interior. The family of rectangle endpoints and the unit coefficient sphere are compact, and the integral is continuous in both. Its minimum is thus strictly positive. Its maximum is finite. Density bounds and the bounds on leaf volume transfer these two inequalities to the conditional population Gram matrix given membership in the leaf.

Each entry of $r_\ell r_\ell^\top$ is bounded on $[-1, 1]^p$. Conditional on the leaf count, the observations in that leaf are independent draws from the conditional covariate distribution. Hoeffding's inequality for

each entry, followed by a union bound over the fixed number of entries and T_n leaves, puts all empirical leaf Gram eigenvalues in a common interval $[c_2, C_2]$, with the failure probability in (28). One may convert entrywise error to operator error by multiplying its maximum by the fixed matrix dimension. Independence between different leaves or trees is not needed for this union bound. The forest Γ_n is an average of these leaf Gram matrices, so has the same bounds. The kernel case uses the one containing cube and the identical argument.

On this event, $\alpha_i \le C/(n_{\text{final}}h^p)$ and $\sum_i \alpha_i = 1$. Bounded r_ℓ and Γ_n^{-1} give

$$\max_i |\omega_i| \le C/(n_{\text{final}}h^p), \qquad \sum_i |\omega_i| \le C.$$

For any coordinate of r_ℓ, matrix multiplication in (9) gives $\sum_i \omega_i r_\ell((X_i - x_0)/h)^\top = \mathbf{e}_1^\top$. This proves normalization and all moment equalities. It follows that $q_n^2 \le \max_i |\omega_i| \sum_i |\omega_i| \le C/(n_{\text{final}}h^p)$. For the lower bound, the number N of observations in the containing cube is at most $Cn_{\text{final}}h^p$ on the same event. Cauchy-Schwarz and normalization give

$$1 = \left(\sum_{i:\omega_i \neq 0} \omega_i \right)^2 \le N \sum_i \omega_i^2.$$

Thus $q_n^2 \ge c/(n_{\text{final}}h^p)$. The supports lie in the containing cube by construction. Combining the finitely many failure bounds proves the lemma. □

B.2 Moment and polynomial-bias bounds

Lemma 10 (Conditional second and fourth moments)**.** *Given a sigma-field $\mathcal{F}$, let $Z_i \in \mathbb{R}^k$ be independent, centered, and bounded by a fixed constant. Let w_i be $\mathcal{F}$-measurable, and put $q^2 = \sum_i w_i^2$. For fixed k,*

$$\mathbb{E}\left(\left\| \sum_i w_i Z_i \right\|^2 \mid \mathcal{F} \right) \le Cq^2, \qquad \mathbb{E}\left(\left\| \sum_i w_i Z_i \right\|^4 \mid \mathcal{F} \right) \le Cq^4.$$

Proof. Cross terms in the second moment vanish by conditional independence and zero means. For a scalar coordinate Z_{ij}, expansion of the fourth power leaves only indices occurring at least twice:

$$\mathbb{E}\left\{ \left(\sum_i w_i Z_{ij} \right)^4 \,\middle|\, \mathcal{F} \right\} = \sum_i w_i^4 \mathbb{E}(Z_{ij}^4 \mid \mathcal{F}) + 6 \sum_{i<l} w_i^2 w_l^2 \mathbb{E}(Z_{ij}^2 \mid \mathcal{F}) \mathbb{E}(Z_{lj}^2 \mid \mathcal{F}).$$

Bounded fourth moments make this at most $C(\sum_i w_i^2)^2$. Finally, $\|z\|^4 = (\sum_j z_j^2)^2 \le k \sum_j z_j^4$. Summing over coordinates proves the vector inequality. Applying this to a concatenation of the two arm vectors retains their within-observation dependence. □

Lemma 11 (Polynomial bias)**.** *On $\mathcal{W}_n$,*

$$\left\| \sum_i \omega_i b_a(X_i) \right\| \le Ch_n^{s_\mu}, \qquad \|B_{\tau,n}\| \le Ch_n^{s_\tau}, \qquad \rho_{a,n} \le Ch_n^{\min(s_H, \ell+1)}.$$

Proof. Both $b_a(x_0)$ and $\vartheta_{x_0}(x_0)$ are zero by the mean equations. Expand either field in its Taylor polynomial of degree $\lceil s \rceil - 1$. Every polynomial term vanishes in the weighted sum by Assumption 4. The remainders have norm at most Ch_n^s at every supported covariate. Their weighted sum is bounded by $Ch_n^s \sum_i |\omega_i|$. This gives the first two bounds. For the Hessian field, the constant term is H_a. Subtract it. Use the Taylor polynomial through degree $\min(\ell, \lceil s_H \rceil - 1)$. If $s_H > \ell + 1$, bounded derivatives of order $\ell + 1$ give remainder $O(h_n^{\ell+1})$. Otherwise the assumed Hölder remainder gives $O(h_n^{s_H})$. Bounded total variation again proves the claim. □

C COMPLETE PROOF OF THE JOINT EXPANSION

Proof of Theorem 4. All steps up to the final unconditional bound condition on $\mathcal{F}_n$ and restrict to $\mathcal{A}_n$. Suppress the index n temporarily. The reference points, nuisance functions, Hessians, and weights are then fixed. Set

$$u_a = \mathrm{Log}_{m_a} \widetilde{m}_a, \quad \widetilde{H}_a^* = \mathsf{P}_{\widetilde{m}_a \to m_a} \widehat{H}_a \mathsf{P}_{m_a \to \widetilde{m}_a}, \quad Z_a = \sum_i \omega_i \xi_{i,a}, \quad B_a = H_a^{-1} \sum_i \omega_i b_a(X_i).$$

Here $\|u_a\| = r_a$. The centered sums satisfy

$$\mathbb{E}(\|Z_a\|^2 \mid \mathcal{F}_n) \le Cq^2, \qquad \mathbb{E}(\|Z_a\|^4 \mid \mathcal{F}_n) \le Cq^4 \tag{29}$$

by Lemma 10. Boundedness follows from overlap, compact log responses, bounded nuisance fits, and isometric transport.

Step 1: conditional mean at the random reference point. Equation (7) remains valid at the realized reference point because the reference point and fitted functions are fixed under this conditioning. Write

$$\delta_a(x') = \mathsf{P}_{\widetilde{m}_a \to m_a} \frac{\widehat{e}_a(x') - e_{0,a}(x')}{\widehat{e}_a(x')} \{\widehat{g}_a(\widetilde{m}_a, x') - g_{0,a}(\widetilde{m}_a, x')\}.$$

Then $\sup \|\delta_a\| \le D_a/\epsilon$. The population Taylor expansion (25) gives, observation by observation,

$$\mathsf{P}_{\widetilde{m}_a \to m_a} \mathbb{E}(\widehat{\psi}_{i,a} \mid \mathcal{F}_n) = b_a(X_i) - H_{0,a}(m_a, X_i) u_a + T_{i,a} + \delta_a(X_i), \quad \|T_{i,a}\| \le C r_a^2.$$

Consequently the transported weighted moment S_a equals

$$S_a := \mathsf{P}_{\widetilde{m}_a \to m_a} \sum_i \omega_i \widehat{\psi}_{i,a} = H_a B_a - H_{a,\mathrm{loc}} u_a + Z_a + d_a + T_a, \tag{30}$$

where $H_{a,\mathrm{loc}} = \sum_i \omega_i H_{0,a}(m_a, X_i)$, $\|d_a\| \le C D_a$, and $\|T_a\| \le C r_a^2$. These last bounds use signed-weight total variation, not positivity of the corrected weights.

Step 2: cancellation of reference-point displacement. Transport the update vector to $T_{m_a}\mathcal{M}$, obtaining $t_a = (\widetilde{H}_a^*)^{-1} S_a$. Spectral clipping and the event $E_a^H \le \lambda/2$ imply

$$\left\|(\widetilde{H}_a^*)^{-1}\right\|_{\mathrm{op}} \le 2/\lambda, \qquad \left\|(\widetilde{H}_a^*)^{-1} - H_a^{-1}\right\|_{\mathrm{op}} \le 2E_a^H/\lambda^2.$$

The second identity follows by writing the inverse difference as $(\widetilde{H}_a^*)^{-1}(H_a - \widetilde{H}_a^*)H_a^{-1}$. Adding u_a to (30) after multiplying by $(\widetilde{H}_a^*)^{-1}$ yields

$$\begin{aligned} u_a + t_a - (B_a + H_a^{-1} Z_a) = \{I - (\widetilde{H}_a^*)^{-1} H_{a,\mathrm{loc}}\} u_a + \{(\widetilde{H}_a^*)^{-1} H_a - I\} B_a \\ + \{(\widetilde{H}_a^*)^{-1} - H_a^{-1}\} Z_a + (\widetilde{H}_a^*)^{-1}(d_a + T_a). \end{aligned} \tag{31}$$

Since $\|H_{a,\mathrm{loc}} - H_a\|_{\mathrm{op}} = \rho_a$, Lemma 11 and the inverse bounds imply

$$\left\|u_a + t_a - (B_a + H_a^{-1} Z_a)\right\| \le C\{D_a + r_a^2 + (E_a^H + \rho_a) r_a + E_a^H (h^{s_\mu} + \|Z_a\|)\}. \tag{32}$$

Also, directly from (30) and boundedness of $H_{a,\mathrm{loc}}$,

$$\|t_a\| \le C(r_a + h^{s_\mu} + D_a + \|Z_a\|). \tag{33}$$

Step 3: endpoint coordinates and the joint logarithm. There are fixed positive constants h_0, r_0, d_0, z_0 such that, whenever their corresponding upper bounds hold and $\max_a \|Z_a\| \le z_0$, both update paths remain in the neighborhoods of Lemma 7, and both updated endpoints belong to $\mathcal{K}$. To see this, use the positive distance of $\mathcal{K}_0$ to $\mathcal{M} \setminus \mathrm{int}\,\mathcal{K}$, the uniform chart radius in that lemma, and (33), then choose the four constants successively small. Thus the safeguard is inactive on the event $\mathcal{E} = \{\max_a \|Z_a\| \le z_0\}$.

Let $s_a = \mathrm{Log}_{m_a} \widehat{\mu}_a$. Equations (26), (32), and (33) give on $\mathcal{E}$

$$\left\|s_a - B_a - H_a^{-1} Z_a\right\| \le C\{D_a + r_a^2 + (E_a^H + \rho_a) r_a + E_a^H (h^{s_\mu} + \|Z_a\|)$$

$$+ h^{2s_\mu} + \|Z_a\|^2\}. \tag{34}$$

Here $D_a^2 \le CD_a$ because D_a is uniformly bounded. The same local composition bound also gives $\|s_a\| \le C(r_a + h^{s_\mu} + D_a + \|Z_a\|)$. Define $R_{\tau,n}$ by the exact identity (17). Applying (27), then using this last bound and (34), proves on $\mathcal{E}$

$$\|R_{\tau,n}\| \le C \sum_a \{D_a + r_a^2 + (E_a^H + \rho_a) r_a + E_a^H (h^{s_\mu} + \|Z_a\|) + h^{2s_\mu} + \|Z_a\|^2\}. \tag{35}$$

Linearity gives $\mathcal{J}(B_0, B_1) = B_{\tau,n}$ and $\mathcal{J}(H_0^{-1} Z_0, H_1^{-1} Z_1) = \sum_i \omega_i v_{i,n}$. The definition of $R_{\tau,n}$ also applies outside $\mathcal{E}$.

Step 4: moments and safeguards. On $\mathcal{E}$, the conditional L^2 norm of the right side of (35) is at most $C(\mathcal{R}_n + q^2 + h^{2s_\mu})$, by (29) and the triangle inequality in L^2. For example, the random terms $E_a^H \|Z_a\|$ and $\|Z_a\|^2$ have conditional L^2 norms at most $CE_a^H q$ and Cq^2, respectively.

On $\mathcal{E}^c$, the reported endpoints still lie in $\mathcal{K}$, because an inadmissible endpoint is replaced by its reference point. Thus the reported effect, the true effect, and their transported difference are uniformly bounded. The bias and the linear moment sum are also uniformly bounded, by bounded moment values and $\sum_i |\omega_i| \le C_w$. Hence $\|R_{\tau,n}\| \le C$ everywhere on this exceptional event. Markov's inequality using fourth moments gives

$$\mathbb{P}(\mathcal{E}^c \mid \mathcal{F}_n) \le \sum_a z_0^{-4} \mathbb{E}(\|Z_a\|^4 \mid \mathcal{F}_n) \le Cq^4.$$

Its contribution to the conditional L^2 norm of the remainder is at most Cq^2. This proves (18) for the safeguarded estimator.

Finally, the bias is bounded by Ch^{s_τ}, and the centered joint sum has conditional second moment at most Cq^2. Squaring the triangle bound and using $q^4 \le Cq^2$, which follows from bounded weight total variation, proves (19). Taking expectations over $\mathcal{A}_n$ gives its stated unconditional contribution. On $\mathcal{A}_n^c$, the compact-valued estimator has bounded transported error, so the remaining contribution is at most $C\mathbb{P}(\mathcal{A}_n^c)$. □

Proof of Corollary 5. Conditional Markov's inequality applied to (19), together with $\mathbb{P}(\mathcal{A}_n^c) \to 0$, gives the asserted stochastic order. Formally, along any subsequence on which the normalized first-stage error is tight, restrict to an event where it is bounded by a fixed constant. On that event the conditional tail at M times the displayed rate is at most C/M^2. First take $M \to \infty$, then let the probability of the restriction's complement tend to zero. The weight bounds replace q_n by $(nh_n^p)^{-1/2}$. For $s = \min(s_\tau, 2s_\mu)$, balancing this term with h_n^s gives $h_n \asymp n^{-1/(2s+p)}$ and the claimed rate. Under the additional expectation and failure-probability conditions, integrate Theorem 4. Each contribution is $O(n^{-2s/(2s+p)})$. □

D ARM-WISE FRÉCHET COMPARISON

This appendix gives the full arm-wise comparison summarized in the main text, including the population bias decomposition, empirical rate, and proofs.

The smoothness of ϑ_{x_0} in Corollary 5 is partly a property of the target itself: a separate plug-in estimator can also inherit first-order cancellation. We now state exactly when this occurs and what separate arm-wise smoothing adds. Decompose the linear map in (11) as

$$\mathcal{J}(u_0, u_1) = J_0 u_0 + J_1 u_1,$$

and define, for $a \in \{0, 1\}$,

$$f_a(x) = J_a H_a^{-1} b_a(x). \tag{36}$$

Then $\vartheta_{x_0}(x) = f_0(x) + f_1(x)$.

For the comparison, use the same local-polynomial degree ℓ and the box kernel $K(u) = \mathbf{1}(\|u\|_\infty \le 1)$. Put $U_h = (X - x_0)/h$ and, for a vector-valued function f, define the population arm-wise local-polynomial intercept operator

$$\Gamma^F_{a,h} = \mathbb{E}\{K(U_h) r_\ell(U_h) r_\ell(U_h)^\top \mid A = a\},$$
$$\mathcal{L}_{a,h} f = \mathbf{e}_1^\top (\Gamma^F_{a,h})^{-1} \mathbb{E}\{K(U_h) r_\ell(U_h) f(X) \mid A = a\}. \tag{37}$$

For sufficiently small h, the Gram matrices are nonsingular by overlap and the local density bounds. Each $\mathcal{L}_{a,h}$ reproduces polynomials through degree ℓ. Let $m^F_{a,h}$ denote the solution near m_a of the population arm-wise local Fréchet score equation

$$\mathcal{L}_{a,h}\{g_{0,a}(m, \cdot)\} = 0. \tag{38}$$

This is the first-order condition for the corresponding population local-polynomial Fréchet fit whenever the local minimizer is interior.

Proposition 12 (Arm-wise Fréchet localization)**.** *Under Assumptions 1 through 3, let* $\bar{s}_H = \min(s_H, \ell + 1)$. *For all sufficiently small h, the local roots in (38) are uniquely defined near* m_a *and*

$$\begin{aligned}\Phi(m^F_{0,h}, m^F_{1,h}) - \tau(x_0) &= \mathcal{L}_{0,h} f_0 + \mathcal{L}_{1,h} f_1 + R_F(h) \\ &= \mathcal{L}_{0,h} \vartheta_{x_0} + (\mathcal{L}_{1,h} - \mathcal{L}_{0,h}) f_1 + R_F(h),\end{aligned} \tag{39}$$

where

$$\|R_F(h)\| \le C\{h^{2s_\mu} + h^{s_\mu + \bar{s}_H}\}. \tag{40}$$

If, in addition, the true propensity $e_{0,1}$ *is locally Hölder of order* $s_e > 0$, *and* $\bar{s}_e = \min(s_e, 1)$, *then*

$$\|(\mathcal{L}_{1,h} - \mathcal{L}_{0,h}) f_1\| \le C h^{s_\mu + \bar{s}_e}, \tag{41}$$

and therefore

$$\left\|\Phi(m^F_{0,h}, m^F_{1,h}) - \tau(x_0)\right\| \le C\{h^{s_\tau} + h^{s_\mu + \bar{s}_e} + h^{2s_\mu} + h^{s_\mu + \bar{s}_H}\}. \tag{42}$$

Let $\widehat{\mu}^F_{a,h}$ be the empirical arm-wise local-polynomial Fréchet root obtained from the observations with $A_i = a$, and put $\widehat{\tau}_F = \mathrm{Log}_{\widehat{\mu}^F_{0,h}} \widehat{\mu}^F_{1,h}$.

Corollary 13 (Rate comparison)**.** *Under Proposition 12, if* $nh^p / \log n \to \infty$ *and the empirical local roots remain in the stated normal neighborhood with probability tending to one, then*

$$\left\| \mathsf{P}_{\widehat{\mu}^F_{0,h} \to m_0} \widehat{\tau}_F - \tau(x_0) \right\| = O_p\{(nh^p)^{-1/2} + h^{s_\tau} + h^{s_\mu + \bar{s}_e} + h^{2s_\mu} + h^{s_\mu + \bar{s}_H}\}. \tag{43}$$

In particular, if $\bar{s}_H \ge s_\mu$ *and* $s_\mu + \bar{s}_e \ge s_\tau$, *separate arm-wise Fréchet regression has no additional deterministic order beyond the GO terms* $h^{s_\tau} + h^{2s_\mu}$. *If* $s_\mu + \bar{s}_e < s_\tau$, *the additional arm-wise term can instead determine the plug-in bias. Proposition 18 gives a noncommuting positive-definite model in which this term has exact order* $h^{s_\mu + \bar{s}_e}$ *and changes the optimized rate.*

The decomposition separates the cancellation in $f_0 + f_1 = \vartheta_{x_0}$ from the additional arm-specific design term. The first component can also benefit a plug-in estimator. GO-learner uses Neyman-orthogonal treatment-specific mean moments and the same pooled final-covariate weights for both arms, while separate arm-wise smoothing can retain the second term in (39) because $X \mid A = 0$ and $X \mid A = 1$ generally induce different local design distributions. Proposition 18 gives a noncommuting positive-definite construction in which this additional term changes the optimized rate.

Proof of Proposition 12. The conditional density of $X \mid A = a$ is bounded above and away from zero near x_0 by Assumption 1. After the rescaling $u = (x - x_0)/h$, the population Gram matrices in (37) therefore have eigenvalues bounded above and away from zero uniformly for sufficiently small h, by the same polynomial-compactness argument used in Lemma 9. Hence $\mathcal{L}_{a,h}$ has uniformly bounded operator norm and reproduces every polynomial through degree ℓ.

Write $u_{a,h} = \mathrm{Log}_{m_a} m^F_{a,h}$. Applying (25) at m_a and then $\mathcal{L}_{a,h}$ gives

$$0 = \mathcal{L}_{a,h} b_a - \mathcal{L}_{a,h}\{H_{0,a}(m_a, \cdot)\} u_{a,h} + R_{a,h}, \qquad \|R_{a,h}\| \le C \|u_{a,h}\|^2. \tag{44}$$

Polynomial reproduction and Assumption 3 imply

$$\|\mathcal{L}_{a,h} b_a\| \le C h^{s_\mu}, \qquad \|\mathcal{L}_{a,h}\{H_{0,a}(m_a,\cdot)\} - H_a\|_{\mathrm{op}} \le C h^{\bar{s}_H}. \tag{45}$$

The derivative of the left side of (44) with respect to u at zero is therefore a perturbation of $-H_a$ of size $O(h^{\bar{s}_H})$. Since H_a has eigenvalues bounded below by λ, the inverse function theorem gives a unique root near zero for all sufficiently small h. Equation (44) then first gives $\|u_{a,h}\| \le C h^{s_\mu}$ and, after adding and subtracting $H_a u_{a,h}$,

$$\left\|u_{a,h} - H_a^{-1}\mathcal{L}_{a,h} b_a\right\| \le C\{h^{2s_\mu} + h^{s_\mu+\bar{s}_H}\}. \tag{46}$$

Applying the joint logarithm expansion (27) to $(u_{0,h}, u_{1,h})$, using linearity of J_0, J_1, proves the first line of (39) and (40). The second line is the identity

$$\mathcal{L}_{0,h} f_0 + \mathcal{L}_{1,h} f_1 = \mathcal{L}_{0,h}(f_0 + f_1) + (\mathcal{L}_{1,h} - \mathcal{L}_{0,h}) f_1.$$

It remains to bound the operator difference. Let $p_a(x)$ denote the conditional density of $X \mid A = a$. On the local cube,

$$\frac{p_1(x)}{p_0(x)} = \frac{\mathbb{P}(A=0)}{\mathbb{P}(A=1)} \frac{e_{0,1}(x)}{e_{0,0}(x)}.$$

Overlap and local Hölder continuity of order s_e imply that this ratio equals a positive constant plus $O(h^{\bar{s}_e})$ uniformly on the cube. Multiplying a design density by a positive constant leaves the intercept operator unchanged. After this rescaling, the two arm-specific design measures differ pointwise by a relative $O(h^{\bar{s}_e})$ factor. Let $P_1(x)$ be the Taylor polynomial of f_1 at x_0 through degree $\lceil s_\mu \rceil - 1$. Because both operators reproduce P_1,

$$(\mathcal{L}_{1,h} - \mathcal{L}_{0,h}) f_1 = (\mathcal{L}_{1,h} - \mathcal{L}_{0,h})(f_1 - P_1).$$

The remainder $f_1 - P_1$ is uniformly $O(h^{s_\mu})$. The numerator moments in (37) for this remainder differ by $O(h^{s_\mu+\bar{s}_e})$, while the two rescaled Gram matrices differ by $O(h^{\bar{s}_e})$ and have uniformly bounded inverses. The inverse identity $G_1^{-1} - G_0^{-1} = G_1^{-1}(G_0 - G_1)G_0^{-1}$ therefore gives (41). Finally $\mathcal{L}_{0,h}\vartheta_{x_0} = O(h^{s_\tau})$ by polynomial reproduction and $\vartheta_{x_0}(x_0) = 0$, proving (42). □

Proof of Corollary 13. Within each treatment arm, overlap and the local density bounds imply that the number of observations in the h-cube is of order nh^p with probability tending to one. The empirical arm-specific local-polynomial Gram matrices satisfy the same eigenvalue and total-variation bounds as Lemma 9, now under the conditional design $X \mid A = a$. At the population root $m_{a,h}^F$, the empirical centered log-score has norm $O_p\{(nh^p)^{-1/2}\}$ by bounded conditional second moments. Bounded log and Hessian terms, fixed polynomial degree, and the local Gram bounds give uniform convergence of the empirical score and derivative on a fixed normal neighborhood of $m_{a,h}^F$. The population derivative is uniformly nonsingular there, so with probability tending to one an empirical root exists in that neighborhood and the usual local root expansion gives

$$d(\widehat{\mu}_{a,h}^F, m_{a,h}^F) = O_p\{(nh^p)^{-1/2}\}.$$

Smoothness of the transported logarithm map on the common compact normal neighborhood converts the two endpoint errors into the same stochastic order for the reported treatment effect. Combining this with (42) proves (43). The two stated comparisons are immediate after noting that $h^{s_\mu+\bar{s}_H} \le h^{2s_\mu}$ when $\bar{s}_H \ge s_\mu$. □

E COVARIANCE CONSISTENCY AND POINTWISE INFERENCE

Proof of Theorem 6. We first work in fixed orthonormal frames at m_0, m_1, transporting all reference-point quantities to these spaces. All matrix comparisons below therefore compare operators on the same space. Every statement is restricted to $\mathcal{A}_n$, whose complement has vanishing probability.

Step 1: fitted linear maps. Let $\widehat{L}_n : T_{m_0}\mathcal{M} \times T_{m_1}\mathcal{M} \to T_{m_0}\mathcal{M}$ be

$$\widehat{L}_n(z_0, z_1) = \mathsf{P}_{\widetilde{m}_0 \to m_0} \widetilde{\mathcal{J}}\left(\widehat{H}_0^{-1}\mathsf{P}_{m_0\to\widetilde{m}_0} z_0, \widehat{H}_1^{-1}\mathsf{P}_{m_1\to\widetilde{m}_1} z_1\right),$$

and let $L(z_0, z_1) = \mathcal{J}(H_0^{-1} z_0, H_1^{-1} z_1)$. The differential and transports vary smoothly on the common compact neighborhoods. Together with the inverse difference identity used in (31), this gives

$$\left\| \widehat{L}_n - L \right\|_{\mathrm{op}} \le C \sum_a (r_{a,n} + E^H_{a,n}) = o_p(1). \tag{47}$$

Both maps are uniformly bounded. Write $W_i = \mathsf{P}_{\widetilde{m}_0 \to m_0} U_i$ and $m_i^U = \mathbb{E}(W_i \mid \mathcal{F}_n)$. The centered vector $W_i - m_i^U$ is exactly $\widehat{L}_n(\xi_{i,0}, \xi_{i,1})$. Bounded moment covariances and (47) imply

$$\left\| \sum_i \beta_i \operatorname{Cov}(W_i \mid \mathcal{F}_n) - \Sigma_n \right\|_{\mathrm{op}} = o_p(1). \tag{48}$$

Indeed, for a covariance matrix C_i of the concatenated moment vector, the difference is $(\widehat{L}_n - L) C_i \widehat{L}_n^* + L C_i (\widehat{L}_n^* - L^*)$. Its norm is at most a constant times $\left\| \widehat{L}_n - L \right\|$, and $\sum_i \beta_i = 1$.

Step 2: conditional moment means. By (25), (7), and $b_a(x_0) = 0$,

$$\max_{i:\omega_i \neq 0} \left\| \mathsf{P}_{\widetilde{m}_a \to m_a} \mathbb{E}(\widehat{\psi}_{i,a} \mid \mathcal{F}_n) \right\| \le C\{ r_{a,n} + h_n^{\min(s_\mu, 1)} + D_{a,n} \} = o_p(1).$$

The exponent here is $\min(s_\mu, 1)$, because this is a pointwise bound before polynomial cancellation. It follows that $\max_i \left\| m_i^U \right\| = o_p(1)$. Thus

$$\left\| \sum_i \beta_i m_i^U \otimes m_i^U \right\|_{\mathrm{op}} \le \max_i \left\| m_i^U \right\|^2 = o_p(1). \tag{49}$$

This replaces a fixed-limit nuisance smoothness or covariance localization assumption.

Step 3: weighted empirical covariance. The variables W_i are independent conditional on $\mathcal{F}_n$, and are uniformly bounded. Moreover,

$$\sum_i \beta_i^2 \le \max_i \beta_i = \left(\frac{\max_i |\omega_i|}{q_n} \right)^2 = o_p(1).$$

For each matrix entry, the conditional variance of $\sum_i \beta_i \{ W_i \otimes W_i - \mathbb{E}(W_i \otimes W_i \mid \mathcal{F}_n) \}$ is bounded by $C \sum_i \beta_i^2$. Summing over the fixed number of entries gives the same bound for its expected squared Frobenius norm. Conditional Markov's inequality, then averaging its truncated upper bound over $\mathcal{F}_n$, proves that this matrix is $o_p(1)$. The same argument yields

$$\sum_i \beta_i (W_i - m_i^U) = o_p(1), \qquad \overline{W} := \sum_i \beta_i W_i = o_p(1),$$

where the second equality also uses $\max_i \left\| m_i^U \right\| \to_p 0$. Now expand the covariance exactly:

$$\begin{aligned} \sum_i \beta_i (W_i - \overline{W}) \otimes (W_i - \overline{W}) = & \sum_i \beta_i \operatorname{Cov}(W_i \mid \mathcal{F}_n) + \sum_i \beta_i m_i^U \otimes m_i^U \\ & + \sum_i \beta_i \{ W_i \otimes W_i - \mathbb{E}(W_i \otimes W_i \mid \mathcal{F}_n) \} - \overline{W} \otimes \overline{W}. \end{aligned}$$

Equations (48) through (49) and the just-established fluctuation bounds prove that this matrix differs from Σ_n by $o_p(1)$.

The left side is $\widetilde{\Sigma}_n$ transported directly from $\widetilde{m}_0$ to m_0. The reported covariance first transports from $\widetilde{m}_0$ to $\widehat{\mu}_0$ and then, for comparison, from $\widehat{\mu}_0$ to m_0. Both $\widetilde{m}_0$ and $\widehat{\mu}_0$ converge to m_0: the former by assumption and the latter by the endpoint bounds in the proof of Theorem 4, since $h_n, q_n, D_{a,n}, r_{a,n}, E^H_{a,n} \to_p 0$. Smoothness of parallel transport implies that the two compositions differ in operator norm by $o_p(1)$. Since $\widetilde{\Sigma}_n$ is bounded, this additional covariance error is $o_p(1)$. This proves the claimed covariance consistency, including its random-anchor comparison.

Step 4: conditional normal limit. On the event $\lambda_{\min}(\Sigma_n) \geq c$, put

$$T_n = \sum_i \frac{\omega_i}{q_n} \Sigma_n^{-1/2} v_{i,n}.$$

It is conditionally centered and has conditional covariance $I_{d_{\mathcal{M}}}$. For fixed $t \in \mathbb{R}^{d_{\mathcal{M}}}$, set $V_i = (\omega_i/q_n) t^\top \Sigma_n^{-1/2} v_{i,n}$. Its conditional mean is zero, $\sum_i \mathbb{E}(V_i^2 \mid \mathcal{F}_n) = \|t\|^2$, and

$$\sum_i \mathbb{E}(|V_i|^3 \mid \mathcal{F}_n) \leq C_t \frac{\max_i |\omega_i|}{q_n} \sum_i \frac{\omega_i^2}{q_n^2} = o_p(1).$$

The elementary expansion $|e^{iz} - 1 - iz + z^2/2| \leq |z|^3/6$ yields

$$\mathbb{E}(e^{iV_i} \mid \mathcal{F}_n) = 1 - \tfrac{1}{2}\mathbb{E}(V_i^2 \mid \mathcal{F}_n) + r_i, \qquad \sum_i |r_i| = o_p(1).$$

Each variance is at most $C_t(\max_i |\omega_i|/q_n)^2$, so its maximum vanishes. Taking the product of the characteristic functions and using $\log(1+z) = z + O(|z|^2)$ proves convergence to $\exp(-\|t\|^2/2)$. The sum of the squared logarithm arguments vanishes because their maximum vanishes and their absolute sum is bounded. This establishes the scalar normal limit for every t, hence the multivariate limit.

For precision when $\mathcal{F}_n$ is random, any subsequence has a further subsequence on which the stated $o_p(1)$ quantities vanish almost surely and the covariance lower bound eventually holds. Conditional on almost every such auxiliary/covariate sequence, the preceding characteristic-function argument is deterministic. Tightness follows from $\mathbb{E}(\|T_n\|^2 \mid \mathcal{F}_n) = d_{\mathcal{M}}$. It therefore gives weak convergence of the conditional laws. The subsequence characterization returns convergence in probability of their integrals against bounded Lipschitz test functions. These are the usual finite-dimensional weak-convergence criteria. See van der Vaart & Wellner (1996).

Step 5: studentization and coverage. Theorem 4 and (23) give

$$q_n^{-1} \Sigma_n^{-1/2} \{\mathsf{P}_{\widehat{\mu}_0 \to m_0} \widehat{\tau} - \tau(x_0)\} = T_n + o_p(1).$$

For the remainder, its conditional L^2 bound divided by q_n tends to zero in probability. Conditional Markov's inequality, truncated at one and integrated, gives the asserted unconditional $o_p(1)$. The bias is treated by Lemma 11. The covariance eigenvalues are bounded above by bounded moment values and below by assumption. The proved operator-norm consistency therefore also gives consistency of the corresponding inverse and inverse square root. Substitution in the displayed quadratic form yields a $\chi^2_{d_{\mathcal{M}}}$ limit. Isometric transport preserves inner products, so that form is exactly the one in (24) evaluated at $\mathsf{P}_{m_0 \to \widehat{\mu}_0} \tau(x_0)$. Continuity of the $\chi^2_{d_{\mathcal{M}}}$ distribution proves coverage. Its smallest empirical covariance eigenvalue is eventually positive with probability tending to one, so the specified inference-failure events have vanishing probability under these assumptions. □

F Concrete first stages and a noncommuting positive-definite model

The main theorems allow any first stages satisfying their displayed error conditions. This section verifies those conditions for bounded local regressions. The additional conditions here define a sufficient model class. They are not imposed on every implementation of the main estimator. Let each block size be between fixed positive multiples of n. For uniform regression bounds, assume the covariate density is bounded above and below on $[0,1]^p$, and overlap holds there. Outcome support and the geometry bounds are uniform on this domain. Whenever a supremum over candidate points occurs below, it is a property of the true regression functions. The fitted regressions are evaluated only at independent realized reference points.

F.1 Positive-weight reference points

Lemma 14 (A reference point from local Fréchet regression)**.** *Suppose $m \mapsto d^2(m,y)/2$ is geodesically λ_{sc}-strongly convex on $\mathcal{K}$ for every supported outcome, with fixed $\lambda_{\mathrm{sc}} > 0$. Let $\widetilde{m}_a$ minimize*

its empirical average over the observations in I_{ref} *with* $A_i = a$ *and* $\|X_i - x_0\|_\infty \le b_p$*, restricting the minimization to* $\mathcal{K}$*. On an empty cell, return a fixed point of* $\mathcal{K}$*. Put* $t_\mu = \min(s_\mu, 1)$*. If* $n_{\mathrm{ref}} b_p^p \ge C \log(1/\eta)$*, then, with probability at least* $1 - \eta$*,*

$$\max_a d(\widetilde{m}_a, m_a) \le C \left\{ b_p^{t_\mu} + \sqrt{\frac{\log(C/\eta)}{n_{\mathrm{ref}} b_p^p}} \right\}. \tag{50}$$

Proof. The arm-specific cell probability is at least a constant times b_p^p, by overlap and the density lower bound. A Chernoff bound makes its count at least $cn_{\mathrm{ref}} b_p^p$, except with probability $C \exp(-cn_{\mathrm{ref}} b_p^p)$. Conditional on the selected covariates and treatments, the empirical gradient at m_a is minus the average of independent bounded log responses. Its conditional mean has norm at most $Cb_p^{t_\mu}$, since each response mean is $b_a(X_i)$, $b_a(x_0) = 0$, and bounded local Hölder derivatives give this pointwise bound. Coordinatewise Hoeffding bounds in a fixed orthonormal frame, followed by a union bound over the two arms and $d_{\mathcal{M}}$ coordinates, bound the centered gradient norm by the second term in (50).

Let F_a be the selected empirical objective. Its strong convexity along the geodesic from m_a to $\widetilde{m}_a$ implies

$$F_a(\widetilde{m}_a) \ge F_a(m_a) + \left\langle \nabla F_a(m_a), \mathrm{Log}_{m_a} \widetilde{m}_a \right\rangle + \tfrac{1}{2} \lambda_{\mathrm{sc}} d^2(\widetilde{m}_a, m_a).$$

Since $F_a(\widetilde{m}_a) \le F_a(m_a)$, Cauchy-Schwarz gives $d(\widetilde{m}_a, m_a) \le 2 \|\nabla F_a(m_a)\| / \lambda_{\mathrm{sc}}$, including the case of a constrained minimizer. Compactness ensures existence, and strong convexity ensures uniqueness. Combining with the gradient and count bounds proves the lemma. □

On the affine-invariant positive-definite manifold one can take $\lambda_{\mathrm{sc}} = 1$. This manifold is complete, simply connected, and has nonpositive sectional curvature (Pennec et al., 2006; Bhatia, 2007). The relevant convexity implication follows from the second variation of energy. Along the geodesic from a fixed y to m, the index form for a variation with endpoint velocity v is $\int_0^1 \{\|\nabla_t J\|^2 - \langle R(J, \dot\gamma)\dot\gamma, J\rangle\}\, dt$, with $J(0) = 0, J(1) = v$. The curvature term is nonpositive inside the inner product, and transport plus Cauchy-Schwarz gives $\int_0^1 \|\nabla_t J\|^2\, dt \ge \|v\|^2$. Thus the squared-distance Hessian is bounded below by the identity. The no-cut-locus property makes the argument global.

F.2 Bounded piecewise-polynomial regressions

Partition $[0,1]^p$ into cubes of side $b = 1/k$, for an integer k. Within each cube, express the predictor in rescaled monomials through degree $\lceil s \rceil - 1$. Fit least squares separately in each cube. The number of monomials is fixed when p, s are fixed. Fit vector or operator responses by summing squared coordinate errors in an orthonormal frame. An empty or poorly conditioned cube returns zero before the specified norm projection. For propensity regression, the response is $\mathbf{1}(A = 1)$, fitted without arm restriction. For the other regressions, fit within each treatment arm. The covariate density conditional on either arm is still bounded above and below by overlap.

Lemma 15 (Uniform regression error at independent anchors)**.** *Suppose a bounded scalar response has a conditional mean in a Hölder ball of order* $s > 0$ *and fixed radius, and a covariate density bounded above and away from zero. For the piecewise-polynomial estimate just defined, projected onto an interval containing the true range, with probability at least* $1 - \eta$*,*

$$\left\| \widehat{f} - f \right\|_\infty \le C \left\{ b^s + \sqrt{\frac{\log(C/\eta) + p \log(1/b)}{n_{\mathrm{reg}} b^p}} \right\}, \tag{51}$$

provided $n_{\mathrm{reg}} b^p \ge C\{\log(C/\eta) + p \log(1/b)\}$*. The same rate holds for fixed-dimensional vector and self-adjoint operator responses, using norm or spectral projections whose constraint sets contain the true conditional means. It holds conditionally on a fixed number of independent random anchors, with the same constants, if the response bounds and true-function Hölder bounds are uniform over the anchors.*

Proof. Let r be the rescaled monomial vector and K its fixed dimension. In a cube C_b, choose the Taylor polynomial $r^\top \theta_b$ for the true mean, with uniform approximation error at most Cb^s. Define

$$G_b = (n_{\mathrm{reg}} b^p)^{-1} \sum_i \mathbf{1}(X_i \in C_b) r_i r_i^\top .$$

The eigenvalues of $\mathbb{E}G_b$ are bounded above and away from zero: for any nonzero coefficient vector the square of its polynomial has a positive integral on the unit cube, and compactness of the unit coefficient sphere gives a common positive lower bound. The density bounds preserve it. Every summand entry is bounded, with second moment at most Cb^p before normalization. Scalar Bernstein's inequality therefore implies, for each entry,

$$\mathbb{P}\{|(G_b - \mathbb{E}G_b)_{jl}| > t\} \le 2 \exp\left\{ -\frac{c n_{\mathrm{reg}} b^p t^2}{1+t} \right\}.$$

A union bound over $K^2 b^{-p}$ entries and cubes, followed by $\|B\|_{\mathrm{op}} \le K \max_{jl} |B_{jl}|$, puts all Gram eigenvalues between two fixed positive constants under the stated count condition. The same concentration bound, with the constant coordinate, controls every normalized cube count.

Write the response as $V_i = f(X_i) + \varepsilon_i$, with $\mathbb{E}(\varepsilon_i \mid X_i) = 0$. On the Gram event, the least-squares normal equations give exactly

$$\widehat{\theta}_b - \theta_b = G_b^{-1} \left[(n_{\mathrm{reg}} b^p)^{-1} \sum_i \mathbf{1}_{C_b}(X_i) r_i \varepsilon_i + (n_{\mathrm{reg}} b^p)^{-1} \sum_i \mathbf{1}_{C_b}(X_i) r_i \{ f(X_i) - r_i^\top \theta_b \} \right].$$

The second bracketed term is bounded coordinatewise by Cb^s, using the count bound and bounded monomials. Each coordinate of the first term is a sum of independent, bounded, mean-zero variables, with total variance at most $Cn_{\mathrm{reg}} b^p$ before normalization. Bernstein's inequality and a union bound over cubes and coordinates bound it by $C\sqrt{\{\log(C/\eta) + p\log(1/b)\}/(n_{\mathrm{reg}} b^p)}$. The smaller linear Bernstein term is absorbed because the denominator exceeds a constant times the logarithm. Bounded G_b^{-1} gives the same coefficient error. Multiplication by the bounded $r(x)$ converts it to the uniform prediction error in (51). The projection cannot increase distance to a true value inside the interval.

For vector responses, apply the scalar argument to each of the finitely many coordinates and project onto a closed norm ball containing the true vector. Euclidean projection is nonexpansive. For self-adjoint matrices, symmetrization is an orthogonal projection in Frobenius norm. Projection onto a fixed convex spectral interval is also a Frobenius projection, so does not increase that error. Operator norm is at most Frobenius norm. These facts give the same rate with a dimension-dependent constant. Conditional on independent anchors, the response maps and their frames are fixed, the observations used for regression remain independent, and all constants remain the stated uniform ones. Apply the preceding proof under that conditioning and integrate its failure bound. This proves the random-anchor assertion without requiring a fitted field over a continuum of candidates. □

The elementary Bernstein and Hoeffding bounds used here are standard bounded-variable concentration inequalities. See Wainwright (2019). The polynomial approximation and reproduction argument is the usual local-polynomial construction (Fan & Gijbels, 1996). The proofs above state the normalizations, Gram bounds, and conversions needed in this application explicitly.

F.3 Hessian calibration and sufficient rates

Let $\varepsilon_{e,n}, \varepsilon_{g,n}, \varepsilon_{h,n}$ bound the uniform errors of the three fitted nuisance functions at the reference points. Let s_g and s_h denote Hölder orders in x of $g_{0,a}(m,x)$ and $H_{0,a}(m,x)$, uniformly for $m \in \mathcal{K}$. These are first-stage regression smoothness conditions, while s_H in Assumption 3 concerns $x \mapsto H_{0,a}(m_a, x)$ at the true endpoint. Use kernel local-polynomial weights of degree at least $\lceil s_h \rceil - 1$ and bandwidth b_H on I_{cal}. A failed calibration Gram matrix returns $\lambda I_{d_{\mathcal{M}}}$ before clipping. Denote

$$v_{H,n}(\eta) = \sqrt{\frac{\log(C/\eta)}{n_{\mathrm{cal}} b_H^p}}.$$

For fixed degrees and positive local density, provided $n_{\mathrm{cal}} b_H^p \geq C \log(C/\eta)$, the following bound holds jointly with the regression and reference-point events:

$$E_{a,n}^H \leq C\{r_{a,n} + b_H^{s_h} + v_{H,n}(\eta) + \varepsilon_{e,n}\varepsilon_{h,n}\} \quad \text{with an additional failure probability at most } \eta. \tag{52}$$

To verify it, condition on $I_{\mathrm{ref}}, I_{\mathrm{reg}}$ and the calibration covariates. Lemma 9 bounds the calibration weights. The conditional mean of $\widehat{\chi}_{i,a}$ equals $H_{0,a}(\widetilde{m}_a, X_i)$ plus the product error in (8). Its local average differs from $H_{0,a}(\widetilde{m}_a, x_0)$ by at most $C(b_H^{s_h} + \varepsilon_e \varepsilon_h)$, using polynomial reproduction and bounded total variation. The centered operator responses are conditionally independent and bounded. A weighted scalar Bernstein inequality for each coordinate, with $\sum w_i^2 \leq C/(n_{\mathrm{cal}} b_H^p)$ and $\max |w_i| \leq C/(n_{\mathrm{cal}} b_H^p)$, bounds their sum by $C v_{H,n}(\eta)$. Its linear logarithmic term is again smaller. Symmetrization and eigenvalue projection cannot increase Frobenius distance to the transported H_a, which is inside $[\lambda/2, 2\Lambda]$. Finally the candidate-point derivative bound gives $\|\mathsf{P}_{\widetilde{m}_a \to m_a} H_{0,a}(\widetilde{m}_a, x_0) \mathsf{P}_{m_a \to \widetilde{m}_a} - H_a\| \leq C r_a$. Adding these errors proves (52). Calibration may be reused for later auxiliary tuning or partition construction. The displayed high-probability inequality remains a joint event. Uniformity over a finite tuning grid follows by allocating η across that grid before this argument.

Proposition 16 (A fully estimated rate with compatible smoothness)**.** *Suppose the preceding sufficient conditions hold with $p = 2$, $s_\mu = s_H = s_g = s_h = 3/2$, $s_\tau = 2$, and a propensity of Hölder order $s_e = 4$. Assume the outcome regressions at every candidate point are Hölder of order $3/2$ with common bounds. Use degree-one final weights, $h_n \asymp n^{-1/6}$, and*

$$b_p \asymp n^{-1/4}, \qquad b_g \asymp b_h \asymp b_H \asymp n^{-1/5}, \qquad b_e \asymp n^{-1/10},$$

with regression degrees at least one for g, h, H and three for the propensity. Rounding the piecewise-polynomial grid sizes to integers does not change these comparisons. Then the fully estimated procedure satisfies

$$\mathbb{E}\,\|\mathsf{P}_{\widehat{\mu}_0 \to m_0} \widehat{\tau} - \tau(x_0)\|^2 = O(n^{-2/3}).$$

With $h_{\mathrm{inf}} \asymp n^{-1/5}$ instead, the same first stages satisfy the bias and first-stage conditions for Theorem 6. That theorem additionally requires its covariance lower bound.

Proof. Take $\eta = n^{-M}$ in the preceding lemmas, with any fixed $M > 4$, and allocate it across all fitted objects. Polynomially many auxiliary candidates or trees only change the constant multiplying $\log n$. On the intersection of their good events,

$$\begin{aligned} r_a &= O(n^{-1/4}\sqrt{\log n}), & \varepsilon_e &= O(n^{-2/5}\sqrt{\log n}), \\ \varepsilon_g + \varepsilon_h &= O(n^{-3/10}\sqrt{\log n}), & E_a^H &= O(n^{-1/4}\sqrt{\log n}), \\ D_a &= O(n^{-7/10}\log n), & \rho_a &= O(h_n^{3/2}). \end{aligned}$$

The reference-point bound uses $t_\mu = 1$. It does not posit a smoother regression than b_a. For $h_n \asymp n^{-1/6}$, $q_n \asymp n^{-1/3}$, $h_n^{s_\mu} \asymp n^{-1/4}$, and $\rho_a = O(n^{-1/4})$. Substitution into every term of (16) gives $\mathcal{R}_n = O(n^{-1/2}\log n)$. The two remaining deterministic biases are $h_n^{s_\tau} = O(n^{-1/3})$ and $h_n^{2s_\mu} = O(n^{-1/2})$. Thus (19) is $O(n^{-2/3} + n^{-1}\log^2 n) = O(n^{-2/3})$ on the joint good event.

All normalized weights satisfy the bounded-total-variation requirement on $\mathcal{W}_n$, all fitted functions are norm bounded, and the reported endpoints are compact valued even on failures. The first-stage failure probability is $O(n^{-M})$, while (28) is smaller than every polynomial for these bandwidths and polynomially many trees. The constants defining $\mathcal{A}_n$ are eventually satisfied on the joint good event. Hence its complement has probability $O(n^{-M})$. The bounded loss there contributes at most $O(n^{-M})$. This establishes the expectation conclusion, including the failures.

For $h_{\mathrm{inf}} \asymp n^{-1/5}$, $q_n \asymp n^{-3/10}$, $h_{\mathrm{inf}}^{s_\tau} = n^{-2/5}$, and $h_{\mathrm{inf}}^{2s_\mu} = n^{-3/5}$. The first-stage terms remain at most $O(n^{-1/2}\log n)$. The terms involving h^{s_μ} and ρ_a only decrease. Each required bias or remainder is therefore $o(q_n)$, and $\max |\omega_i|/q_n = O((n h_{\mathrm{inf}}^p)^{-1/2}) \to 0$. The other consistency conditions follow from the displayed rates. This verifies all stated conditions except the separate covariance nondegeneracy condition, as claimed. □

The moment identities continue to hold with a misspecified nuisance. Their rate consequences depend on which fitted function converges to its truth. For example, if the propensity estimator

converges uniformly at $n^{-2/5}\sqrt{\log n}$ while the two regression errors are only bounded, $D_a = O(n^{-2/5}\sqrt{\log n})$, which is still smaller than $n^{-1/3}$. In contrast, if only the regressions converge at $n^{-3/10}\sqrt{\log n}$ and the propensity error stays bounded, this product need not be smaller than $n^{-1/3}$. Consistency can persist while that particular accelerated rate fails. An inference choice $h \asymp n^{-1/4}$ gives $q_n \asymp n^{-1/4}$, under which the latter product is negligible, with the other displayed first-stage bounds. These comparisons motivate keeping robustness and rate experiments separate.

F.4 A NONCOMMUTING MODEL WITH FIRST-ORDER CANCELLATION

Proposition 17 (An explicit noncommuting positive-definite model)**.** *On $\mathbb{S}^r_{++}$, $r \geq 2$, with the affine-invariant metric, there exist bounded noncommuting outcome laws and covariate-dependent treatment assignment satisfying the sufficient conditions of Proposition 16, with positive-definite conditional covariance of $\mathrm{Log}_{m_a} Y(a)$ at x_0. In particular, b_0, b_1 can have order $s_\mu = 3/2$ while ϑ_{x_0} has order $s_\tau = 2$.*

Proof. Take $p = 2$, X uniform on $[0,1]^2$, and $x_0 = (1/2, 1/2)$. Choose distinct noncommuting matrices m_0, m_1 in a bounded positive-definite region. For each arm choose an orthonormal tangent basis E_{aj}, $j = 1, \ldots, d_{\mathcal{M}}$, and fix a sufficiently small $\sigma > 0$. The outcome atoms are

$$y_{aj,\pm} = \mathrm{Exp}_{m_a}(\pm\sigma E_{aj}), \qquad j = 1, \ldots, d_{\mathcal{M}}.$$

Under the equal probabilities $1/(2d_{\mathcal{M}})$, their mean log at m_a is zero. Global strong convexity on this Hadamard manifold makes m_a their unique global Fréchet mean. Let

$$H_a = \frac{1}{2d_{\mathcal{M}}} \sum_{j,\pm} \mathcal{H}_{m_a}(y_{aj,\pm}).$$

Its eigenvalues are positive and finite. Decompose the joint differential as $\mathcal{J}(u_0, u_1) = J_0 u_0 + J_1 u_1$. The derivative $J_1 = \mathrm{D}_{m_1} \mathrm{Log}_{m_0} m_1$ is invertible, because the exponential and logarithm are global smooth inverses on this manifold. Also $J_0 = -\mathcal{H}_{m_0}(m_1)$ is invertible by strong convexity. Choose $v_0 \neq 0$, put $v_1 = -J_1^{-1} J_0 v_0$, and choose $w_1 \neq 0$. Thus $v_1 \neq 0$ and $J_1 w_1 \neq 0$.

For small positive constants c_b, c_f, define

$$B(x) = c_b |x_1 - 1/2|^{3/2}, \qquad F(x) = c_f (x_2 - 1/2)^2,$$
$$b_0(x) = H_0 v_0 B(x), \qquad b_1(x) = H_1\{v_1 B(x) + w_1 F(x)\}.$$

Specify each conditional outcome law by probabilities

$$p_{aj,\pm}(x) = \frac{1}{2d_{\mathcal{M}}} \pm \frac{\langle b_a(x), E_{aj} \rangle}{2\sigma} \tag{53}$$

for the corresponding fixed atom. Choose c_b, c_f small enough that $\sup_x \|b_a(x)\| \leq \sigma/(2d_{\mathcal{M}})$. Then all probabilities are at least $1/(4d_{\mathcal{M}})$, and they sum to one. Since the logarithms of the atoms at m_a are $\pm\sigma E_{aj}$, their conditional mean is exactly

$$\sum_j \sigma E_{aj}\{p_{aj,+}(x) - p_{aj,-}(x)\} = b_a(x).$$

At x_0, these means are zero and the Hessians are precisely the defined H_a. At other covariates, strong convexity supplies unique global conditional means. Bounded atoms place them in a fixed compact geodesically convex set. By reducing c_b, c_f, their means remain near m_a, inside the interior of a common working compact set. Smooth log maps and finitely many fixed atoms give all uniform geometric envelopes. The support includes a full tangent basis and is not contained in a commuting diagonal submanifold.

For every candidate $m \in \mathcal{K}$, both true regression functions are finite sums of fixed smooth vector or operator responses with coefficients $p_{aj,\pm}(x)$. Those coefficients are Hölder of order $3/2$, uniformly in x, so the same is true of the regressions, uniformly in m. The two displayed b_a have a nonzero $|x_1 - 1/2|^{3/2}$ component. In contrast,

$$\vartheta_{x_0}(x) = J_0 v_0 B(x) + J_1\{v_1 B(x) + w_1 F(x)\} = J_1 w_1 F(x),$$

which is Hölder of order two, while b_0 and b_1 remain Hölder of order $3/2$.

For treatment assignment take $e_{0,1}(x) = 1/2 + (1/5)\sin(2\pi x_1)\sin(2\pi x_2)$, which is smooth of every finite order and lies in $[3/10, 7/10]$. Generate the potential outcomes conditionally independently given X, and draw treatment independently of them given X. This enforces consistency and unconfoundedness in the observed sample.

At x_0, the conditional covariance of $\mathrm{Log}_{m_a} Y(a)$ is exactly $\sigma^2 I_{d_{\mathcal{M}}}/d_{\mathcal{M}}$. With the moments evaluated at the true nuisances, within-arm outcome variation in arm one contributes

$$\frac{1}{e_{0,1}(x_0)} J_1 H_1^{-1} \frac{\sigma^2}{d_{\mathcal{M}}} I_{d_{\mathcal{M}}} (H_1^{-1})^* J_1^*$$

to the joint effect covariance. This is positive definite since J_1, H_1 are invertible. Continuity of the finite-support conditional laws makes its minimum eigenvalue bounded below on a sufficiently small query neighborhood. The fitted regressions and propensities in Proposition 16 converge uniformly at the realized reference points, which converge to m_a. Boundedness then makes the conditional covariance of the fitted moments converge uniformly to the true covariance there. Since $\sum_i \beta_i = 1$, the same positive lower bound eventually holds for Σ_n with probability tending to one. This verifies the final nondegeneracy requirement as well. □

F.5 Sharpness of the arm-wise smoothing term

Proposition 18 (A sharp arm-wise comparison)**.** *On $\mathbb{S}^r_{++}$, $r \geq 2$, with the affine-invariant metric, there exist bounded noncommuting outcome laws with $p = 2$, $s_\mu = s_H = s_g = s_h = 3/2$, $s_\tau = 2$, and propensity smoothness $s_e \in (1/14, 1/2)$ such that the following hold. First, a fully estimated GO-learner with $h_n \asymp n^{-1/6}$ and bounded nonparametric first stages satisfying the displayed smoothness orders has squared risk $O(n^{-2/3})$. Second, for the degree-one box-kernel arm-wise local Fréchet plug-in, the leading squared bias has exact order h^{3+2s_e} and the leading variance has order $(nh^2)^{-1}$. The resulting leading bias-variance expression is minimized at $h \asymp n^{-1/(5+2s_e)}$ and has order*

$$n^{-(3+2s_e)/(5+2s_e)}. \tag{54}$$

For $s_e = 1/6$, this leading rate is $n^{-5/8}$, compared with the $O(n^{-2/3})$ squared-risk upper bound for GO-learner.

Proof. Use the atom construction in Proposition 17, with the same noncommuting endpoints, tangent bases, Hessians, and decomposition $\mathcal{J}(u_0, u_1) = J_0 u_0 + J_1 u_1$. Choose $v_0 \neq 0$, $v_1 = -J_1^{-1} J_0 v_0$, and $w_1 \neq 0$ as before, so $J_0 v_0 + J_1 v_1 = 0$. Replace the shared endpoint function by the odd Hölder function

$$B(x) = c_b \operatorname{sgn}(x_1 - 1/2)|x_1 - 1/2|^{3/2}, \qquad F(x) = c_f (x_2 - 1/2)^2,$$

and keep

$$b_0(x) = H_0 v_0 B(x), \qquad b_1(x) = H_1\{v_1 B(x) + w_1 F(x)\}.$$

The atom probabilities are again given by (53). Sufficiently small c_b, c_f make them uniformly positive. The outcome regressions and Hessian regressions are Hölder of order $3/2$, while

$$\vartheta_{x_0}(x) = J_1 w_1 F(x)$$

is Hölder of order two.

Take

$$e_{0,1}(x) = \frac{1}{2} + \kappa \operatorname{sgn}(x_1 - 1/2)|x_1 - 1/2|^{s_e}, \tag{55}$$

with $\kappa > 0$ small enough for overlap. This propensity is Hölder of order s_e and is continuous at x_0.

We first calculate the arm-wise local-linear bias. Write $u = (x_1 - 1/2)/h$. The common rough component is $B(x) = c_b h^{3/2} \operatorname{sgn}(u)|u|^{3/2}$. Under (55), the two arm-specific local design densities differ, up to a common multiplicative constant, by the odd perturbations $\pm\kappa h^{s_e} \operatorname{sgn}(u)|u|^{s_e}$. For a

symmetric box kernel, the two-dimensional degree-one normal equations separate in the x_2 coordinate, and the intercept calculation reduces to the one-dimensional basis $(1, u)$. Direct integration gives

$$\mathcal{L}_{a,h}B = \pm C_{3/2,s_e} h^{3/2+s_e} + o(h^{3/2+s_e}), \qquad C_{s_\mu,s_e} = \frac{2\kappa c_b (s_e-1)(s_\mu-1)}{(s_e+s_\mu+1)(s_e+2)(s_\mu+2)}. \quad (56)$$

The sign is opposite across the two arms and $C_{3/2,s_e} \neq 0$ for $s_e \in (0, 1/2)$. Since $J_1 v_1 = -J_0 v_0$, the rough contribution to (39) is

$$J_0 v_0 \{\mathcal{L}_{0,h}B - \mathcal{L}_{1,h}B\},$$

whose norm is asymptotic to a positive constant times $h^{3/2+s_e}$. The F contribution is $O(h^2)$ and the geometric/Hessian remainders are $O(h^3)$, both smaller because $3/2 + s_e < 2$. Thus the squared bias is asymptotic to a positive constant times h^{3+2s_e}.

At x_0, the conditional covariance of $\mathrm{Log}_{m_a} Y(a)$ is positive definite. The arm-wise local-linear weights have squared ℓ_2 norm of order $(nh^2)^{-1}$, and the derivative $J_1 H_1^{-1}$ is nonsingular. The corresponding first-order plug-in term therefore has variance bounded above and below by positive constants times $(nh^2)^{-1}$. The deterministic terms omitted from the leading bias are $O(h^2) + O(h^3)$, which are smaller than $h^{3/2+s_e}$ for $s_e \in (0, 1/2)$. The leading bias-variance expression is therefore of order $h^{3+2s_e} + (nh^2)^{-1}$. Balancing these terms gives $h \asymp n^{-1/(5+2s_e)}$ and (54).

It remains to verify that the separation is compatible with fully estimated GO nuisance functions. Use the same reference-point, vector/operator-regression, and Hessian-calibration bandwidths as Proposition 16, except estimate the s_e-Hölder propensity by bounded piecewise-constant regression with bandwidth $b_e \asymp n^{-1/(2s_e+2)}$. Lemma 15 gives, up to logarithmic factors,

$$\varepsilon_e = O_p\{n^{-s_e/(2s_e+2)}\}, \qquad \varepsilon_g + \varepsilon_h = O_p(n^{-3/10}).$$

Hence the nuisance product has exponent

$$\frac{s_e}{2s_e+2} + \frac{3}{10} > \frac{1}{3}$$

exactly when $s_e > 1/14$. The remaining reference-point estimation and Hessian terms are the same or smaller than those in Proposition 16. With $h_n \asymp n^{-1/6}$, $q_n \asymp n^{-1/3}$, $h_n^{s_\tau} = n^{-1/3}$, and $h_n^{2s_\mu} = n^{-1/2}$, Theorem 4 therefore gives squared risk $O(n^{-2/3})$. This proves the fully estimated separation for every $s_e \in (1/14, 1/2)$. □

G Implementation and tuning

G.1 Data roles and estimators

Use a seeded permutation to allocate approximately one quarter of the observations to each of $I_{\mathrm{ref}}, I_{\mathrm{reg}}, I_{\mathrm{cal}}, I_{\mathrm{final}}$. Distribute any remainder in that order. Covariate transformations are fitted on auxiliary observations and then frozen. In simulations the known covariate domain is already $[0, 1]^p$. Every object used to choose a partition or tuning parameter is computed without (A_i, Y_i) for $i \in I_{\mathrm{final}}$.

For each query, fit the positive-kernel reference points of Lemma 14. The exact mathematical definition is the constrained Fréchet minimizer. The positive-definite implementation uses a geodesic descent method with objective-decreasing backtracking and a recorded gradient or objective-gap tolerance. All returned reference points must belong to $\mathcal{K}$. A solver failure is recorded and the corresponding fit is not treated as a successful statistical estimate.

At each reference point, choose an orthonormal tangent basis and form the vector and symmetric-operator responses in (5) and (6). The fully specified first-stage option is the piecewise-polynomial least-squares procedure in Lemma 15. Use a norm-ball projection for the vector fit, and self-adjoint spectral projection for the operator fit. Set the radius from the prescribed compact response and geometry bounds. These operations and squared-norm losses are invariant under a rotation of the tangent basis. No relation between two arbitrarily chosen coordinate frames is used: vectors are

moved between anchors by parallel transport. The propensity is a least-squares regression of A, clipped to $[\epsilon, 1-\epsilon]$.

Hessian calibration uses the same formula (9), applied to I_{cal} with box-kernel base weights, its own bandwidth b_H, and degree ℓ_H. Average $\widehat{\chi}_{i,a}$ with these weights, symmetrize, and clip eigenvalues to $[\lambda/2, 2\Lambda]$. A calibration failure returns $\lambda I_{d_{\mathcal{M}}}$ and a flag. This supplies all weights and responses in the endpoint algorithm. The constants $\epsilon, \lambda, \Lambda$ and projection radius are fixed implementation inputs. On a prescribed bounded affine-invariant positive-definite domain, $\lambda = 1$ is available and a finite Λ follows from the compact operator bound. The thresholds are fixed before fitting. Sensitivity analyses may vary them explicitly.

G.2 An optional query-specific honest forest

The forest supplies α_i, not an additional identifying assumption. Fix a query x_0, its two reference points, and two auxiliary Hessians. Define the auxiliary vector response

$$Z_i = \widetilde{\mathcal{J}}(\widehat{H}_0^{-1}\widehat{\psi}_{i,0}, \widehat{H}_1^{-1}\widehat{\psi}_{i,1}), \qquad i \in I_{\mathrm{cal}}.$$

All Z_i lie in the one tangent space $T_{\widetilde{m}_0}\mathcal{M}$. These observations may have been used in calibration, because both operations remain auxiliary to the final sample. No population split-consistency claim is needed for Lemma 9.

Each tree begins at $[0,1]^p$ and uses all auxiliary split observations. There is no tree subsampling. At a node, eligible coordinates have side length greater than the target h. Independently for each tree and node, sample a uniformly chosen subset of size $\min\{\lceil p/2\rceil, \text{number eligible}\}$. On each sampled coordinate, candidate cuts occur at fractions $1/3, 1/2, 2/3$ of that node's coordinate interval. Candidates with fewer than $k_{\mathrm{split}} = \max\{5, \lceil \log n\rceil\}$ auxiliary observations in either child are excluded from score maximization. For each remaining candidate, use

$$\widehat{\pi}_-\widehat{\pi}_+ \left\|\overline{Z}_- - \overline{Z}_+\right\|^2,$$

where $\widehat{\pi}_\pm$ are the child proportions within the parent and $\overline{Z}_\pm$ are their sample means. Select its maximum, with ties broken by coordinate index and then cut location. This is exactly the reduction in average within-child squared vector error: expanding each squared deviation around its child mean gives the between-child term above. If no candidate has adequate counts, split the longest eligible side at its midpoint, again using coordinate-index ties. Empty auxiliary nodes follow this same geometric rule.

Stop only when all side lengths are at most h. Every terminal side is greater than $h/3$, since a side is split only while it exceeds h, and every selected fraction is at least $1/3$. A path has depth at most $p\{1 + \lceil \log(1/h)/\log(3/2)\rceil\}$. Thus the tree is finite and has exactly the geometry used by the weight lemma. A fixed $T_n = 100$ trees is the default. Polynomially growing tree counts are covered by the lemma. Random coordinate subsets provide the tree randomization. For $p = 1$ the trees can coincide, in which case repeated copies do not change their weights.

For each final observation, count membership in each query leaf and average the normalized leaf weights. Empty final query leaves trigger the weight failure flag. No final outcome is examined in this check. A deterministic midpoint partition with the same stopping scale gives the fixed-partition comparison. The forest is query-specific because its auxiliary vector responses use query-specific reference points. Its computation must therefore be charged per query. Local-polynomial correction is applied after the base weights are formed.

G.3 Estimation tuning and inference tuning

The following validation rule makes point-estimation tuning reproducible. Fix a finite candidate set of bandwidths and degrees before fitting. One practical bandwidth set on $[0,1]^p$ is $\{0.05, 0.075, 0.10, 0.15, 0.20, 0.30\}$, with degree one for the proposed estimator. The experiments below use degree one. Higher polynomial degrees, when used, must meet Assumption 3. The explicit sufficient-rate schedule in Proposition 16 is used in the additional inference rule reported in Appendix H.4.

First choose nuisance regression tuning by a seeded training/validation split inside I_{reg}. Minimize validation squared error for the propensity, squared tangent norm for g, and squared Frobenius norm

for h. Refit the chosen regressions on all of I_{reg}. For the reference-point bandwidth, use the analogous split inside I_{ref}, minimizing validation squared geodesic distance within each arm, then refit. Fix all grids and tie rules in the experimental configuration. Uniform finite-grid bounds allocate the failure probability over every candidate.

To choose the final estimation bandwidth at x_0, use preliminary Hessians obtained by spectrally clipping $\widehat{h}_a(\widetilde{m}_a, x_0)$, without calibration outcomes. Form the response Z_i above on I_{cal} using these preliminary Hessians. Conditional on the reference-point and regression blocks, these are independent vector responses in a fixed tangent space. Split I_{cal} into equally sized fitting and validation parts. For every candidate (h, ℓ), fit local-polynomial regression of Z_i on the fitting part. Evaluate at validation covariates in the fixed cube $\|X - x_0\|_\infty \le 1/4$. The validation loss is

$$\widehat{L}(h, \ell) = \frac{1}{|V_{x_0}|} \sum_{i \in V_{x_0}} \left\| Z_i - \widehat{f}^{\mathrm{fit}}_{h,\ell}(X_i) \right\|^2, \tag{57}$$

where V_{x_0} is this validation subset. A candidate with a failed prediction Gram matrix has loss $+\infty$. Choose the minimizer, resolving equal losses by larger bandwidth and then smaller degree. If $|V_{x_0}| < 5$, or all losses are infinite, use the largest bandwidth and smallest degree and report a tuning-failure flag. This rule estimates the predictive error of the auxiliary response Z_i using auxiliary observed data. The target effect and counterfactual outcomes are excluded from tuning. The same kernel-based bandwidth rule is used for the kernel and optional forest versions, so their comparison does not change the tuning budget. After selecting tuning, calibrate the Hessians on all of I_{cal}, using the prespecified calibration bandwidth and degree, and build any final auxiliary forest. Reusing this auxiliary block does not expose final outcomes.

For centered inference, specify a separate bandwidth before looking at final outcomes. With class smoothness orders s_τ, s_μ, a sufficient deterministic choice is $h_{\mathrm{inf}} = c_h n^{-\kappa}$, where

$$\max\left\{ \frac{1}{2s_\tau + p}, \frac{1}{4s_\mu + p} \right\} < \kappa < \frac{1}{p},$$

provided the first-stage remainder is also $o_p\{(nh_{\mathrm{inf}}^p)^{-1/2}\}$. The two strict lower inequalities make both localization biases smaller than the standard error. They do not by themselves control nuisance errors. Proposition 16 provides one fully verified choice. In applications with unknown smoothness, the class assumptions and sensitivity across prespecified undersmoothing choices should be reported. The present theorem covers prespecified undersmoothing and does not establish unrestricted post-selection coverage or adaptive bias correction. In particular, a validation-selected MSE bandwidth generally does not satisfy the centered Wald conditions.

G.4 Positive-definite matrix operations

For $r \times r$ positive-definite matrices, $d_{\mathcal{M}} = r(r+1)/2$. The affine-invariant inner product and basic maps are (Pennec et al., 2006)

$$\begin{aligned}
\langle U, V \rangle_S &= \mathrm{tr}(S^{-1} U S^{-1} V), \\
d(S, T) &= \left\| \log(S^{-1/2} T S^{-1/2}) \right\|_F, \\
\mathrm{Exp}_S U &= S^{1/2} \exp(S^{-1/2} U S^{-1/2}) S^{1/2}, \\
\mathrm{Log}_S T &= S^{1/2} \log(S^{-1/2} T S^{-1/2}) S^{1/2}.
\end{aligned}$$

For $C = S^{-1/2} T S^{-1/2}$, put $E = S^{1/2} C^{1/2} S^{-1/2}$. Parallel transport is $\mathsf{P}_{S \to T} U = E U E^\top$. These formulas give an isometry because $E S E^\top = T$, so direct substitution in the inner product cancels the congruence factors. The frame $S^{1/2} F_j S^{1/2}$, where F_j ranges over diagonal unit matrices and normalized symmetric off-diagonal matrices, is orthonormal.

Let $G(S, T) = \mathrm{Log}_S T$. The affine-invariant connection on a tangent vector field $V(S)$ is

$$\nabla_U V = \mathrm{D}V(S)[U] - \tfrac{1}{2}\{U S^{-1} V + V S^{-1} U\}.$$

Consequently the Hessian response and the transported endpoint differential are

$$\mathcal{H}_S(T)[U] = -\mathrm{D}_S G(S, T)[U] + \tfrac{1}{2}\{U S^{-1} G + G S^{-1} U\}, \tag{58}$$

$$\mathcal{J}_{S,T}(U,V) = \mathrm{D}_S G(S,T)[U] + \mathrm{D}_T G(S,T)[V] - \tfrac{1}{2}\{US^{-1}G + GS^{-1}U\}. \tag{59}$$

The connection term in (59) differentiates the transported vector at its moving base point. Differentiating the entries of G alone omits this term and computes a different derivative. Equation (58) represents the Hessian by its action on a tangent basis. Its matrix has entries $\langle E_j, \mathcal{H}_S(T)[E_k]\rangle_S$. Symmetrize this $d_{\mathcal{M}} \times d_{\mathcal{M}}$ matrix before calibration.

Matrix-function derivatives can be computed by spectral divided differences (Higham, 2008). If $S = Q\,\mathrm{diag}(\lambda_j)Q^\top$, then

$$\mathrm{D}f(S)[U] = Q\{F \circ (Q^\top U Q)\}Q^\top, \qquad F_{jk} = \begin{cases} \{f(\lambda_j) - f(\lambda_k)\}/(\lambda_j - \lambda_k), & \lambda_j \neq \lambda_k, \\ f'(\lambda_j), & \lambda_j = \lambda_k. \end{cases}$$

Here $\circ$ is the entrywise product. The derivative value applies whenever eigenvalues coincide, including off-diagonal indices. Near coincidence use the continuous limit or a stable scalar evaluation. Applying this formula to the square root, inverse square root, logarithm, and exponential, followed by the ordinary product rule, evaluates (58) and (59). A separate Jacobi boundary-value solver is therefore unnecessary in this implementation.

G.5 Numerical safeguards

Fréchet optimization uses the convergence criteria stated in Appendix H.2. Gram-matrix validity, endpoint safeguards, Hessian clipping, covariance nondegeneracy, and tuning failures are recorded for every fit. A failed point estimate is retained as a failed fit, and an unconstructed confidence ellipsoid is counted as noncoverage. All planned point estimates and all 21,000 reported confidence ellipsoids in the completed experiments were successfully constructed.

H Experimental details and additional results

This appendix gives the generators and implementation choices for Section 6, followed by additional estimation, inference, nuisance-robustness, and arm-wise smoothing results. All numerical results below come from the same completed run.

H.1 Data-generating processes

Matrix-outcome benchmark. Let

$$F_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad F_3 = 2^{-1/2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

These matrices form an orthonormal tangent basis at the identity under the affine-invariant metric. For X uniform on $[0,1]^2$, set

$$\begin{aligned} B(x) &= 0.4\sin(2\pi x_1)(F_1 - F_2) + 0.3(x_2 - 1/2)F_3, \\ V(x) &= 0.25(1 + x_1)(F_1 - F_2) + 0.2\sin(\pi x_2)F_3, \\ \mu_0(x) &= \exp B(x), \qquad \mu_1(x) = \mathrm{Exp}_{\mu_0(x)}\{\mathsf{P}_{I \to \mu_0(x)} V(x)\}. \end{aligned}$$

Generate independent tangent-noise coordinates $\varepsilon_{aj} \sim U[-\sqrt{3}, \sqrt{3}]$ for $a \in \{0,1\}$ and $j = 1,2,3$, and put $\varepsilon_a = \sum_j \varepsilon_{aj} F_j$. The potential outcomes are

$$Y(a) = \mathrm{Exp}_{\mu_a(X)}\{0.3\,\mathsf{P}_{I \to \mu_a(X)}\varepsilon_a\}.$$

Their conditional mean logarithms at $\mu_a(x)$ are zero, giving the prescribed conditional Fréchet means. Treatment is conditionally independent of the potential outcomes given X, with

$$e_{0,1}(x) = \tfrac{1}{2} + \tfrac{1}{5}\sin\{\pi(x_1 + x_2)\}, \qquad Y = Y(A).$$

The exact target is $\tau(x) = \mathsf{P}_{I \to \mu_0(x)} V(x)$. The endpoints generally do not commute, including at the two off-center estimation queries. At the inference query $x_0 = (1/2, 1/2)$, $\mu_0(x_0) = I$. The outcome noise still spans the full three-dimensional tangent space. The zero-effect inference experiment sets $V(x) = 0$ and leaves the noise and treatment assignment unchanged.

Finite-support model for nuisance robustness. We use the construction of Proposition 17 with $d_{\mathcal{M}} = 3$, atom radius $\sigma = 0.2$, and fixed noncommuting centers

$$m_0 = \mathrm{diag}(e^{0.4}, e^{-0.4}), \qquad m_1 = \exp\begin{pmatrix} 0 & 0.35 \\ 0.35 & 0 \end{pmatrix}.$$

The six atoms in arm a are $y_{aj,\pm} = \mathrm{Exp}_{m_a}(\pm 0.2 E_{aj})$, where $E_{aj} = m_a^{1/2} F_j m_a^{1/2}$. Write H_a for the equal-probability average of their Hessian responses and decompose the endpoint differential as $\mathcal{J}(u_0, u_1) = J_0 u_0 + J_1 u_1$. In these orthonormal frames, set

$$v_0 = (1, 0, 0)^\top, \qquad w_1 = (0, 0, 1)^\top, \qquad v_1 = -J_1^{-1} J_0 v_0.$$

The common coefficient used by the supplied generator is

$$c = \frac{0.2}{12(1+M)},$$
$$M = \max\Big\{2^{-3/2}\|H_0 v_0\|, 2^{-3/2}\|H_1 v_1\| + \tfrac{1}{4}\|H_1 w_1\|\Big\}.$$

For these fixed matrices, this gives $c = 0.010315827179120341$. Define the score fields

$$B(x) = 12c(x_1 - 1/2)^2, \qquad F(x) = 12c(x_2 - 1/2)^2,$$
$$b_0(x) = H_0 v_0 B(x), \qquad b_1(x) = H_1\{v_1 B(x) + w_1 F(x)\}.$$

Start with the probabilities in (53). Add $q(x) = 0.24(x_2 - 1/2)^2$ to each of the two probabilities in direction $j = 1$ and subtract it from each probability in direction $j = 3$. This preserves their sum and their endpoint score at m_a, while making the Hessian regression covariate-dependent. All probabilities remain positive. The propensity is

$$e_{0,1}(x) = 0.25 + 1.6(x_1 - 1/2)^2.$$

At x_0, $B = F = q = 0$, so m_0, m_1 are the exact conditional Fréchet means. The correct nuisance classes use the fixed basis $(1, x_1, x_2, x_1^2, x_1 x_2, x_2^2)$. This basis represents the propensity and both regressions exactly at any realized reference point. Wrong classes use only an intercept. Their coefficients are fitted on the same finite regression block as the correct classes.

H.2 Training and reporting protocol

Sample sizes and data use. Every sample-size experiment uses $n = 1000 \times 2^j$ for $j = 0, \ldots, 6$. A seeded permutation divides the GO observations equally into $I_{\mathrm{ref}}, I_{\mathrm{reg}}, I_{\mathrm{cal}}, I_{\mathrm{final}}$. Baselines receive the same total n. Only observed treatments and outcomes are passed to the fitted learners. Ground-truth functions are evaluated afterward, except in the explicitly labeled oracle-nuisance diagnostic. The benchmark uses 100 independent repetitions with paired data across methods. Inference uses 500 repetitions for each of the nonzero- and zero-effect models. Nuisance robustness uses 100 repetitions per sample size. Seeds are fixed before fitting and independent of worker scheduling.

GO-learner. The reference points use positive box-kernel Fréchet smoothing. The final local-polynomial degree is one. Nuisance least-squares regressions use degree three for the propensity and degree one for the vector and Hessian responses within each cell. The six candidate cell resolutions are $1, \ldots, 6$ per covariate axis. The six bandwidth candidates are $\{0.05, 0.075, 0.10, 0.15, 0.20, 0.30\}$. Nuisance fits and reference-point bandwidths use internal 75/25 training/validation splits, then refit on their full designated blocks. Geodesic validation for reference points and baselines uses a seeded subset of at most 40 held-out observations selected independently of their outcomes. Final-bandwidth selection uses the 50/50 split of I_{cal} and all validation covariates in the fixed radius-$1/4$ cube around the query, as defined in Appendix G.3. Hessian calibration uses degree one and bandwidth $n^{-1/5}$.

Propensities are clipped to $[0.1, 0.9]$. Vector regressions are projected to a norm ball of radius 10. Fitted Hessian eigenvalues are clipped to $[1, 10]$, and calibrated Hessians to $[0.5, 20]$. Endpoints are constrained to the geodesic ball of radius 6 around I. Fréchet optimization uses gradient tolerance 10^{-7} and at most 100 iterations.

Table 2: **MSE at each fixed query.** Mean squared transported tangent-space error is reported in units of 10^{-3}, with MCSE in parentheses over 100 independent repetitions. Bold marks the smallest mean in each row. "Arm-wise" denotes arm-wise local-linear Fréchet regression and "RF-weighted" denotes RF-weighted local-constant Fréchet regression.

n	GO	Arm-wise	RF-weighted
Query $x = (0.5, 0.5)$			
1,000	40.168 (6.386)	**12.123** (**1.219**)	13.686 (1.188)
2,000	13.554 (1.387)	**9.984** (**1.021**)	13.358 (1.303)
4,000	6.161 (0.764)	**5.890** (**0.535**)	11.606 (1.005)
8,000	**3.543** (**0.421**)	4.558 (0.506)	8.718 (0.687)
16,000	**2.017** (**0.131**)	3.163 (0.380)	8.048 (0.711)
32,000	**1.439** (**0.095**)	1.781 (0.193)	7.588 (0.557)
64,000	1.068 (0.070)	**0.943** (**0.079**)	7.221 (0.612)
Query $x = (0.25, 0.5)$			
1,000	35.466 (4.499)	14.353 (1.299)	**13.431** (**1.174**)
2,000	15.983 (2.430)	12.162 (1.066)	**9.264** (**0.677**)
4,000	**6.898** (**0.676**)	12.008 (1.161)	8.399 (0.884)
8,000	**3.445** (**0.436**)	7.901 (0.867)	7.737 (0.867)
16,000	**2.127** (**0.146**)	7.907 (1.112)	7.022 (0.809)
32,000	**1.600** (**0.137**)	4.880 (0.619)	4.776 (0.530)
64,000	**1.109** (**0.077**)	3.043 (0.413)	5.282 (0.479)
Query $x = (0.75, 0.5)$			
1,000	45.129 (5.812)	15.459 (1.307)	**12.278** (**1.092**)
2,000	24.700 (12.506)	16.404 (2.308)	**11.037** (**0.815**)
4,000	**7.056** (**0.754**)	14.127 (1.565)	8.776 (0.783)
8,000	**3.235** (**0.227**)	8.593 (0.892)	7.543 (0.827)
16,000	**1.696** (**0.117**)	6.687 (0.939)	6.448 (0.809)
32,000	**1.401** (**0.125**)	5.126 (1.088)	5.126 (0.521)
64,000	**0.951** (**0.063**)	3.026 (0.432)	4.761 (0.534)

Arm-wise baselines. Local-linear Fréchet regression selects an arm-specific bandwidth by held-out squared geodesic loss, refits using the full arm, and applies the logarithm to the two fitted means. The validation set is capped at 40 observations per arm, as specified above. The RF baseline implements the local-constant estimator of Qiu et al. (2024): random-forest weights are averaged first and then used in a weighted Fréchet objective. It is an independent Python implementation for the affine-invariant metric. Each of 100 trees samples $\lceil n_a^{0.8} \rceil$ observations without replacement from an arm of size n_a. Each node considers one randomly chosen covariate and its $1/3$, $1/2$, and $2/3$ empirical quantiles, with a minimum child size of five. Splits minimize within-child Fréchet loss. Six candidate maximum depths, equally spaced after rounding from 3 to $\max\{8, \lceil \log_2 n_{a,\mathrm{train}} \rceil\}$, are selected by held-out geodesic loss using the same validation cap. Prediction averages the leaf observation weights across trees before solving a weighted Fréchet problem.

Inference and augmentation ablation. The two main inference rules share their fitted first stages, reference points, and calibrated Hessians. They differ only in the final bandwidth. The additional rule based on Proposition 16 uses reference bandwidth $n^{-1/4}$, regression cell resolutions obtained by rounding $n^{1/10}$ for the propensity and $n^{1/5}$ for the vector and Hessian regressions, calibration bandwidth $n^{-1/5}$, and final inference bandwidth $n^{-1/5}$. All bandwidth constants equal one. The separate robustness experiment uses the fixed quadratic or intercept-only nuisance classes, reference bandwidth $n^{-1/4}$, calibration bandwidth $n^{-1/5}$, and final bandwidth $n^{-1/6}$. Its regression-only ablation replaces the augmented evaluations by $\widehat{\psi}_{i,a}^{\mathrm{reg}} = \widehat{g}_a(\widetilde{m}_a, X_i)$ and $\widehat{\chi}_{i,a}^{\mathrm{reg}} = \widehat{h}_a(\widetilde{m}_a, X_i)$. It retains the same remaining components and is evaluated only for point estimation.

Monte Carlo summaries. Let $e_r(x)$ denote the transported estimation error in independent repetition r. For the three-query benchmark, the loss is first averaged across queries within each repetition. The MCSE is the sample standard deviation of these repetition-level losses divided by the square root of the repetition count. Paired comparisons apply the same calculation to within-repetition differences between methods. At a fixed query with R independent repetitions, squared Monte Carlo

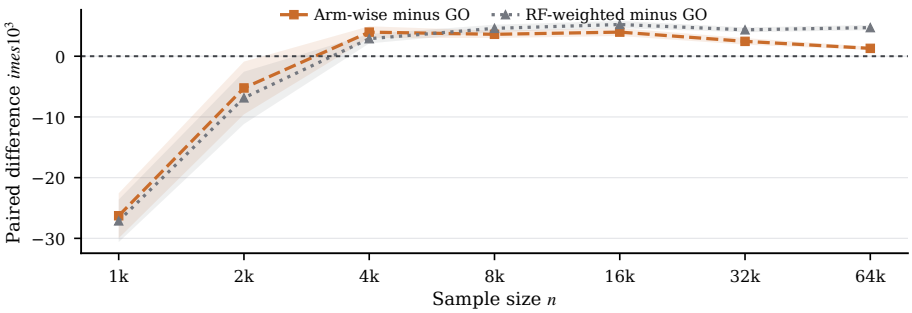


Figure 2: **Paired MSE differences.** Baseline-minus-GO MSE is reported in units of 10^{-3} after averaging the three queries within each repetition. Shaded regions are ± 1 paired MCSE over 100 repetitions. Positive values favor GO-learner. The dashed horizontal line marks equality.

bias satisfies

$$\widehat{B}^2_{\mathrm{MC}}(x) = \left\| R^{-1} \sum_{r=1}^{R} e_r(x) \right\|^2, \qquad \mathbb{E}\{\widehat{B}^2_{\mathrm{MC}}(x)\} = \|\mathbb{E} e_r(x)\|^2 + \frac{\mathrm{tr}\{\mathrm{Var}(e_r(x))\}}{R}. \tag{60}$$

It can therefore be positive when the finite-sample estimator bias is zero. The exported summaries estimate the second term. When the reported value is below this estimate, the simulation cannot distinguish a nonzero finite-sample bias from Monte Carlo variation at the available number of repetitions. Coverage uses all planned repetitions as the denominator and counts an unconstructed ellipsoid as noncoverage. Wilson intervals describe Monte Carlo uncertainty in the coverage probability, separately from the parameter confidence sets.

H.3 Additional estimation results

Accuracy across queries. Table 2 reports MSE separately at the three fixed queries. The center query is the one exception to the large-sample average ordering: at $n = 64000$, arm-wise Fréchet regression has lower MSE there, while GO-learner remains lower at the two off-center queries. Figure 2 complements the query-wise table with paired repetition-level differences averaged over the three queries. For every $n \geq 4000$, both baseline-minus-GO differences exceed three paired MCSEs.

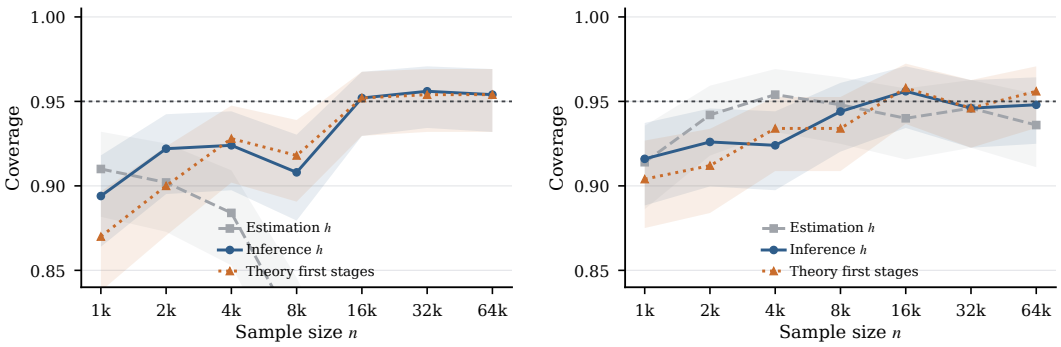
H.4 Additional pointwise inference results

Coverage for nonzero and zero effects. Figure 3 reports all three inference rules in both the nonzero- and zero-effect models. "Estimation h" uses validation-selected first stages and h_{est}. "Inference h" keeps those first stages fixed and changes only the final bandwidth to h_{inf}. "Theory first stages" uses the theoretical first-stage schedule together with h_{inf}. All $7 \times 3 \times 2 \times 500 = 21{,}000$ confidence ellipsoids were successfully constructed. In the zero-effect model, the vector confidence set remains well defined at $\tau(x_0) = 0$.

Bias and confidence-set size. Figure 4 isolates the tradeoff between centering and confidence-set size for the two practical rules. At $n = 64000$, the average estimation bandwidth is 0.232, compared with 0.109 for inference. The corresponding squared Monte Carlo biases are 5.53×10^{-4} and 2.49×10^{-5}. The inference bandwidth therefore reduces squared Monte Carlo bias while producing larger confidence ellipsoids because fewer local observations receive appreciable weight. The bandwidth constant is prespecified and is not selected or recalibrated from the observed coverage.

H.5 Additional nuisance-robustness results

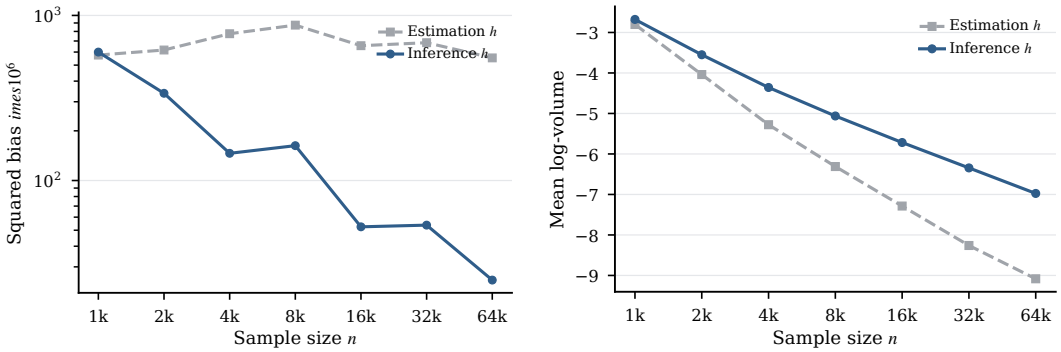
Model specification. We write C for a correctly specified model class and W for an intercept-only class with nonzero population approximation error at the query. The same specification label is applied to both regression functions g and h. In a C cell the true nuisance belongs to the fitted



(a) Nonzero effect.

(b) Zero effect.

Figure 3: **Pointwise coverage across inference rules.** Empirical coverage of nominal 95% vector confidence sets over 500 repetitions. Shaded regions are 95% Wilson intervals, and the dashed line is the nominal level.



(a) Squared Monte Carlo bias, units of 10^{-6}.

(b) Confidence-set size.

Figure 4: **Inference-bandwidth diagnostics in the nonzero-effect model.** The two rules share the same fitted first stages and differ only in the final bandwidth. Panel (a) reports squared Monte Carlo bias in units of 10^{-6}. Panel (b) reports the mean log-volume of the confidence ellipsoid in an orthonormal tangent frame.

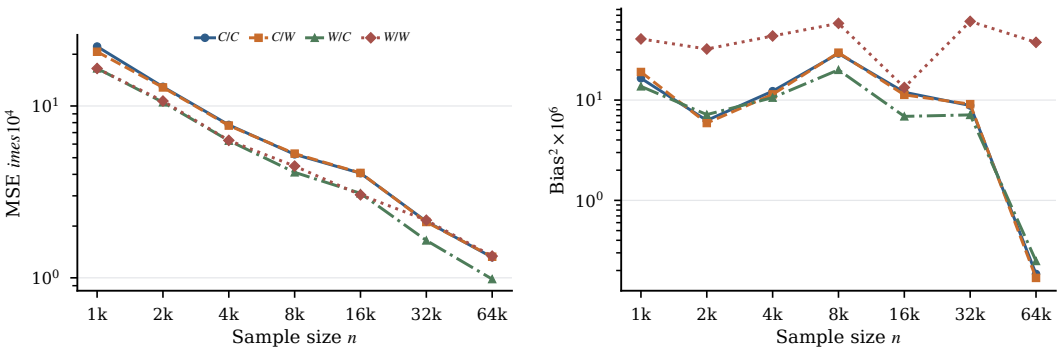
class, while its coefficients are estimated from the finite regression block. Population calculations at all 700 realized reference-point pairs verify C membership and nonzero approximation error for every W component. Proposition 2 gives zero conditional moment error when one fitted nuisance equals its truth. A fitted nuisance in a C cell generally still has finite-sample estimation error, so the product term need not equal zero.

Bias reduction and variance. Figure 5 shows the full sample-size paths for GO-learner and for the regression-only ablation. With augmentation, squared Monte Carlo bias is small at the largest sample size whenever either nuisance class is correct, while the W/W case retains substantial bias. Without augmentation, the C/W squared Monte Carlo bias remains near 1.25×10^{-4} at $n = 64000$, compared with 1.69×10^{-7} for GO-learner. In the correct-regression cases, the regression-only procedure has lower MSE, consistent with the additional variance introduced by the residual correction. At $n = 64000$, the three GO cells containing at least one correct nuisance class have squared Monte Carlo bias at most 2.49×10^{-7}, below the estimated $R^{-1}\,\mathrm{tr}\{\mathrm{Var}(e_r)\}$ contribution in (60), which is of order 10^{-6} for $R = 100$. With 100 repetitions, the simulation cannot distinguish a nonzero finite-sample estimator bias from Monte Carlo variation in those cells.

H.6 Arm-specific smoothing diagnostic

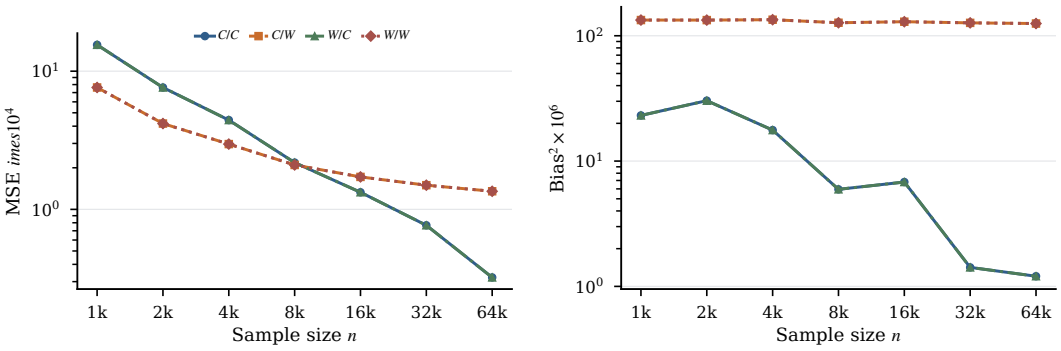
Proposition 18 identifies

$$D(h) = \|\{\mathcal{L}_{1,h} - \mathcal{L}_{0,h}\} f_1\|$$



(a) GO-learner: MSE.

(b) GO-learner: squared Monte Carlo bias.



(c) Without augmentation: MSE.

(d) Without augmentation: squared Monte Carlo bias.

Figure 5: **Nuisance misspecification across sample sizes.** C means that the true nuisance belongs to the fitted class, with coefficients estimated from data. W is an intercept-only class with nonzero population approximation error at the query. The first letter refers to e and the second to (g, h). MSE and squared Monte Carlo bias are reported in units of 10^{-4} and 10^{-6}. Each point summarizes 100 repetitions.

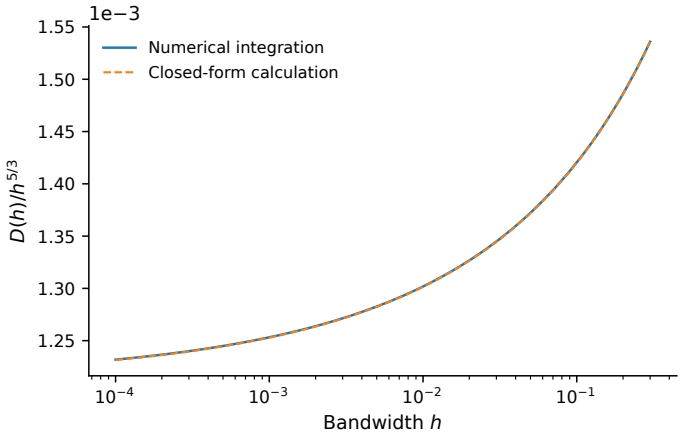


Figure 6: **Arm-specific population smoothing difference.** Numerical and closed-form evaluations of $D(h)/h^{5/3}$ in the construction of Proposition 18. Stabilization as h decreases verifies the order of the additional arm-wise bias term.

as the additional population bias term created by different treatment-specific local covariate distributions. We evaluate this term directly in the noncommuting positive-definite construction of Proposition 18. Numerical integration and an independent closed-form calculation agree, and $D(h)/h^{5/3}$ stabilizes as h decreases, matching the order $h^{s_\mu+\bar{s}_e}$ with $s_\mu = 3/2$ and $\bar{s}_e = 1/6$.