PREDICTIVE AFFINITY LEARNING FOR LAPLACIAN EIGENMAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

In many scientific studies, reducing high-dimensional measurements to a few coordinates simplifies prediction and visualization. Laplacian Eigenmaps construct coordinates from a neighborhood graph, but nuisance features can distort its distances and the standard eigenmap loss can favor almost disconnected graphs. In this article, we study how to learn a feature-weighted distance for Laplacian Eigenmaps using a predictive criterion. We propose predictive affinity learning, which predicts excluded features from neighbor averages and constructs spectral coordinates from the selected features. Theoretical results show that, under the stated identification and regularity conditions, the method removes independent nuisance features and its spectral representation converges to that of the population predictive distance, even when the same observations estimate the distance and construct the graph. The selected features persist under small departures from independence when the population criterion strictly separates the selected and excluded coordinates. Through simulations and multiview data analyses, we identify settings where the learned distance improves neighborhood recovery and prediction over isotropic eigenmaps.

1 INTRODUCTION

Nuisance features can obscure variation needed for prediction and classification. Spectral coordinates and clusters depend on a neighborhood graph (Ng et al., 2002; von Luxburg, 2007). CITE-seq measures surface proteins and RNA in the same cells (Stoeckius et al., 2017). Protein representations should preserve variation that predicts RNA, although antibody panels can include isotype controls for nonspecific binding (10x Genomics, 2018). Figure 1 illustrates a synthetic surface with 48 independent nuisance coordinates: unweighted Laplacian Eigenmaps (isotropic LE) distort the four latent quadrants and principal component analysis (PCA) mixes them, whereas our predictive affinity preserves them.

Laplacian Eigenmaps and diffusion maps construct coordinates from local similarities (Belkin & Niyogi, 2003; Coifman & Lafon, 2006). Local scaling (Zelnik-Manor & Perona, 2005), adaptive windows (Xu et al., 2024), and feature-weighted distances such as Soft Laplacian Eigenmap (Fathinezhad et al., 2025) adjust these similarities. Adaptive kernelized diffusion maps select parameters by spectral criteria and report bandwidth degeneracy (Aboussaad et al., 2026). Learning the graph remains difficult: we prove that the eigenmap loss approaches zero for almost disconnected graphs even with symmetric bistochastic affinities, positive off-diagonal entries, and positive row entropy. A small eigenmap loss alone does not establish an informative representation. We assess distances through prediction of measurements excluded from their construction.

Prediction and reconstruction have also been used to learn distances and graphs. Metric learning for kernel regression (MLKR) uses responses (Weinberger & Tesauro, 2007), while supervised metric and similarity methods use labels or pairwise relations (Bach & Jordan, 2006; Ying & Li, 2012). In nonparametric regression, cross-validated bandwidths smooth out irrelevant regressors (Hall et al., 2007), and Rodeo keeps large bandwidths for irrelevant variables (Lafferty & Wasserman, 2008). Sparse clustering weights features by their contribution to a clustering criterion (Witten & Tibshirani, 2010). Unlike these methods, the label-free criterion studied here requires neither a prespecified response view nor a clustering objective. Smooth graph learning estimates edges from signal variation (Kalofolias, 2016; Dong et al., 2016; Egilmez et al., 2017). Self-smoothing and self-diffusion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

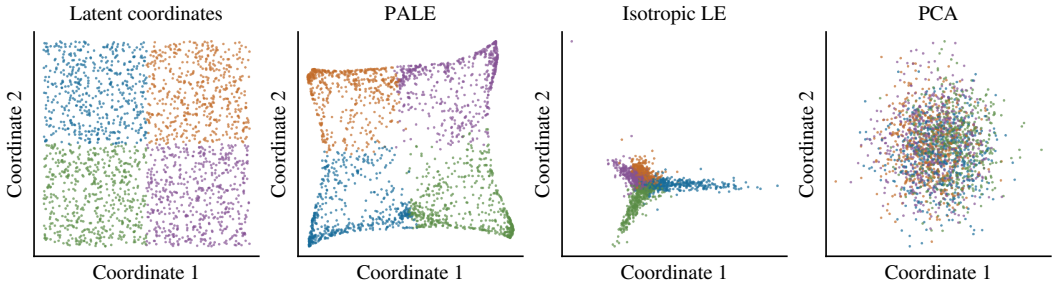

Figure 1: Two-dimensional representations of the surface design with 48 independent nuisance coordinates. Colors mark the four latent quadrants.

transform similarities, and unsupervised multiple-kernel learning uses reconstruction and prescribed local geometry (Jiang et al., 2011; Wang & Tu, 2012; Zhuang et al., 2011). Self-supervised multiview metric learning studies prediction under a latent-factor model (Wang, 2023). Cross-channel autoencoders and denoising predict excluded measurements, and SLAPS learns adjacency by feature denoising (Zhang et al., 2017; Batson & Royer, 2019; Fatemi et al., 2021). SUBLIME uses contrastive agreement (Liu et al., 2022), while GraphMAE2 learns representations on an observed graph (Hou et al., 2023). These results do not establish spectral convergence for the feature-exclusion criterion studied here when the same observations estimate the distance and construct the graph.

In this article, we propose predictive affinity learning, abbreviated PALE, for Laplacian Eigenmaps. We learn feature weights by minimizing regularized prediction error for randomly excluded coordinates or a prespecified response view. We remove weights near their lower bound and construct spectral coordinates from the selected features.

We establish three results. First, independent nuisance kernels cannot improve population prediction in product-view models, and nuisance weights attain their lower bound under verified curvature and multiplier conditions. Second, we bound metric and spectral estimation errors when feature selection and graph construction share an estimated distance. Uniform spectral bounds over a compact metric family permit this dependence. Bistochastic kernel theory studies normalization (Marshall & Coifman, 2019; Landa & Cheng, 2023; Cheng & Landa, 2024). Fixed-distance theory studies graph-Laplacian and spectral approximation (Belkin & Niyogi, 2008; Singer, 2006; Hein et al., 2005; Belkin & Niyogi, 2007; García Trillos & Slepčev, 2018; Calder & García Trillos, 2022; Cheng & Wu, 2022; Cheng et al., 2027). Spectral series regression relates an eigenbasis to conditional-mean approximation (Lee & Izbicki, 2016). Third, we prove support stability under small departures from independence with strict population separation. For smooth signal views and independent nuisance coordinates, Theorem 10 shows that, under its ridge and embedding conditions, every limit of the population metrics as bandwidth decreases minimizes the roughness of the held-out signal conditional means plus the limiting ridge penalty.

Augmentation invariant manifold learning gives spectral and classification guarantees from augmented observations (Wang, 2025); our predictive metric does not require augmented copies. Section T gives further comparisons. Simulations and two molecular data analyses assess neighborhood recovery and prediction.

Our main spectral result assumes fixed feature dimension, bounded measurements, an identified metric, a smooth compact manifold for selected features, and a positive eigengap. The remainder of the article is organized as follows. Sections 2–4 give the failure result, methods, and theory. Sections 5 and 6 present numerical results and discussion. The Supplementary Materials contain proofs, further theory including full symmetric metrics and new observations, implementation details, complete numerical results, and a posterior-moment extension for classification with measurements available only during training.

Algorithm 1 Predictive affinity learning for Laplacian Eigenmaps

Require: Blocks $I_1, \dots, I_R$ of standardized metric observations, representation observations $x_1, \dots, x_n$, tuning values $\delta, \rho, \mathcal{H}, k, J, t_N, k_n, p$, and optionally a response view V

1: For each block, draw J independent masks that select k coordinates uniformly, or set every mask to $\mathbf{1}$ when V is prespecified.
2: On each block I_r and mask ξ, scale the kernel (3) to the affinity (4) and evaluate the loss (5) for the omitted columns or V.
3: Compute $(\widehat{a}, \widehat{h})$ satisfying (7) for the criterion (6), with each block loss averaged over its J masks.
4: Select $\widehat{S}$ and the representation metric $\widehat{a}^{\mathrm{rep}}$ by (8).
5: Form the symmetric k_n-nearest-neighbor affinity $\widehat{A}$ under $d_{\widehat{a}^{\mathrm{rep}}}$ and its Laplacian $\widehat{L}$.
6: **return** the coordinates $\widehat{Y}$ in (9)

2 Why the eigenmap loss cannot learn its own affinity

For a connected graph with symmetric bistochastic affinity $W \in \mathbb{R}^{n \times n}$, the Laplacian is $L = I_n - W$, where I_n is the identity. Write its eigenvalues as $0 = \lambda_1(L) \le \cdots \le \lambda_n(L)$. Proposition 9 gives the minimum p-dimensional eigenmap loss $\Phi_p(W) = \sum_{r=2}^{p+1} \lambda_r(L)$. The following theorem shows that degree-one normalization does not bound this loss away from zero. Here $\mathbf{1}$ is an all-ones vector and diag denotes diagonal extraction or construction.

Theorem 1. *Let $p \ge 1$, let $s \ge 3$, and put $n = (p+1)s$. For every $\varepsilon \in (0, 1)$, there is a symmetric affinity W_ε such that*

$$W_\varepsilon \ge 0, \qquad \mathrm{diag}(W_\varepsilon) = 0, \qquad W_\varepsilon \mathbf{1} = \mathbf{1},$$

all off-diagonal entries are positive, and

$$\Phi_p(W_\varepsilon) = \frac{pn}{n-1}\varepsilon.$$

The entropy of every row converges to $\log(s-1)$ as $\varepsilon \downarrow 0$.

Theorem 1 shows that weak intergroup connections reduce the loss while within-group mass preserves positive row entropy. Proposition 10 gives the same limit as bandwidth tends to zero when a feature is constant within equally sized groups and distinct across them. Degree penalties and row-entropy bounds below $\log(s-1)$ do not prevent this degeneracy, motivating a predictive criterion.

3 Predictive affinity learning

Consider measurements $X = (S, U)$ with independent signal and nuisance components and a response view V excluded from the distance:

$$P_{S,U} = P_S \otimes P_U, \qquad V = g(S) + \zeta, \qquad \mathbb{E}(\zeta \mid S, U) = 0. \tag{1}$$

Here $g(S)$ is the conditional mean, ζ the error, P_Z the distribution of a random vector Z, $\otimes$ the product distribution, and $\mathbb{E}, \mathbb{P}$ expectation and probability. Write vector ℓ_q norms as $\|\cdot\|_q$. Proposition 1 shows that a product kernel's population risk equals $\mathbb{E}\|\zeta\|_2^2$ plus signal smoothing error, independently of the nuisance kernel. Nuisance weights reach their lower bound if reallocating their trace mass to signal decreases regularized risk. Propositions 28 and 45 verify decreasing unpenalized risk in the signal weight. In CITE-seq, X contains proteins and V an RNA panel.

3.1 Prediction across measurement views

We learn the distance, select features, and construct spectral coordinates in four steps, summarized in Algorithm 1. Repeated masks improve the mask average within each block without adding observations.

Step 1: Masked distances. Let $X \in \mathbb{R}^{n \times m}$ have standardized observation rows $x_i^\top$, with scales fixed in advance or estimated from an independent pilot sample. For $0 < \delta < 1$, define

$$\mathcal{A}_\delta = \left\{ a \in \mathbb{R}^m : a_j \ge \delta, \quad \mathbf{1}^\top a = m \right\}.$$

The trace fixes total weight and the lower bound ensures positive definiteness. A mask $\xi \in \{0,1\}^m$ selects k coordinates uniformly, $1 \le k < m$, with $\pi = k/m$. As an index set, $\xi = \{j : \xi_j = 1\}$. Define

$$d^2_{a,\xi}(x_i, x_j) = \frac{1}{\pi} \sum_{\ell=1}^{m} \xi_\ell a_\ell (x_{i\ell} - x_{j\ell})^2. \tag{2}$$

The factor π^{-1} makes this unbiased for the full squared distance.

Step 2: Prediction loss. For bandwidth $h > 0$, define

$$K_{ij}(a, h; \xi) = \begin{cases} \exp\{-d^2_{a,\xi}(x_i, x_j)/(2h^2)\}, & i \neq j, \\ 0, & i = j. \end{cases} \tag{3}$$

For $n \ge 3$, Proposition 12 gives a unique positive symmetric scaling $u(a, h; \xi)$ with

$$W(a, h; \xi) = \operatorname{diag}(u) K(a, h; \xi) \operatorname{diag}(u), \qquad W(a, h; \xi)\mathbf{1} = \mathbf{1}. \tag{4}$$

The affinity is undirected with unit row sums (Sinkhorn, 1964; Sinkhorn & Knopp, 1967; Knight, 2008). Let $\xi^c = \{j : \xi_j = 0\}$ and let X_{ξ^c} contain the $m - k$ omitted columns. Their neighbor-prediction loss is

$$\ell_n(a, h; \xi) = \frac{1}{n|\xi^c|} \left\| \{I_n - W(a, h; \xi)\} X_{\xi^c} \right\|_F^2. \tag{5}$$

Here $\|\cdot\|_F$ is the Frobenius norm. The zero diagonal excludes self-prediction. The omitted columns do not enter the distance. A coordinate thus cannot reduce this loss by making its neighbors agree on its own values, and Proposition 1 shows that a kernel on nuisance coordinates that are independent of the signal and leave the conditional mean of the predicted quantities unchanged cannot improve the population prediction. A prespecified response matrix $V \in \mathbb{R}^{n \times q_V}$ replaces X_{ξ^c}, with denominator nq_V and $\xi = \mathbf{1}$, $\pi = 1$, because V is excluded from the distance.

Step 3: Metric estimation. Partition $N = Rb$ observations into disjoint blocks $I_1, \ldots, I_R$ of size b, with independent masks ξ_r. For ridge $\rho \ge 0$ and a finite bandwidth set $\mathcal{H}$, define

$$\widehat{\mathcal{R}}_\rho(a, h) = \frac{1}{R} \sum_{r=1}^{R} \ell_{I_r}(a, h; \xi_r) + \frac{\rho}{2m} \|a - \mathbf{1}\|_2^2, \qquad a \in \mathcal{A}_\delta, \quad h \in \mathcal{H}. \tag{6}$$

Here ℓ_{I_r} is (5) on block I_r. The ridge favors equal weights and belongs to the population target. We estimate weights and bandwidth with objective error $\eta_R \ge 0$:

$$\widehat{\mathcal{R}}_\rho(\widehat{a}, \widehat{h}) \le \inf_{a \in \mathcal{A}_\delta,\, h \in \mathcal{H}} \widehat{\mathcal{R}}_\rho(a, h) + \eta_R. \tag{7}$$

Differentiating (4) requires one positive definite solve per mask. Sequential quadratic programming gives feasible candidates; Lipschitz branch-and-bound or verified curvature certifies η_R. Section F gives a screening rule that fixes excluded weights at δ on a face containing a global minimizer.

For threshold $t_N \in (0, 1-\delta)$, retain weights above $\delta + t_N$, remove the other features, and normalize retained weights to average one:

$$\widehat{S} = \{j : \widehat{a}_j \ge \delta + t_N\}, \qquad \widehat{a}_j^{\mathrm{rep}} = \frac{|\widehat{S}| \widehat{a}_j}{\sum_{\ell \in \widehat{S}} \widehat{a}_\ell} \mathbf{1}\{j \in \widehat{S}\}. \tag{8}$$

Here $\mathbf{1}\{\cdot\}$ denotes an indicator. The set is nonempty because fitted weights average to one.

Step 4: Spectral coordinates. The selected weights define a graph distinct from W. For intrinsic dimension $d \ge 2$, choose a nonincreasing Lipschitz kernel φ supported on $[0, 1]$, with $\varphi(1/2) > 0$ and $\int_{\mathbb{R}^d} \varphi(\|u\|_2)\, du = 1$. Write $d^2_w(x, x') = \sum_j w_j (x_j - x'_j)^2$ for $w \in [0, \infty)^m$. On n representation observations, let $\widehat{r}_i$ be the k_nth-neighbor distance under $\widehat{a}^{\mathrm{rep}}$ and $\widehat{r}_{ij} = \widehat{r}_i \vee \widehat{r}_j$. Here $\vee, \wedge$ mean maximum and minimum, and tr means trace. Define

$$\widehat{A}_{ij} = \varphi\left\{ \frac{d_{\widehat{a}^{\mathrm{rep}}}(x_i, x_j)}{\widehat{r}_{ij}} \right\}, \quad i \neq j, \qquad \widehat{A}_{ii} = 0, \qquad \widehat{D} = \operatorname{diag}(\widehat{A}\mathbf{1}), \qquad \widehat{L} = \widehat{D} - \widehat{A}.$$

Zero radii use a positive floor tending to zero. The coordinate matrix $\widehat{Y} \in \mathbb{R}^{n \times p}$ solves

$$\widehat{Y} \in \operatorname*{arg\,min}_{Y^\top Y = I_p,\, Y^\top \mathbf{1} = 0} \mathrm{tr}(Y^\top \widehat{L} Y). \tag{9}$$

The graph may use the metric observations. Theorem 19 requires $k_n/n \to 0$ and $k_n^{1+4/d} / \{n^{4/d} \log n\} \to \infty$. Proposition 39 gives choices for k_n; Section R.1 avoids dense distance storage. Proposition 48 bounds perturbations of the fixed-graph extension to new observations.

4 MAIN THEORETICAL RESULTS

4.1 THE POPULATION PREDICTIVE METRIC

Let $X_\xi = (X_j : \xi_j = 1) \sim P_\xi$ and let $k_{a,h,\xi}$ be (3) before diagonal deletion. Bounded measurements (Condition 1) bound the kernel above and away from zero, giving a unique population scaling (Marshall & Coifman, 2019):

$$u_{a,h,\xi}(x) \int k_{a,h,\xi}(x,x')u_{a,h,\xi}(x')\,dP_\xi(x') = 1.$$

For square-integrable functions f, the population bistochastic smoother is

$$T_{a,h,\xi}f(x) = u_{a,h,\xi}(x) \int k_{a,h,\xi}(x,x')u_{a,h,\xi}(x')f(x')\,dP_\xi(x').$$

Unlike a row-normalized smoother, which is self-adjoint with respect to a degree-weighted distribution, $T_{a,h,\xi}$ is a positive self-adjoint contraction on $L^2(P_\xi)$. Lemma 5 uses this property for an inverse bound that does not deteriorate as h decreases, and it shows that, under the conditions of Theorem 9, for scalar C^6 functions f, $(T_{a,h,\xi}f - f)/h^2$ approaches one half of the density-weighted Laplace–Beltrami operator applied to f. For square-integrable X, define $m_\xi(x) = \mathbb{E}(X_{\xi^c} \mid X_\xi = x)$, apply $T_{a,h,\xi}$ coordinatewise, and average over masks using $\mathbb{E}_\xi$. The population risk is

$$\mathcal{R}_{\infty,\rho}(a,h) = \mathbb{E}_\xi \left[\frac{1}{|\xi^c|} \mathbb{E}\|X_{\xi^c} - T_{a,h,\xi}m_\xi(X_\xi)\|_2^2 \right] + \frac{\rho}{2m}\|a - \mathbf{1}\|_2^2. \tag{10}$$

For a square-integrable response view V, use $m_\xi(x) = \mathbb{E}(V \mid X_\xi = x)$ and replace $|\xi^c|$ by q_V. Write (a_0, h_0) for a minimizer.

In model (1), let T and T_S be the full-feature and signal-feature population operators, and let $\|\cdot\|_{L^2(P_S)}$ be the root mean squared norm under P_S. The following proposition identifies the nuisance kernel's effect.

Proposition 1. *Suppose $P_{S,U} = P_S \otimes P_U$ and $k\{(s,u),(s',u')\} = k_S(s,s')k_U(u,u')$. If both factor kernels admit symmetric bistochastic scalings, then*

$$Tg(s,u) = T_S g(s), \qquad \mathbb{E}\|V - Tg(S,U)\|_2^2 = \mathbb{E}\|\zeta\|_2^2 + \|g - T_S g\|_{L^2(P_S)}^2.$$

The two quantities do not depend on k_U.

Proposition 1 excludes improvement from the independent nuisance kernel. Transferring its trace mass to signal yields boundary nuisance weights when regularized risk decreases. Proposition 25 verifies uniqueness, curvature, and boundary multipliers; Corollary 10 reduces these conditions to one-dimensional risk inequalities for uniform masks in a repeated-pair model.

4.2 METRIC ESTIMATION

Assume independent copies of X, or (X, V) for a response view. Coordinate masks are independent of the observations, include and omit features, and have common inclusion probability $\pi \geq \pi_- > 0$. Blocks have size $b \geq 3$; a response view has $q_V \geq 1$ coordinates. We bound measurements by $B > 0$ and normalize bandwidth by $\sqrt{m}$.

Condition 1 (Bounded measurements and prediction scale)**.** $\|X\|_\infty \leq B$ *almost surely. For a prespecified response view,* $\|V\|_\infty \leq B$ *almost surely. For constants* $0 < \tau_- < \tau_+ < \infty$*,*

$$\tau_- \leq h/\sqrt{m} \leq \tau_+ \quad \textit{for every } h \in \mathcal{H}.$$

Condition 1 and these design choices ensure that each off-diagonal kernel entry is bounded away from zero. This bound allows Lemma 4 to derive bounds for the scaling and its derivatives using the bounds B, π_-, and τ_-. For a fixed prescribed view with $\xi = \mathbf{1}$, Corollary 17 relaxes the response boundedness to a finite second moment and, under its stated conditions, provides a $\sqrt{N}$ limit distribution for the metric estimator with a consistent covariance estimator.

We next assume that the population predictive risk identifies the optimal metric and bandwidth.

Condition 2 (Predictive identification). *The minimizer (a_0, h_0) is unique. For constants $r_0, \kappa, \Delta_h > 0$,*

$$\mathcal{R}_{\infty,\rho}(a, h_0) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \geq \kappa \|a - a_0\|_2^2, \qquad \|a - a_0\|_2 \leq r_0,$$

and

$$\inf_{a \in \mathcal{A}_\delta,\, h \neq h_0} \{\mathcal{R}_{\infty,\rho}(a, h) - \mathcal{R}_{\infty,\rho}(a_0, h_0)\} \geq \Delta_h.$$

Condition 2 permits the optimal feature weights to lie on the lower boundary δ. Propositions 25 and 27 verify these metric and bandwidth conditions for product-view models.

Let $\mathcal{R}_{b,\rho}$ be the expected block risk. Proposition 31 gives $\sup_{a,h} |\mathcal{R}_{b,\rho} - \mathcal{R}_{\infty,\rho}| \leq C_{\mathrm{bl}}\{m \log(mb)/b\}^{1/2}$. The following theorem separates sampling, block approximation, and optimization errors for $0 < \alpha < 1$.

Theorem 2. *Suppose Condition 1 holds. There is a constant C_1, depending only on B, π_-, and τ_-, such that, with probability at least $1 - \alpha$,*

$$\sup_{a \in \mathcal{A}_\delta,\, h \in \mathcal{H}} |\widehat{\mathcal{R}}_\rho(a, h) - \mathcal{R}_{b,\rho}(a, h)| \leq C_1 \sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}}.$$

If m is fixed, every estimator satisfying (7) *obeys*

$$\mathcal{R}_{\infty,\rho}(\widehat{a}, \widehat{h}) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \leq 2C_1 \sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}} + 2C_{\mathrm{bl}} \sqrt{\frac{m \log(mb)}{b}} + \eta_R.$$

For ratios to the stated rate, $O_{\mathbb{P}}$ denotes boundedness in probability, $o_{\mathbb{P}}$ convergence to zero in probability, and O, o their deterministic versions. Theorem 2 gives population-risk consistency for fixed m as $R, b \to \infty$ and $\eta_R \to 0$ in probability; fixed b targets finite-block risk. With Condition 2, fixed $m, \mathcal{H}$ and mask distribution, and J independent masks per block, Theorem 6 gives $\mathbb{P}(\widehat{h} = h_0) \to 1$ and metric error $O_{\mathbb{P}}\{(Rb)^{-1/2} + b^{-1} + (RJ)^{-1/2} + \eta_R^{1/2}\}$. The terms measure observation variation, block bias, mask variation, and optimization error.

4.3 The selected spectral representation

Define the active coordinates and their separation from the lower boundary by

$$S_0 = \{j : a_{0j} > \delta\}, \qquad \beta_0 = \min_{j \in S_0} (a_{0j} - \delta),$$

The population representation metric is

$$a_{0j}^{\mathrm{rep}} = \frac{|S_0| a_{0j}}{\sum_{\ell \in S_0} a_{0\ell}} \mathbf{1}\{j \in S_0\}.$$

To analyze the spectral convergence, we model the selected measurements on a d-dimensional manifold $\mathcal{M}$ with a latent state $Z \sim P$. We put

$$\sigma_\varphi = \int_{\mathbb{R}^d} u_1^2 \varphi(\|u\|_2)\, du.$$

Let $\phi_{S_0} : \mathcal{M} \to \mathbb{R}^{|S_0|}$ denote the selected observation map. For a metric w supported on S_0, the embedding is

$$F_w(z) = \mathrm{diag}(w_{S_0})^{1/2} \phi_{S_0}(z).$$

Let g_w be the Riemannian metric induced by this embedding, and let p_w denote the sampling density with respect to its volume measure. For a smooth function f, the continuum operator of the symmetric nearest-neighbor graph is

$$\mathcal{L}_w f = -\frac{\sigma_\varphi}{2p_w} \mathrm{div}_{g_w}\{p_w^{1-2/d} \nabla_{g_w} f\}. \tag{11}$$

Here ∇_{g_w}, div_{g_w}, and dV_{g_w} denote the gradient, divergence, and volume measure for g_w. Because $dP = p_w\, dV_{g_w}$, this operator is self-adjoint and nonnegative on the square-integrable space $L^2(P)$. The unweighted selected observation map induces $g_{\mathbf{1},S_0}$. We impose the following standard manifold conditions on the selected coordinates.

Condition 3 (Selected manifold model)**.** *$|S_0|$ and d are fixed, $d \geq 2$, $\mathcal{M}$ is a compact connected C^∞ manifold without boundary, and $X_{S_0} = \phi_{S_0}(Z)$ for a C^∞ embedding ϕ_{S_0}. The distribution of Z has a C^6 density bounded above and away from zero with respect to the Riemannian volume measure of $g_{\mathbf{1},S_0}$. The coordinates $X_{S_0^c}$ are bounded, with no further restriction.*

Condition 3 gives uniform density, curvature, injectivity-radius, and ellipticity bounds over compact positive metric families. Excluded coordinates need not be functions of the latent state Z.

Let $0 = \lambda_1(w) < \lambda_2(w) \leq \cdots$ denote the eigenvalues of (11). We let $\mathcal{E}_p(w)$ be the span of eigenfunctions numbered 2 through $p+1$, and we define the eigengap as

$$\gamma_p = \lambda_2(a_0^{\mathrm{rep}}) \wedge \{\lambda_{p+2}(a_0^{\mathrm{rep}}) - \lambda_{p+1}(a_0^{\mathrm{rep}})\}.$$

For an orthonormal basis $\psi_2, \ldots, \psi_{p+1}$ of $\mathcal{E}_p(a_0^{\mathrm{rep}})$, let $\Phi_0 = (\psi_{j+1}(Z_i))_{i\leq n,\, j\leq p}$, where Z_i is the latent state of representation observation i. We normalize the estimated coordinates $\widehat{Y}$ as in (9), and an orthogonal $p \times p$ matrix O aligns the coordinates. For J independent masks per block, the estimator minimizes the corresponding average of (6) with the objective tolerance η_R. For a deterministic sequence $\bar{\eta}_R > 0$ satisfying $\eta_R = O_{\mathbb{P}}(\bar{\eta}_R)$, we define the overall metric estimation error rate as

$$\omega_{R,J,b} = \left\{\frac{\log(2m|\mathcal{H}|)}{R}\right\}^{1/4} + \left\{\frac{\log(2m|\mathcal{H}|)}{RJ}\right\}^{1/4} + \left\{\frac{m\log(mb)}{b}\right\}^{1/4} + \bar{\eta}_R^{1/2}.$$

Theorem 2 and Corollary 3, combined with local quadratic growth, bound the metric error by $\omega_{R,J,b}$. Theorem 6 provides a sharper rate under fixed-dimensional conditions. The following corollary establishes correct support recovery and spectral convergence when this metric error is sufficiently small.

Corollary 1. *Suppose Conditions 1, 2, and 3 hold, m is fixed, and $\gamma_p > 0$. If*

$$\omega_{R,J,b} = o(t_N), \qquad t_N = o(\beta_0),$$

and k_n satisfies Theorem 19, then

$$\mathbb{P}(\widehat{S} = S_0) \to 1, \qquad \inf_{O^\top O = I_p} \left\|\widehat{Y} - n^{-1/2}\Phi_0 O\right\|_F = O_{\mathbb{P}}\left\{\frac{(k_n/n)^{1/d} + \omega_{R,J,b}}{\gamma_p}\right\} = o_{\mathbb{P}}(1).$$

Corollary 1 separates graph approximation from metric estimation, with both errors divided by the eigengap. Thresholding removes excluded coordinates from the limiting geometry. Uniform extension of the bounds of Calder & García Trillos (2022) permits the same observations to estimate the metric and construct the graph; related graph-cut limits appear in García Trillos (2019).

Additional results delimit the target of the label-free criterion. Let ρ_h be the ridge weight at bandwidth h. If $\rho_h/h^4 \to \rho_\star \geq 0$ as $h \downarrow 0$, Theorem 9 gives the conditional-mean roughness limit for smooth embedded views. With mutually independent nuisance coordinates independent of the signal, Theorem 10 gives this limit on the signal manifold under its embedding conditions, after multiplying the excess over irreducible risk by h^{-4}. Every limit of the population minimizers minimizes this roughness plus the limiting ridge penalty. Under the conditions of Proposition 36, selection scores and unpenalized risk curvature are bounded by C/m for a fixed number of signal coordinates, where C does not depend on m. Theorem 18 gives support recovery with increasing m for a prespecified response view and fixed risk weights. Section A.2 gives sharper risk bounds, Section G gives estimation of the roughness minimizer and observation-level inference, and Section B bounds the prediction error for a later response.

4.4 Dependence between selected and excluded measurements

We next study departures from independence. Write $\mathcal{R}^P_{\infty,\rho}$ for risk under the complete-observation distribution P, with fixed masks, penalty, and bandwidth set. Let P_0 have unique minimizer (a_0, h_0), support S_0, and separation β_0. Define $d_{\mathrm{TV}}(P, P_0) = \sup_A |P(A) - P_0(A)|$ over measurable sets.

Support stability requires a positive first-order cost for increasing an excluded weight. With $\partial_j = \partial/\partial a_j$ and trace multiplier λ_0, strict complementarity means

$$\partial_j \mathcal{R}^{P_0}_{\infty,\rho}(a_0, h_0) + \lambda_0 = 0 \quad (j \in S_0), \qquad \min_{j \notin S_0}\{\partial_j \mathcal{R}^{P_0}_{\infty,\rho}(a_0, h_0) + \lambda_0\} > 0. \tag{12}$$

Table 1: Mean label-free neighborhood recovery over 100 data sets with 48 independent nuisance coordinates.

	Surface			Torus		
Method	Trust.	Acc.	Error	Trust.	Acc.	Error
PALE	0.9954	0.972	0.160	0.9992	0.953	0.069
Prediction, all features	0.9888	0.925	0.244	0.9975	0.936	0.137
Isotropic LE	0.8061	0.676	0.828	0.9650	0.828	0.496
PCA	0.5265	0.301	1.275	0.9963	0.930	0.179
SSFS	0.7415	0.579	0.886	0.5021	0.252	1.389
SLAPS-DAE FP	0.8158	0.686	0.781	0.9792	0.861	0.372
SLAPS-DAE MLP	0.7663	0.625	0.740	0.8246	0.636	0.796

Omit the minimum when S_0 contains all coordinates. Propositions 25 and 29 verify this population condition. The following theorem establishes support stability.

Theorem 3. *Suppose Conditions 1 and 2 hold under P_0, and* (12) *holds. Let m, $\mathcal{H}$, ρ, δ, and the mask distribution be fixed. There exist $\epsilon_0, C > 0$ such that, for every P satisfying the same measurement bound and $\epsilon = d_{\mathrm{TV}}(P, P_0) \le \epsilon_0$, the population minimizer (a_P, h_P) is unique and*

$$h_P = h_0, \qquad \{j : a_{Pj} > \delta\} = S_0, \qquad \|a_P - a_0\|_2 \le C\epsilon.$$

Condition 2 holds at (a_P, h_0) with constants uniform over these distributions.

Theorem 3 preserves boundary weights exactly while active weights change by $O(\epsilon)$. The estimator need not know P_0 or ϵ. Constants depend on reference curvature, bandwidth gap, and complementarity margin, and are not uniform as dimension increases or bandwidth tends to zero. Without strict complementarity, support stability is not guaranteed.

Corollary 2 adds this perturbation error to spectral estimation when selected measurements retain their marginal distribution. Section A.1 proves the results and gives a dependent model preserving the conditional response mean.

5 Numerical results

We assess feature selection, neighborhood recovery, and prediction, with 100 replications per reported simulation setting.

The surface and torus designs assess neighborhood recovery without labels or a response view, with 48 independent nuisance coordinates. Table 1 reports mean trustworthiness with ten neighbors (Trust.), seven-nearest-neighbor accuracy for latent quadrants or sectors (Acc.), and aligned distance from the signal-only graph eigenspace (Error). Mean trustworthiness is 0.9954 and 0.9992 for PALE, compared with 0.8061 and 0.9650 for isotropic LE and 0.8158 and 0.9792 for SLAPS-DAE FP. PALE also outperforms spectral self-supervised feature selection (SSFS) (Segal et al., 2024). SLAPS-DAE uses a denoising autoencoder with fully parameterized adjacency (FP) or a multilayer-perceptron adjacency generator (MLP). In the first of the 100 surface data sets, PALE preserves the four latent quadrants, isotropic coordinates distort them, and PCA mixes them (Figure 1). Excluding target features from prediction distances reduces reference-eigenspace error from 0.244 to 0.160 and from 0.137 to 0.069 relative to prediction of all features. Table 2 gives standard deviations and continuity. These designs assess neighborhood recovery. Their global objective bounds do not certify the asymptotic optimization tolerance.

On the fixed handwritten-digits data, 100 label-free block/mask allocations reduce pixel prediction error from 0.7523 to 0.6301 at the same bandwidth. Classification accuracy is 97.603%, compared with 98.275% for isotropic LE and 97.559% for Laplacian Score with LE (Table 21). This transductive assessment, after exploratory design selection, shows that lower omitted-feature prediction error need not improve class separation.

The exact response-view sphere model of Theorem 8 identifies population weights and the conditional mean represented by the first three coordinates. Across six dimension and sample-size settings

with 100 independent data sets each, exact support recovery occurs in 97 to 99 data sets at $N = 1024$ and all 100 at $N = 4096$ in every dimension (Table 6). Every fit strictly satisfies the prescribed global objective tolerance. These linear-response experiments use $\tau = 300$, making the prediction kernel nearly constant. At $m = 192$, mean spectral and conditional-mean prediction errors decrease from 0.1448 to 0.0466 and from 0.0338 to 0.0023, respectively. At $N = 4096$, prediction error is 0.0010 for the oracle-support graph, 0.0527 for partial least squares (PLS), and 0.9829 for isotropic eigenmaps (Table 7). Across four dependence levels, including independence, the dependent study recovers the exact signal set in all 400 samples at $N = 4096$, with mean prediction error between 0.0020 and 0.0023 (Tables 8 and 9).

To assess nonlinear prediction, we use quadratic responses whose Pearson correlations with every coordinate equal zero. At $\tau = 0.5$, PALE recovers the signal coordinates in all 100 independent data sets at each $\delta \in \{0.01, 0.02, 0.1\}$. Mean conditional-mean error is 0.0140, compared with 1.0012 for correlation screening, 1.0139 for isotropic LE, and 0.0144 for diagonal MLKR (Table 20). These fits have no global objective-gap certificate. Proposition 50 proves a finite-bandwidth predictive advantage in this model, with support identification restricted to the group-isotropic metric family.

In CITE-seq and independently collected peripheral blood mononuclear cell (PBMC) data (Stoeckius et al., 2017; 10x Genomics, 2018), we predict RNA from proteins, excluding the RNA response panel from the distance. With ten spectral coordinates, PALE reduces paired mean squared error by 0.0023 and 0.0175 relative to isotropic eigenmaps and 0.0272 and 0.0282 relative to PLS, respectively. All four observation-bootstrap intervals exclude zero (Table 12). In PBMC, the metric retains 10.71 of 17 antibody measurements on average and excludes all three isotype controls in every fit. Repeated partitions describe variation within each collection. MLKR spectral has lower error at both common dimensions in both collections (Table 11); the gains over isotropic geometry do not establish superiority over response-based metric learning.

Table 13 reports time for metric fitting, certification, graph construction, eigendecomposition, and prediction. With the same block implementation, PALE is faster than isotropic eigenmaps at every sample size, including metric fitting: 2.357 versus 9.522 seconds at $n = 8192$. Exact blockwise distances reduce memory and time without changing the graph or predictions, although search remains quadratic. Table 4 reports coverage of 0.97 and 0.94 for nominal 95 percent metric intervals when the response has an infinite fourth moment.

Section J evaluates a separate posterior-moment extension for classification with measurements available only during training.

6 DISCUSSION

Under the stated conditions, we control final spectral error even when metric estimation and graph construction use the same observations. Under the conditions of Theorem 3, the population predictive support remains exactly stable under small changes in the observation distribution measured in total variation.

We recommend using a scientifically prespecified response view when available: Theorem 18 extends support recovery, under its conditions, to increasing feature dimension. Without such a view, the masked criterion removes independent nuisance coordinates at fixed dimension under the conditions of Section 4.

The label-free criterion can preserve nuisance variation that predicts held-out measurements. Identifying and removing predictive but scientifically irrelevant features remains unresolved.

AI USE STATEMENT

We used generative AI tools to assist with code development, literature searches, and manuscript editing. The authors take full responsibility for the contents of this submission.

ETHICS STATEMENT

For our real-data analyses, we use the public Prokaryotic and Outdoor Scene multiview datasets, UCI handwritten digits, molecular count matrices from GEO accession GSE100866, and the 10x Genomics PBMC dataset. We did not recruit participants or collect new human data, and the molecular analyses do not attempt to identify donors. We study this method strictly for representation and prediction, and we do not propose it for clinical decisions.

REPRODUCIBILITY STATEMENT

The Supplementary Materials provide proofs of the principal results in Section A, further statistical statements in Section G, and auxiliary proofs in Sections H and K. Section R specifies the scaling, differentiation, and optimization certificates. Sections I, J, and U specify the data distributions, comparison methods, uncertainty calculations, and evaluation designs. The accompanying code records the 100-replication outputs, separate development and production settings, data sources, registered confirmation predictions, local optimization diagnostics, and available global objective bounds.

REFERENCES

10x Genomics. 10k PBMCs from a healthy donor: Gene expression with a panel of TotalSeq-B antibodies. Public dataset, Cell Ranger 3.0.0, 2018. URL https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-gene-expression-and-cell-surface-protein-3-standard-3-0-0. Published November 19, 2018. Accessed August 31, 2026.

Othmane Aboussaad, Adam Miraoui, Boumediene Hamzi, and Houman Owhadi. Adaptive kernel selection for kernelized diffusion maps. arXiv:2604.18402v1, 2026. URL https://arxiv.org/abs/2604.18402v1.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533. PMLR, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Mikhail Belkin and Partha Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems 19*, pp. 129–136. MIT Press, 2007.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008. doi: 10.1016/j.jcss.2007.08.006.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs. *Applied and Computational Harmonic Analysis*, 60:123–175, 2022. doi: 10.1016/j.acha.2022.02.004.

Xiuyuan Cheng and Boris Landa. Bi-stochastically normalized graph Laplacian: Convergence to manifold Laplacian and robustness to outlier noise. *Information and Inference: A Journal of the IMA*, 13(4):iaae026, 2024. doi: 10.1093/imaiai/iaae026.

Xiuyuan Cheng and Nan Wu. Eigen-convergence of Gaussian kernelized graph Laplacian by manifold heat interpolation. *Applied and Computational Harmonic Analysis*, 61:132–190, 2022. doi: 10.1016/j.acha.2022.06.003.

Xiuyuan Cheng, Yixuan Tan, and Nan Wu. Improved convergence rate of kNN graph Laplacians: Differentiable self-tuned affinity. *Applied and Computational Harmonic Analysis*, 86:101919, 2027. doi: 10.1016/j.acha.2026.101919.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning Laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016. doi: 10.1109/TSP.2016.2602809.

Hilmi E. Egilmez, Eduardo Pavez, and Antonio Ortega. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017. doi: 10.1109/JSTSP.2017.2726975.

Bahare Fatemi, Layla El Asri, and Seyed Mehran Kazemi. SLAPS: Self-supervision improves structure learning for graph neural networks. In *Advances in Neural Information Processing Systems 34*, pp. 22667–22681. Curran Associates, Inc., 2021.

Fatemeh Fathinezhad, Peyman Adibi, and Bijan Shoushtarian. Soft laplacian eigenmap: A novel regularized graph-based method for dimensionality reduction with local geometry preservation. *Advances in Data Analysis and Classification*, 2025. doi: 10.1007/s11634-025-00658-1.

Nicolás García Trillos. Variational limits of k-NN graph-based functionals on data clouds. *SIAM Journal on Mathematics of Data Science*, 1(1):93–120, 2019. doi: 10.1137/18M1188999.

Nicolás García Trillos and Dejan Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018. doi: 10.1016/j.acha.2016.09.003.

Peter Hall, Qi Li, and Jeffrey S. Racine. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics*, 89(4):784–789, 2007. doi: 10.1162/rest.89.4.784.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. From graphs to manifolds: Weak and strong pointwise consistency of graph Laplacians. In *Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pp. 470–485. Springer, 2005. doi: 10.1007/11503415_32.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. GraphMAE2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM Web Conference 2023*, pp. 737–746, 2023. doi: 10.1145/3543507.3583379.

Jiayan Jiang, Bo Wang, and Zhuowen Tu. Unsupervised metric learning by self-smoothing operator. In *2011 International Conference on Computer Vision*, pp. 794–801, 2011. doi: 10.1109/ICCV.2011.6126318.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 920–929, 2016.

Philip A. Knight. The Sinkhorn–Knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. doi: 10.1137/060659624.

John Lafferty and Larry Wasserman. Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008. doi: 10.1214/009053607000000811.

Boris Landa and Xiuyuan Cheng. Robust inference of manifold density and geometry by doubly stochastic scaling. *SIAM Journal on Mathematics of Data Science*, 5(3):589–614, 2023. doi: 10.1137/22M1516968.

Ann B. Lee and Rafael Izbicki. A spectral series approach to high-dimensional nonparametric regression. *Electronic Journal of Statistics*, 10(1):423–463, 2016. doi: 10.1214/16-EJS1112.

Yixin Liu, Yu Zheng, Daokun Zhang, Hongxu Chen, Hao Peng, and Shirui Pan. Towards unsupervised deep graph structure learning. In *Proceedings of the ACM Web Conference 2022*, pp. 1392–1403, 2022. doi: 10.1145/3485447.3512186.

Nicholas F. Marshall and Ronald R. Coifman. Manifold learning with bi-stochastic kernels. *IMA Journal of Applied Mathematics*, 84(3):455–482, 2019. doi: 10.1093/imamat/hxy065.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pp. 849–856. MIT Press, 2002.

Daniel Segal, Ofir Lindenbaum, and Ariel Jaffe. Spectral self-supervised feature selection. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=t0EJiOd9Lg`.

Amit Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006. doi: 10.1016/j.acha.2006.03.004.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. doi: 10.2140/pjm.1967.21.343.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017. doi: 10.1038/nmeth.4380.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. doi: 10.1007/s11222-007-9033-z.

Bo Wang and Zhuowen Tu. Affinity learning via self-diffusion for image segmentation and clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2312–2319, 2012. doi: 10.1109/CVPR.2012.6247942.

Shulei Wang. Self-supervised metric learning in multi-view data: A downstream task perspective. *Journal of the American Statistical Association*, 118(544):2454–2467, 2023. doi: 10.1080/01621459.2022.2057317.

Shulei Wang. Augmentation invariant manifold learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):978–1000, 2025. doi: 10.1093/jrsssb/qkaf003.

Kilian Q. Weinberger and Gerald Tesauro. Metric learning for kernel regression. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 612–619, 2007.

Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010. doi: 10.1198/jasa.2010.tm09415.

Jiayi Xu, Qirui Huang, Yang Zhao, Jihong Pei, and Xuan Yang. Adaptive local neighborhood window Laplacian eigenmaps. In *Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition*, pp. 1–7, 2024. doi: 10.1145/3704323.3704335.

Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13:1–26, 2012.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pp. 1601–1608. MIT Press, 2005.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Jinfeng Zhuang, Jialei Wang, Steven C. H. Hoi, and Xiangyang Lan. Unsupervised multiple kernel learning. In *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pp. 129–144, 2011.

SUPPLEMENTARY MATERIALS

A PROOFS OF THE PRINCIPAL RESULTS

The main results separate predictive identification, estimation error, and spectral approximation. We present their proofs first. Sharper bounds, along with the posterior-moment and continuous-model results, follow. Sections H and K provide the auxiliary scaling and uniform graph estimates.

Proof of Theorem 1. Partition the n vertices into $p+1$ groups of size s. Let

$$B = \operatorname{diag}(B_s, \dots, B_s), \qquad B_s = \frac{J_s - I_s}{s-1}, \qquad C = \frac{J_n - I_n}{n-1},$$

where J_k is the $k \times k$ all-ones matrix. Define

$$W_\varepsilon = (1-\varepsilon)B + \varepsilon C.$$

Both B and C are symmetric, bistochastic, and have zero diagonal. Every off-diagonal entry of W_ε is positive.

The matrices B and C commute because $B\mathbf{1} = \mathbf{1}$ and B is symmetric. On the constant vector, both have eigenvalue one. On the p-dimensional space of vectors that are constant within each block and sum to zero, B has eigenvalue one and C has eigenvalue $-1/(n-1)$. Thus $I_n - W_\varepsilon$ has eigenvalue $\frac{n}{n-1}\varepsilon$ with multiplicity p. On a within-block contrast, the eigenvalue of W_ε is $-\frac{1-\varepsilon}{s-1} - \frac{\varepsilon}{n-1}$. The corresponding Laplacian eigenvalue exceeds the block-contrast eigenvalue by $(1-\varepsilon)\left(1 + \frac{1}{s-1}\right)$, which is nonnegative. Proposition 9 now gives

$$\Phi_p(W_\varepsilon) = \frac{pn}{n-1}\varepsilon.$$

Every row has $s-1$ entries equal to

$$w_{\text{in}} = \frac{1-\varepsilon}{s-1} + \frac{\varepsilon}{n-1}$$

and $n-s$ entries equal to $w_{\text{out}} = \varepsilon/(n-1)$. Its entropy is

$$-(s-1)w_{\text{in}} \log w_{\text{in}} - (n-s)w_{\text{out}} \log w_{\text{out}}.$$

The second term tends to zero and the first tends to $\log(s-1)$. This completes the proof. □

Proof of Proposition 1. Let u_S and u_U be the population scaling functions of the two factor kernels. Their product satisfies the scaling equation for $k_S k_U$. Uniqueness gives the factorized bistochastic operator $T = T_S \otimes T_U$. Because $T_U\mathbf{1} = \mathbf{1}$,

$$Tg(s,u) = T_S g(s).$$

Write $V - T_S g(S) = \zeta + g(S) - T_S g(S)$. The inner-product cross term has expectation zero by $\mathbb{E}(\zeta \mid S, U) = 0$. Taking the squared Euclidean norm and expectation proves the risk identity. This completes the proof. □

For posterior-targeted prediction, teacher class scores replace the response view in (5). Let $e_C \in \mathbb{R}^K$ be the one-hot class vector, $\eta_X(X) = \mathbb{E}(e_C \mid X)$, and $\eta_{XV}(X,V) = \mathbb{E}(e_C \mid X, V)$. For square-integrable teacher scores $s_0(X,V)$, put $\mu_0(X) = \mathbb{E}\{s_0(X,V) \mid X\}$. Orthogonality gives, for square-integrable $g(X) \in \mathbb{R}^K$,

$$\mathbb{E}\|s_0(X,V) - g(X)\|_2^2 = \mathbb{E}\|s_0(X,V) - \mu_0(X)\|_2^2 + \mathbb{E}\|\mu_0(X) - g(X)\|_2^2.$$

If $s_0 = \eta_{XV}$, the tower property gives $\mu_0 = \eta_X$ almost surely. The second term then measures prediction of deployable class probabilities.

We next evaluate how estimating the posterior responses affects the predictive metric. Let G and $\widehat{\mathsf{G}}$ denote the ideal and estimated $N \times K$ response matrices, respectively, with rows in the probability

simplex. We write $\widehat{\mathcal{R}}_{\rho}^{\mathsf{G}}$ and $\widehat{\mathcal{R}}_{\rho}^{\widehat{\mathsf{G}}}$ for the corresponding empirical criteria, and we write $\mathcal{R}_{b,\rho}^{\mathsf{G}}$ and $\mathcal{R}_{\infty,\rho}^{\mathsf{G}}$ for the expected block criterion and the population risk. To separate the error sources, we define the root-mean-square score error r_G, the uniform empirical deviation D_N, and the block approximation error B_b as

$$r_G = \frac{\|\widehat{\mathsf{G}} - \mathsf{G}\|_F}{\sqrt{NK}}, \qquad D_N = \sup_{a,h} |\widehat{\mathcal{R}}_{\rho}^{\mathsf{G}}(a,h) - \mathcal{R}_{b,\rho}^{\mathsf{G}}(a,h)|, \qquad B_b = \sup_{a,h} |\mathcal{R}_{b,\rho}^{\mathsf{G}}(a,h) - \mathcal{R}_{\infty,\rho}^{\mathsf{G}}(a,h)|.$$

For a given optimization tolerance $\eta_N \geq 0$, the following theorem bounds the difference between the population risk of the estimated metric and the minimum population risk.

Theorem 4. *Let (a_0, h_0) minimize $\mathcal{R}_{\infty,\rho}^{\mathsf{G}}$. If $\widehat{\mathcal{R}}_{\rho}^{\widehat{\mathsf{G}}}(\widehat{a}, \widehat{h}) \leq \inf_{a,h} \widehat{\mathcal{R}}_{\rho}^{\widehat{\mathsf{G}}}(a,h) + \eta_N$, then*

$$\mathcal{R}_{\infty,\rho}^{\mathsf{G}}(\widehat{a}, \widehat{h}) - \mathcal{R}_{\infty,\rho}^{\mathsf{G}}(a_0, h_0) \leq 2D_N + 2B_b + 16r_G + 8r_G^2 + \eta_N.$$

Theorem 4 bounds the excess risk by terms representing score estimation (r_G), empirical sampling (D_N), block approximation (B_b), and optimization (η_N). When this bound tends to zero in probability, Condition 2 for the ideal posterior risk and compactness guarantee consistency for the ideal posterior metric. With probability tending to one, the metric error is at most a constant times the square root of the bound. Using this result, Theorem 19 and Corollary 7 bound the spectral and excess-classification errors under their respective conditions.

Proof of Theorem 4. For a matrix A, $\|A\|_{\mathrm{op}}$ denotes its largest singular value. Let $\mathbb{W}_{a,h}$ be block diagonal with the fitted bistochastic matrix on each prediction block, and put $\mathbb{A}_{a,h} = I_N - \mathbb{W}_{a,h}$. Every block is symmetric and bistochastic, which gives

$$\|\mathbb{W}_{a,h}\|_{\mathrm{op}} \leq 1, \qquad \|\mathbb{A}_{a,h}\|_{\mathrm{op}} \leq 2.$$

Write $E = \widehat{\mathsf{G}} - \mathsf{G}$. Because every row of G is in the probability simplex, $\|\mathsf{G}\|_F \leq \sqrt{N}$, while $\|E\|_F = \sqrt{NK}\, r_G$. Uniformly over (a,h),

$$\left|\widehat{\mathcal{R}}_{\rho}^{\widehat{\mathsf{G}}}(a,h) - \widehat{\mathcal{R}}_{\rho}^{\mathsf{G}}(a,h)\right| \leq \frac{2\|\mathbb{A}_{a,h}\mathsf{G}\|_F \|\mathbb{A}_{a,h}E\|_F + \|\mathbb{A}_{a,h}E\|_F^2}{NK} \leq 8r_G + 4r_G^2.$$

The ridge terms are equal and cancel.

Put $\vartheta = (a,h)$, $\widehat{\vartheta} = (\widehat{a}, \widehat{h})$, and $\vartheta_0 = (a_0, h_0)$. The approximate minimizing property gives

$$\begin{aligned}\mathcal{R}_{\infty,\rho}^{\mathsf{G}}(\widehat{\vartheta}) - \mathcal{R}_{\infty,\rho}^{\mathsf{G}}(\vartheta_0) &\leq 2B_b + 2D_N + 2\sup_{\vartheta} \left|\widehat{\mathcal{R}}_{\rho}^{\widehat{\mathsf{G}}}(\vartheta) - \widehat{\mathcal{R}}_{\rho}^{\mathsf{G}}(\vartheta)\right| + \eta_N \\ &\leq 2B_b + 2D_N + 16r_G + 8r_G^2 + \eta_N.\end{aligned}$$

This completes the proof. □

Proof of Theorem 2. The ridge terms cancel. For independent, identically distributed blocks, Theorem 5 uses the trace constraint and the C/b log-kernel Lipschitz bound to give

$$\sup_{a,h} |\widehat{\mathcal{R}}_{\rho}(a,h) - \mathcal{R}_{b,\rho}(a,h)| \leq C_1 \sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}}$$

with probability at least $1 - \alpha$.

If m is fixed, let the left side be D_R and let

$$B_b = \sup_{a,h} |\mathcal{R}_{b,\rho}(a,h) - \mathcal{R}_{\infty,\rho}(a,h)|.$$

The approximate minimizing property gives

$$\mathcal{R}_{\infty,\rho}(\widehat{a}, \widehat{h}) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \leq 2D_R + 2B_b + \eta_R.$$

Proposition 31 bounds B_b and proves the second assertion. This completes the proof. □

Proof of Corollary 1. Corollary 13 and $\eta_R = O_{\mathbb{P}}(\bar{\eta}_R)$ give

$$\|\widehat{a} - a_0\|_2 = O_{\mathbb{P}}(\omega_{R,J,b}).$$

The two threshold conditions and Corollary 19 give

$$\mathbb{P}(\widehat{S} = S_0) \to 1, \qquad \|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2 = O_{\mathbb{P}}(\omega_{R,J,b}).$$

The simultaneous form of Theorem 19 applies whether the graph sample is independent or equals the metric sample. Substitution of the representation-metric rate proves the stated order. This completes the proof. □

A.1 Distribution perturbations of the predictive metric

To establish the regularity needed for support stability, we first control the population smoother and its first two metric derivatives under perturbations of the complete-observation distribution. The following analysis does not assume smooth densities or independence among the measurement coordinates. Throughout this subsection, we fix the dimension, the mask distribution, and the prediction bandwidth set.

For a twice continuously differentiable function f on the metric domain, we define the norm $\|f\|_{C^2}$ as the maximum over the suprema of $|f|$, its first coordinate derivatives, and its second coordinate derivatives. Because the metric domain has a lower boundary, we define these boundary derivatives through the Gaussian-kernel extension to an open neighborhood of $\mathcal{A}_\delta$. The following lemma bounds all three quantities using only the measurement bound.

Lemma 1. *Suppose P and Q satisfy the same bound in Condition 1. With the same mask distribution, penalty, and bandwidth set,*

$$\max_{h \in \mathcal{H}} \|\mathcal{R}^P_{\infty,\rho}(\cdot, h) - \mathcal{R}^Q_{\infty,\rho}(\cdot, h)\|_{C^2} \le C d_{\mathrm{TV}}(P, Q).$$

The constant is independent of P and Q on this bounded observation space.

Lemma 1 controls the score and curvature, not merely the objective values. Under local quadratic growth, a bound on the risk alone gives a square-root perturbation bound but does not guarantee that the active set remains unchanged. The positive lower bound on the kernel is essential to the proof.

Proof of Lemma 1. Fix a mask and bandwidth. Let z denote a complete observation in a compact box $\mathcal{Z}$, and let $y(z)$ be the omitted measurements or the prescribed response, of dimension $q \ge 1$. Extend the observed-coordinate kernel to $\mathcal{Z} \times \mathcal{Z}$ by ignoring the response coordinates. On a fixed open neighborhood of $\mathcal{A}_\delta$, choose $0 < c_k < 1$ such that

$$c_k \le k_a(z, z') \le 1, \qquad \max_j \|\partial_j \log k_a\|_\infty \le C, \qquad \partial_{jl} \log k_a = 0.$$

Such a neighborhood exists because $\delta > 0$, $\mathbf{1}^\top a = m$, the measurements are bounded, and the prediction scales are bounded away from zero. Constants below are uniform over this neighborhood, the finite mask support, and $h \in \mathcal{H}$.

We give a scaling argument that also permits distributions supported on proper subsets of $\mathcal{Z}$. For a probability distribution H, define on $C(\mathcal{Z})$

$$\mathcal{F}_H(v)(z) = -\log \int k_a(z, z') e^{v(z')}\, dH(z').$$

For a function f, put $\operatorname{osc}(f) = \sup f - \inf f$. The derivative of $\mathcal{F}_H$ at v is minus a Markov operator whose row measures dominate $c_k e^v\, dH / \int e^v\, dH$. Hence

$$\operatorname{osc}\{D\mathcal{F}_H(v)[f]\} \le (1 - c_k) \operatorname{osc}(f), \qquad \operatorname{osc}\{\mathcal{F}_H(v) - \mathcal{F}_H(w)\} \le (1 - c_k) \operatorname{osc}(v - w).$$

The second bound follows by integrating the first along the segment from w to v. Because $\mathcal{F}_H(v + c) = \mathcal{F}_H(v) - c$, the map is a contraction on the complete space $C(\mathcal{Z})$ modulo constant functions, equipped with the oscillation norm. Its unique fixed class has a representative v satisfying $\mathcal{F}_H(v) = v + c$. Replacing v by $v + c/2$ gives a fixed point. Thus $u_H = e^v$ is the unique positive continuous symmetric scaling.

Put $A_H = \int u_H \, dH$. The scaling equation gives

$$(A_H)^{-1} \le u_H(z) \le (c_k A_H)^{-1}, \qquad 1 \le A_H \le c_k^{-1/2}, \qquad c_k^{1/2} \le u_H(z) \le c_k^{-1}.$$

Define $w_H(z,z') = u_H(z) k_a(z,z') u_H(z')$ and $T_H f(z) = \int w_H(z,z') f(z') \, dH(z')$. This is a Markov operator with stationary distribution H, and $w_H \ge c_k^2$. For $\beta = c_k^2/2$, write

$$T_H = \beta \Pi_H + (1-\beta) Q_H, \qquad \Pi_H f = \int f \, dH,$$

where Q_H is Markov and has stationary distribution H. On mean-zero functions, $\|T_H^r f\|_\infty \le (1-\beta)^r \|f\|_\infty$. Thus

$$(I + T_H)^{-1} f = \tfrac{1}{2} \Pi_H f + \sum_{r=0}^{\infty} (-T_H)^r (f - \Pi_H f), \qquad \|(I+T_H)^{-1}\|_{\infty \to \infty} \le \tfrac{1}{2} + 2/\beta.$$

The series converges in the maximum norm. This proves the required inverse bound on all of $C(\mathcal{Z})$.

Let $H_t = Q + t(P-Q)$, $0 \le t \le 1$, and $\mu = P - Q$. Write $v_t = \log u_{H_t}$ and suppress the subscripts a, t on w and T. The map

$$(v, a, t) \longmapsto e^{v(z)} \int k_a(z,z') e^{v(z')} \, dH_t(z') - 1$$

is smooth as a map into $C(\mathcal{Z})$. At a solution its derivative in v is $I + T$. The inverse bound and the implicit function theorem give joint differentiability in a, t. The local solutions agree by uniqueness. At either endpoint the same derivative argument uses a local signed-measure extension of the equation; all bounds are evaluated for $0 \le t \le 1$.

For a metric coordinate j, put $c_j = \partial_j \log k_a$ and $G_j(z,z') = c_j(z,z') + \partial_j v_t(z) + \partial_j v_t(z')$. Differentiation of the row equation gives

$$(I+T)\partial_j v_t = -\int w(\cdot, z') c_j(\cdot, z') \, dH_t(z'),$$
$$(I+T)\partial_{jl} v_t = -\int w(\cdot, z') G_j(\cdot, z') G_l(\cdot, z') \, dH_t(z').$$

The stochastic row sums and inverse bound imply $\|\partial_j v_t\|_\infty + \|\partial_{jl} v_t\|_\infty \le C$. In particular, w, its first two metric derivatives, and the corresponding integral operators have uniformly bounded maximum norms.

A dot denotes a derivative in t. With $b_t(z) = \int w(z,z') \, d\mu(z')$,

$$(I+T)\dot{v}_t = -b_t, \qquad \|b_t\|_\infty \le C d_{\mathrm{TV}}(P,Q), \qquad \|\dot{v}_t\|_\infty \le C d_{\mathrm{TV}}(P,Q).$$

Here the total mass of the variation measure of μ is $2 d_{\mathrm{TV}}(P,Q)$. Differentiating this equation once and twice in the metric gives

$$(I+T)\partial_j \dot{v}_t = -\partial_j b_t - (\partial_j T)\dot{v}_t,$$
$$(I+T)\partial_{jl}\dot{v}_t = -\partial_{jl} b_t - (\partial_{jl} T)\dot{v}_t - (\partial_j T)\partial_l \dot{v}_t - (\partial_l T)\partial_j \dot{v}_t.$$

Each derivative of b_t integrates a bounded derivative of w against μ. The preceding inverse and derivative bounds give

$$\max_{r \le 2} \max_{j_1, \dots, j_r} \|\partial_{j_1 \cdots j_r} \dot{v}_t\|_\infty \le C d_{\mathrm{TV}}(P,Q).$$

Differentiating $w(z,z') = e^{v_t(z) + v_t(z')} k_a(z,z')$ now gives the same bound for $\partial_{j_1 \cdots j_r} \dot{w}$, for $r \le 2$.

The population prediction can be written without a conditional density:

$$f_{a,t}(z) = \int w(z,z') y(z') \, dH_t(z').$$

It equals the population smoother of the conditional mean by Proposition 21. Its metric derivatives through order two are bounded componentwise. Because

$$\dot{f}_{a,t}(z) = \int \dot{w}(z,z') y(z') \, dH_t(z') + \int w(z,z') y(z') \, d\mu(z'),$$

the same mixed-derivative bound holds for $\dot{f}_{a,t}$. Put $\ell_{a,t}(z) = q^{-1}\|y(z) - f_{a,t}(z)\|_2^2$. The measurement bound implies, for every metric derivative ∂^r of order at most two,

$$\|\partial^r \ell_{a,t}\|_\infty \le C, \qquad \|\partial^r \dot{\ell}_{a,t}\|_\infty \le C d_{\mathrm{TV}}(P, Q).$$

The derivative in t of its expected loss is

$$\frac{d}{dt}\int \partial^r \ell_{a,t}\, dH_t = \int \partial^r \dot{\ell}_{a,t}\, dH_t + \int \partial^r \ell_{a,t}\, d\mu.$$

Its absolute value is at most $C d_{\mathrm{TV}}(P, Q)$. Integrate over $t \in [0, 1]$ and average over the fixed mask distribution. The ridge and its derivatives cancel. This completes the proof. □

Proof of Theorem 3. Write $F_P(a, h) = \mathcal{R}^P_{\infty,\rho}(a, h)$, $S = S_0$, and $I = S^c$. The constants in this proof do not depend on P within a sufficiently small total variation neighborhood of P_0. Lemma 1 gives

$$\max_{h\in\mathcal{H}} \|F_P(\cdot, h) - F_{P_0}(\cdot, h)\|_{C^2} \le C\epsilon.$$

Uniqueness and compactness imply a positive reference risk gap outside every fixed neighborhood of a_0 at h_0. Condition 2 gives a positive gap for every other bandwidth. Hence every minimizer under P has bandwidth h_0 and lies in an arbitrarily small fixed neighborhood of a_0 when ϵ is sufficiently small.

For symmetric matrices, $A \succeq B$ means $v^\top(A - B)v \ge 0$ for every vector v. Let B_S have orthonormal columns spanning $\{v : v_I = 0,\ \mathbf{1}^\top v = 0\}$. On the face $a_I = \delta\mathbf{1}$, the coordinates in S are interior at a_0. Reference stationarity and quadratic growth give

$$B_S^\top \nabla F_{P_0}(a_0, h_0) = 0, \qquad B_S^\top \nabla^2 F_{P_0}(a_0, h_0) B_S \succeq 2\kappa I.$$

For the second assertion, apply quadratic growth to $a_0 + tv$ and $a_0 - tv$ and let $t \to 0$. Continuity and the C^2 bound imply that the restricted Hessian of F_P is bounded below by κI on a fixed small face neighborhood. The minimum on a smaller closed face ball is interior for sufficiently small ϵ, because its gradient at a_0 is $O(\epsilon)$ and its outward radial derivative is positive. This minimum, denoted by a_P, is unique on the face ball. Integration of the restricted gradient yields

$$\kappa\|a_P - a_0\|_2^2 \le \left|\{B_S^\top \nabla F_P(a_0, h_0)\}^\top B_S^\top (a_P - a_0)\right| \le C\epsilon\|a_P - a_0\|_2.$$

Thus $\|a_P - a_0\|_2 \le C\epsilon$. If $|S| = 1$, the face is a single point and this conclusion is immediate.

To verify that no nearby point outside the face is a minimizer, put

$$d_j(a, P) = \partial_j F_P(a, h_0) - |S|^{-1} \sum_{l\in S} \partial_l F_P(a, h_0), \qquad j \in I.$$

Strict complementarity, continuity, and the C^1 bound give $d_j(a, P) \ge c > 0$ on a fixed neighborhood of a_0. For a feasible a in a sufficiently small neighborhood, let

$$v = \sum_{j\in I}(a_j - \delta), \qquad \bar{a}_I = \delta\mathbf{1}, \qquad \bar{a}_S = a_S + (v/|S|)\mathbf{1}.$$

Choose the neighborhood to ensure that $\bar{a}$ and the segment from $\bar{a}$ to a stay in the preceding neighborhoods. Both points are feasible. Along this segment,

$$\nabla F_P\{\bar{a} + t(a - \bar{a}), h_0\}^\top (a - \bar{a}) = \sum_{j\in I}(a_j - \delta) d_j\{\bar{a} + t(a - \bar{a}), P\} \ge cv.$$

Restricted strong convexity then gives

$$F_P(a, h_0) - F_P(a_P, h_0) \ge cv + \frac{\kappa}{2}\|\bar{a} - a_P\|_2^2.$$

Because $\|a - \bar{a}\|_2^2 \le 2v^2$ and v is uniformly bounded in this neighborhood, the right side is at least $c'\|a - a_P\|_2^2$ for a fixed $c' > 0$. If I is empty, restricted strong convexity gives this inequality directly. Every global minimizer was already localized to this neighborhood. It follows that a_P is the unique global minimizer, its support is S, and local quadratic growth holds uniformly.

Finally, the other-bandwidth gap decreases by at most $2C\epsilon$. It remains positive uniformly after decreasing ϵ_0. The positive coordinates of a_P exceed $\delta + \beta_0/2$ for this choice. This completes the proof. □

For the spectral comparison, keep the marginal distribution of the selected measurements fixed. This changes their dependence with the other measurements without changing the reference manifold or its sampling distribution. The next corollary permits a sequence of such observation distributions. All observations within experiment n are independent draws from P_n, and the metric observations may also form the graph. The estimator, its optimization tolerance, and $\omega_{R,J,b}$ are as defined before Corollary 1.

Corollary 2. *Suppose Theorem 3 holds, Condition 3 holds under P_0, and $\gamma_p > 0$. For a sufficiently small fixed ϵ_0, suppose*

$$d_{\mathrm{TV}}(P_n, P_0) = \epsilon_n \le \epsilon_0, \qquad (P_n)_{X_{S_0}} = (P_0)_{X_{S_0}}, \qquad \omega_{R,J,b} = o(t_N), \qquad t_N = o(\beta_0).$$

If k_n satisfies Theorem 19, then

$$\mathbb{P}(\widehat{h} = h_0,\ \widehat{S} = S_0) \to 1, \qquad \inf_{O^\top O = I_p} \|\widehat{Y} - n^{-1/2}\Phi_0 O\|_F = O_{\mathbb{P}}\left\{ \frac{(k_n/n)^{1/d} + \omega_{R,J,b} + \epsilon_n}{\gamma_p} \right\}.$$

Corollary 2 separates estimation error from the change in the population distance. A fixed small dependence perturbation permits consistent support selection. The bound includes an $O(\epsilon_n)$ term for comparison with the reference eigenspace. If $\epsilon_n \to 0$, this term also tends to zero. For each fixed perturbed distribution, the coordinates converge to its own population eigenspace by Corollary 1. These conclusions concern the predictive-distance estimator; they do not replace the posterior-moment assumptions.

Proof of Corollary 2. Let (a_n, h_0) be the population minimizer under P_n. Theorem 3 gives the same support S_0, a beta-min value at least $\beta_0/2$, and uniform curvature and bandwidth gaps. The sampling and block-approximation bounds in Theorem 2, Corollary 3, and Proposition 31 depend only on the common measurement and prediction-scale bounds. In particular, with probability at least $1 - \alpha$, their uniform criterion error is bounded by a constant times

$$\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}} + \sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{RJ}} + \sqrt{\frac{m\log(mb)}{b}}.$$

For an exact mask average the second term is omitted. This is a derived probability event, not an additional assumption. Uniform population gaps first localize the approximate minimizer to h_0 and the quadratic-growth neighborhood. The minimizing inequality and the derived event give

$$\mathbb{P}(\widehat{h} = h_0) \to 1, \qquad \|\widehat{a} - a_n\|_2 = O_{\mathbb{P}}(\omega_{R,J,b}).$$

The threshold conditions imply $\mathbb{P}(\widehat{S} = S_0) \to 1$. Normalization on this face is uniformly Lipschitz, because its denominator is bounded away from zero. Write a_n^{rep} for the normalized population metric. Then

$$\|\widehat{a}^{\mathrm{rep}} - a_n^{\mathrm{rep}}\|_2 = O_{\mathbb{P}}(\omega_{R,J,b}), \qquad \|a_n^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2 \le C\epsilon_n.$$

The selected marginal is the reference marginal. Thus the representation observations have the same selected-manifold distribution for every n. Proposition 37 preserves the external gap on a sufficiently small fixed neighborhood of a_0^{rep}. Lemma 13 applies simultaneously to every metric in that neighborhood, including the fitted metric. Lemma 12 and Proposition 37 compare its population eigenfunction evaluations with Φ_0. The triangle inequality gives

$$\inf_{O^\top O = I_p} \|\widehat{Y} - n^{-1/2}\Phi_0 O\|_F \le O_{\mathbb{P}}\{(k_n/n)^{1/d}/\gamma_p\} + O_{\mathbb{P}}\{(\omega_{R,J,b} + \epsilon_n)/\gamma_p\}.$$

No conditioning on the fitted metric is used. The argument applies to triangular sequences P_n. This completes the proof. □

The following dependent model preserves the conditional response mean and both measurement marginals. Let P_0 satisfy (1) and Theorem 3, with S_0 indexing S. Choose nonconstant bounded functions b_S, b_U with $|b_S|, |b_U| \le 1$ and $\mathbb{E}_0 b_S(S) = \mathbb{E}_0 b_U(U) = 0$, where $\mathbb{E}_0$ denotes expectation under P_0. For $|\epsilon| < 1$, define

$$\frac{dP_\epsilon}{dP_0}(S, U, V) = 1 + \epsilon b_S(S) b_U(U).$$

Independence and centering make this a positive density with integral one. Integrating over U or S preserves the other marginal. Because the density ratio depends only on (S, U), it cancels upon conditioning, giving $\mathbb{E}_\epsilon(V \mid S, U) = g(S)$. Direct integration yields

$$d_{\mathrm{TV}}(P_\epsilon, P_0) = \frac{|\epsilon|}{2}\mathbb{E}_0|b_S(S)|\,\mathbb{E}_0|b_U(U)|, \qquad \mathrm{cov}_\epsilon\{b_S(S), b_U(U)\} = \epsilon\mathbb{E}_0 b_S(S)^2\mathbb{E}_0 b_U(U)^2.$$

For sufficiently small nonzero ϵ, the measurements are dependent but Theorem 3 preserves their predictive support. If Condition 3 holds under P_0 and $\gamma_p > 0$, the unchanged selected marginal permits Corollary 2 under its stated tuning conditions. Under coordinate masking, the same density perturbation applies to $P_0 = P_S \otimes P_U$ without a response variable.

The sphere model in Theorem 8 gives a concrete example. Fix its number of blocks and use the singleton bandwidth set $\mathcal{H} = \{\sqrt{m}\tau\}$ with the sufficiently large τ in that theorem. Its proof gives strict exchange margins and curvature on the signal face. Set $b_S = X_1^{(1)}/\sqrt{s}$ and $b_U = X_1^{(2)}/\sqrt{s}$. Each factor has mean zero, absolute value at most one, and second moment $1/s$. Their covariance under P_ϵ equals ϵ/s^2, and the first block remains the selected set for sufficiently small $|\epsilon|$. The spectral comparison is with the independent reference distribution; for fixed nonzero ϵ, the new population weights need not equal the reference weights.

A.2 Sharper risk and metric bounds

Because the feature weights satisfy a trace constraint, the sampling error depends only logarithmically on the number of measured features. To establish this bound, we allow the lower boundary to be $0 \le \delta < 1$, and we consider independent blocks that may follow different distributions. Furthermore, we allow the rows within each block to be dependent. In the following theorem and Corollary 3, $\mathcal{R}_{b,\rho}$ denotes the average of the expected block criteria, which includes the expectation over the prescribed masks. This definition matches the previous one when the blocks are identically distributed. We continue to use the same block sizes, mask design, and measurement bounds as in Condition 1.

Theorem 5. *Under the block design specified above, for every $0 < \alpha < 1$, with probability at least $1 - \alpha$,*

$$\sup_{a \in \mathcal{A}_\delta,\, h \in \mathcal{H}} |\widehat{\mathcal{R}}_\rho(a, h) - \mathcal{R}_{b,\rho}(a, h)| \le C\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}}.$$

The constant C depends only on B, π_-, and τ_-.

Theorem 5 provides the sampling bound used in Theorem 2, but it does so under this weaker block design. The constant C is strictly uniform in both the block size and the number of response coordinates. This result specifically targets the finite-block risk. In contrast, the population approximation in Proposition 31 still depends on the total feature dimension.

Proof of Theorem 5. Put $k_- = \exp\{-2B^2/(\pi_-\tau_-^2)\}$ and let C_u be the upper scaling constant in Lemma 4. Fix one block and one mask. Let $Z \in \mathbb{R}^{b\times q}$ be its response matrix, with $q \ge 1$, and put $E = Z - WZ$. Write $\eta_{ij} = \log K_{ij}$ off the diagonal. Let H be a symmetric perturbation of η with zero diagonal, and let z be the resulting derivative of the log scaling vector. Write $\circ$ for entrywise multiplication. The differentiated row equations give

$$\dot{W} = W \circ \{H + z\mathbf{1}^\top + \mathbf{1}z^\top\}, \qquad (I_b + W)z = -(W \circ H)\mathbf{1}.$$

On symmetric matrices with zero diagonal, use the inner product $\langle A, B\rangle_W = \sum_{i\ne j} W_{ij}A_{ij}B_{ij}$ and its norm $\|A\|_W = \langle A, A\rangle_W^{1/2}$. Let $\mathsf{P}_W H = H + z\mathbf{1}^\top + \mathbf{1}z^\top$ off the diagonal and set its diagonal to zero. Then $(W \circ \mathsf{P}_W H)\mathbf{1} = 0$. For every symmetric A satisfying $(W \circ A)\mathbf{1} = 0$,

$$\langle A, H - \mathsf{P}_W H\rangle_W = -2\sum_i z_i \sum_{j\ne i} W_{ij}A_{ij} = 0.$$

Thus P_W is an orthogonal projection for this inner product. Differentiating the loss gives

$$D\ell_b(\eta)[H] = \langle G, \mathsf{P}_W H\rangle_W, \qquad G = -\frac{EZ^\top + ZE^\top}{bq}.$$

Only the off-diagonal part of G enters this identity. Because $|Z_{ij}| \le B$, $|E_{ij}| \le 2B$, and $\sum_{ij} W_{ij} = b$, we have $\|G\|_W \le 4B^2/\sqrt{b}$. Lemma 4 gives $W_{ij} \le C_u^2/b$ for symmetric kernels with off-diagonal entries in $[k_-, 1]$. Projection contraction yields

$$|D\ell_b(\eta)[H]| \le \|G\|_W \|H\|_W \le \frac{4B^2 C_u}{b} \|H\|_F.$$

Integrating along segments in $[\log k_-, 0]^{b(b-1)}$ gives the C/b Frobenius Lipschitz bound.

For fixed h, write $p_j = a_j/m$. The vector p lies in the probability simplex, including when $\delta = 0$. Write $\ell_r(a, h) = \ell_{I_r}(a, h; \xi_r)$, and let X_{ruj} denote feature j of observation u in block r. For block r, define

$$v_{r,uv,j} = -\frac{m\xi_{rj}(X_{ruj} - X_{rvj})^2}{2\pi h^2}, \qquad \eta_{r,uv}(a, h) = \sum_{j=1}^{m} p_j v_{r,uv,j}.$$

Set $v_{r,uu,j} = 0$. Each entry satisfies $|v_{r,uv,j}| \le M_0 = 2B^2/(\pi_- \tau_-^2)$. Let ε_r and ε_{ruv} be independent Rademacher signs. Conditional on the blocks and masks, the vector contraction inequality of Maurer (2016, Theorem 3) gives

$$\mathbb{E}_\varepsilon \sup_{a \in \mathcal{A}_\delta} \frac{1}{R} \sum_{r=1}^{R} \varepsilon_r \ell_r(a, h) \le \frac{C}{bR} \mathbb{E}_\varepsilon \max_{1 \le j \le m} \sum_{r=1}^{R} \sum_{u,v=1}^{b} \varepsilon_{ruv} v_{r,uv,j} \le C\sqrt{\frac{\log(2m)}{R}}.$$

Each sum is sub-Gaussian with variance proxy $Rb^2 M_0^2$, which gives the second inequality. Apply contraction on a countable dense subset and extend by continuity. The argument also applies to $-\ell_r$.

One-sided symmetrization bounds the expectation of each of $\sup_a(\widehat{\mathcal{R}}_\rho - \mathcal{R}_{b,\rho})$ and $\sup_a(\mathcal{R}_{b,\rho} - \widehat{\mathcal{R}}_\rho)$ by the same order. The deterministic ridge cancels in both differences. Because $0 \le \ell_r \le 4B^2$, replacing one independent block changes either supremum by at most $4B^2/R$. The bounded-difference inequality and a union bound over both signs and $h \in \mathcal{H}$ prove the theorem. This completes the proof. □

This proof controls the expected finite-block criterion without requiring a strictly positive coordinate floor or row independence within each block. However, the population approximation and the observation-level limits still require independent rows. Furthermore, identifying the population metric and guaranteeing the convergence of the selected representation remain separate conclusions.

To distinguish the numerical integration over masks from the sampling of new blocks, let $\widetilde{\mathcal{R}}_{R,J}$ be the empirical average of the criteria over J independent masks per block, including the ridge term. Let $\overline{\mathcal{R}}_R$ be its exact conditional expectation over the masks. The next result separates these two sources of variation.

Corollary 3. *Under the block design of Theorem 5, with probability at least* $1 - \alpha$*,*

$$\sup_{a,h} |\overline{\mathcal{R}}_R(a, h) - \mathcal{R}_{b,\rho}(a, h)| \le C\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}}.$$

Conditionally on the observation blocks, with probability at least $1 - \alpha$*,*

$$\sup_{a,h} |\widetilde{\mathcal{R}}_{R,J}(a, h) - \overline{\mathcal{R}}_R(a, h)| \le C\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{RJ}}.$$

Both suprema are over $\mathcal{A}_\delta \times \mathcal{H}$*;* $\delta = 0$ *is allowed.*

Corollary 3 shows that neither the sampling error nor the mask integration error depends linearly on the feature dimension. With fixed observation blocks, increasing the number of masks J improves only the mask integration bound.

Proof of Corollary 3. Jensen's inequality moves the mask expectation outside the supremum in the symmetrized process for $\overline{\mathcal{R}}_R$. The proof of Theorem 5 then applies to R independent blocks. Conditionally on these blocks, the RJ masks are independent, although their response matrices may differ. The same vector contraction argument applies to these RJ terms. Each loss is bounded by $4B^2$, and one mask changes the normalized supremum by at most $4B^2/(RJ)$. Bounded differences and the union bound over bandwidths and signs prove the second assertion. This completes the proof. □

Unless stated otherwise, the remaining results assume the original design where observations are independent and identically distributed. To bound the excess risk of an approximate empirical minimizer, we compare its criterion value to the value at (a_0, h_0) and add the uniform deviations of both points. Combining the preceding sampling bounds with Proposition 31 establishes this population excess-risk bound, which includes the logarithmic sampling term. When the excess-risk bound tends to zero in probability, Condition 2 and compactness place the estimator in a region with local quadratic growth with probability tending to one. There, the square root of the excess-risk bound controls the metric error. This logic yields the first two terms of Corollary 13, leaving its finite-block approximation term and fixed-m identification premise unchanged.

When the feature dimension is fixed, differentiating the objective yields a sharper metric rate than using the risk bound alone. The following theorem applies to the criterion in (6), evaluated using either J independent masks per block or the exact mask average, under the objective tolerance given in (7). We maintain the response and feature bounds from Condition 1.

Theorem 6. *Suppose Conditions 1 and 2 hold. If m, $\mathcal{H}$, and the mask support are fixed, $R, b \to \infty$, and $\eta_R = o_{\mathbb{P}}(1)$, then*

$$\mathbb{P}(\widehat{h} = h_0) \to 1, \qquad \|\widehat{a} - a_0\|_2 = O_{\mathbb{P}}\Big\{(Rb)^{-1/2} + b^{-1} + (RJ)^{-1/2} + \eta_R^{1/2}\Big\}.$$

For an exact mask average, omit $(RJ)^{-1/2}$.

Theorem 6 explicitly separates the errors arising from observation variation, the bias of finite blocks, mask integration, and optimization tolerance. This guarantee permits the optimal weights to lie on the boundary, and it does not require an empirical curvature condition. With the exact mask average, $R/b = O(1)$ and $\eta_R = O_{\mathbb{P}}\{(Rb)^{-1}\}$ give the total-observation rate $O_{\mathbb{P}}\{(Rb)^{-1/2}\}$. Corollary 16 subsequently identifies the additional block growth required to perform inference centered exactly at the population metric. The constants in this result assume a fixed dimension m and a stable prediction bandwidth; we state the increasing-dimensional results separately.

Proof of Theorem 6. Write $N = Rb$, let P be the distribution of a complete observation, and let P_b and P_N be the empirical distributions of one block and of all observations. For fixed h_0, put $S_{\xi,a}(Q) = \nabla_a L_{\xi,a,h_0}(Q)$, where $L_{\xi,a,h_0}(Q)$ is the unpenalized population prediction risk under Q, defined in the proof of Lemma 8. Write D_Q for differentiation with respect to the distribution argument. On a cube $D \subset \mathbb{R}^{d_0}$ containing the complete observations, choose an integer $s > d_0/2$. Let $H^s(D)$ be the Sobolev space of functions with square-integrable weak derivatives through order s, and define $\|\nu\|_* = \sup_{\|f\|_{H^s(D)} \le 1} |\int f \, d\nu|$. On a fixed neighborhood of a_0 in $\mathcal{A}_\delta$, the lemma gives joint smoothness in the measure, metric, and diagonal-deletion argument t. Boundedness of centered point masses in its Sobolev dual Hilbert space and independence give

$$\mathbb{E}\|P_b - P\|_*^2 \le C/b, \qquad \mathbb{E}\|P_b - P\|_*^4 \le C/b^2.$$

In the fourth-moment expansion, only terms with at most two distinct indices have nonzero expectation. Replacing b by N gives the bounds for P_N. Taylor expansion at $(P, 0, a)$, evaluated at $(P_b, 1/b, a)$, gives

$$\nabla_a \ell_b(a, h_0; \xi) - S_{\xi,a}(P) = D_Q S_{\xi,a}(P)[P_b - P] + E_{\xi,b}(a), \qquad \mathbb{E}\sup_a \|E_{\xi,b}(a)\|_2 \le C/b.$$

The supremum and derivative bounds are uniform over this neighborhood and the finite mask support. On the dual-norm neighborhood, the remainder is bounded by $C\{\|P_b - P\|_*^2 + b^{-1}\}$. Bounded differences and Lemma 4 control the exponentially small complement.

Averaging the linear terms over masks and R blocks applies the derivative to $P_N - P$; the ridge derivative cancels. For the criterion $\widehat{\mathcal{R}}_\rho^{\mathrm{ex}}$ with exact mask averages, the operator bound and Markov's inequality give

$$\sup_a \|\nabla_a \widehat{\mathcal{R}}_\rho^{\mathrm{ex}}(a, h_0) - \nabla_a \mathcal{R}_{\infty,\rho}(a, h_0)\|_2 = O_{\mathbb{P}}\{N^{-1/2} + b^{-1}\}.$$

The remainders may depend on the leading average.

Conditional on the observations, the RJ centered mask scores are independent, bounded, and Lipschitz in a by Lemma 4. On 2^{-l}-nets of cardinality at most $C2^{ml}$, the bounded-variable maximal

inequality for successive increments, summed over $l \geq 0$, bounds the conditional expected supremum by $\frac{C}{\sqrt{RJ}} \sum_{l=0}^{\infty} 2^{-l}\sqrt{l+1} \leq \frac{C}{\sqrt{RJ}}$. Because the vector dimension is fixed, the same bound holds for the score norm. Thus the local score error is of order $q_R = N^{-1/2} + b^{-1} + (RJ)^{-1/2}$ in probability, with the last term omitted for exact averages.

Bounded block losses, their uniform Lipschitz constants, and Proposition 31 give uniform risk consistency, also for block-averaged masks. Uniqueness, compactness, the bandwidth gap, and $\eta_R = o_{\mathbb{P}}(1)$ give $\widehat{a} \to a_0$ and $\mathbb{P}(\widehat{h} = h_0) \to 1$. Let $d = \widehat{a} - a_0$. With probability tending to one, the segment from a_0 to $\widehat{a}$ lies in that neighborhood. The minimizing inequality, integration of the score, and Condition 2 give

$$\kappa\|d\|_2^2 \leq \mathcal{R}_{\infty,\rho}(\widehat{a}, h_0) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \leq O_{\mathbb{P}}(q_R)\|d\|_2 + \eta_R.$$

Solving this quadratic inequality proves the rate, including at the boundary. This completes the proof. □

A.3 Posterior-moment estimator

When auxiliary measurements are available only during training, the posterior-moment estimator uses them to estimate class moments for classification. Let $\mathcal{L}$ denote the indices of labeled triples (X_i, V_i, C_i), and let $\mathcal{U}$ denote the indices of paired unlabeled observations (X_i, V_i). We fit a class-score teacher on $\mathcal{L}$ using both sets of measurements, and we apply this teacher to $\mathcal{U}$. To filter out uncertain predictions, we select the observations whose confidence scores fall at or above an empirical quantile corresponding to a prespecified fraction, including all ties at this threshold. Let $\mathcal{U}_r$ denote the indices of these selected observations, and let $\widehat{p}_i$ be their corresponding score vectors. For each selected observation $i \in \mathcal{U}_r$, we define the responsibility $\widehat{z}_i$ as either the soft probability vector $\widehat{p}_i$ or the hard one-hot vector of its largest score. We fix this choice of soft or hard responsibilities before evaluation. For a fixed label weight $\omega > 0$, we define the raw responsibility of observation i for class k as

$$\widehat{r}_{ik}^{\,\mathrm{raw}} = \begin{cases} \omega \mathbf{1}(C_i = k), & i \in \mathcal{L}, \\ \widehat{z}_{ik}, & i \in \mathcal{U}_r. \end{cases}$$

We normalize the raw responsibilities by dividing them by $c_\omega = \max(1, \omega)$, setting $\widehat{r}_{ik} = \widehat{r}_{ik}^{\,\mathrm{raw}}/c_\omega$. This ensures that every normalized responsibility lies in $[0, 1]$. Because this common scaling affects all terms equally, it completely cancels out from the empirical class proportions, means, variance ratios, and pooled covariance, leaving the final estimator unchanged. To estimate the moments, we use only the deployable measurements X_i. We define the observation weights $\widehat{r}_{i+} = \sum_k \widehat{r}_{ik}$ and the total weight $\widehat{n}_+ = \sum_{i,k} \widehat{r}_{ik}$, and we estimate the moments as

$$\begin{aligned} &\widehat{n}_k = \sum_i \widehat{r}_{ik}, && \widehat{\pi}_k = \widehat{n}_k/\widehat{n}_+, && \widehat{\mu}_{kj} = \widehat{n}_k^{-1} \sum_i \widehat{r}_{ik} X_{ij}, \\ &\widehat{\mu}_j = \widehat{n}_+^{-1} \sum_i \widehat{r}_{i+} X_{ij}, && \widehat{B}_j = \sum_k \widehat{\pi}_k (\widehat{\mu}_{kj} - \widehat{\mu}_j)^2, && \widehat{W}_j = \widehat{n}_+^{-1} \sum_{i,k} \widehat{r}_{ik} (X_{ij} - \widehat{\mu}_{kj})^2, \\ & && \widehat{\theta}_j = \widehat{B}_j / \widehat{W}_j. \end{aligned} \tag{13}$$

To select the most informative features, the posterior-moment estimator identifies the s coordinates with the largest Fisher ratios $\widehat{\theta}_j$ and groups them into the active set $\widehat{S}_q$.

When the responsibilities are fixed, applying an affine transformation $(X_{ij} - c_j)/s_j$ with $s_j > 0$ scales both $\widehat{B}_j$ and $\widehat{W}_j$ by $1/s_j^2$. Their ratio $\widehat{\theta}_j$ remains identical, even if the centering and scaling constants were estimated from the labeled observations. While this invariance holds for moment-based coordinate selection, it does not establish invariance of the initial teacher fitting, the subsequent covariance regularization, or nonlinear clipping operations.

Using the same responsibilities, we estimate a pooled within-class covariance matrix $\widehat{\Sigma}_{\widehat{S}_q}$ on the selected coordinates. For a prespecified covariance ridge $\lambda_N > 0$, a score-affinity bandwidth $h_g > 0$, and a uniform prediction prior, we define the class scores $\widehat{g}_k(x)$, their vector

$\widehat{g}(x) = (\widehat{g}_1(x), \ldots, \widehat{g}_K(x))^\top$, and the score-based affinity $\widehat{A}_g(x, x')$ as

$$\widehat{g}_k(x) = x_{\widehat{S}_q}^\top (\widehat{\Sigma}_{\widehat{S}_q} + \lambda_N I)^{-1} \widehat{\mu}_{k,\widehat{S}_q} - \frac{1}{2} \widehat{\mu}_{k,\widehat{S}_q}^\top (\widehat{\Sigma}_{\widehat{S}_q} + \lambda_N I)^{-1} \widehat{\mu}_{k,\widehat{S}_q},$$
$$\widehat{A}_g(x, x') = \exp\{-\|\widehat{g}(x) - \widehat{g}(x')\|_2^2/(2h_g^2)\}. \tag{14}$$

To further regularize the covariance, we fix a diagonal shrinkage parameter $0 \le \tau_{\mathrm{cov}} \le 1$ and replace $\widehat{\Sigma}_{\widehat{S}_q}$ in (14) with $(1 - \tau_{\mathrm{cov}})\widehat{\Sigma}_{\widehat{S}_q} + \tau_{\mathrm{cov}} \operatorname{diag}(\widehat{\Sigma}_{\widehat{S}_q})$, where diag sets the off-diagonal entries to zero. This shrinkage alters the common covariance used for prediction without affecting the univariate coordinate scores in (13). We specify both covariance regularization parameters in the numerical studies.

The native classifier assigns an observation to the class that maximizes the score $\widehat{g}_k(x)$. To build the spectral representation, we take the affinity in (14), remove its diagonal, and compute d_g nonconstant eigenvectors from its symmetric degree normalization. Let $U_i \in \mathbb{R}^{d_g}$ denote the resulting spectral coordinate for observation i. Because this graph includes labeled, paired unlabeled, and evaluation covariates, it is transductive. For a fixed label weight $\omega_g > 0$, we define the regression targets and weights: for a labeled observation $i \in \mathcal{L}$, we set $t_i = e_{C_i}$ and $v_i = \omega_g$; for an unlabeled observation $i \in \mathcal{U}$, we set t_i to the prespecified hard or soft teacher target and $v_i = 1$. Given a ridge penalty $\alpha > 0$, the spectral prediction solves

$$(\widehat{b}, \widehat{B}) = \underset{b \in \mathbb{R}^K,\, B \in \mathbb{R}^{d_g \times K}}{\arg\min} \sum_{i \in \mathcal{L} \cup \mathcal{U}} v_i \|t_i - b - U_i B\|_2^2 + \alpha \|B\|_F^2 \tag{15}$$

and assigns an evaluation observation to the class that maximizes the corresponding coordinate of $\widehat{b} + U_i \widehat{B}$. Importantly, for a fitted model, both the native prediction and every class score used to construct the graph depend exclusively on the deployable measurements X. For ablation comparisons, we include a direct pseudo-label linear discriminant analysis that uses all deployable coordinates without selection, as well as the earlier posterior-response metric that replaces the held-out matrix in (5) with teacher scores. We evaluate both ablations because neither estimates the sparse posterior moments in (13).

The label-free criterion can retain scientifically irrelevant features that predict the target. The posterior-moment method instead describes the population after confidence screening, which can differ from the original population. Proposition 5 bounds empirical screening error for soft responsibilities.

A.4 Posterior-moment theory

For the separate moment estimator, Proposition 2 equates ideal posterior-weighted and label-weighted moments after screening. With independent observations, fixed class count, sub-Gaussian coordinates, and positive screened class masses and within-class variances, Theorem 7 bounds every Fisher-score error with probability $1 - \alpha$ by a constant times $\{\log(Km^2/\alpha)/N\}^{1/2} + \log(Km^2/\alpha)/N + \epsilon_N$, when sufficiently small. Here $0 < \alpha < 1$, N counts labeled and paired unlabeled observations, and ϵ_N measures covariate-weighted responsibility error. A population gap exceeding twice this bound ensures exact selection. Corollary 4 bounds classifier disagreement under covariance and margin conditions. Proposition 3 and Corollary 5 bound graph eigenspace error under score bounds and a positive eigengap. Proposition 4 transfers this error to predictions fitted to the same targets with ridge bounded away from zero.

The ideal posterior response correctly identifies the class moments of the deployable measurements in the population. We define a fixed population retention weight $q_0(X, V) \in [0, 1]$ with a strictly positive expectation $c_q = \mathbb{E} q_0(X, V) > 0$. We then define the screened distribution P_q, which has a density q_0/c_q relative to the original distribution of (X, V, C), and we write $\mathbb{E}_q$ and $\mathbb{P}_q$ for its corresponding expectation and probability. To ensure the moments are well-defined, we assume $\mathbb{E}\{q_0(X, V)\|X\|_2^2\} < \infty$, and we require the class probabilities and within-class variances to remain strictly positive after confidence screening. The screened population moments are

$$\pi_{qk} = \mathbb{P}_q(C = k), \qquad \mu_{qkj} = \mathbb{E}_q(X_j \mid C = k), \qquad \mu_{qj} = \mathbb{E}_q X_j,$$

and

$$B_{qj} = \sum_{k=1}^K \pi_{qk}(\mu_{qkj} - \mu_{qj})^2, \qquad W_{qj} = \sum_{k=1}^K \pi_{qk} \mathbb{E}_q\{(X_j - \mu_{qkj})^2 \mid C = k\}, \qquad \theta_{qj} = B_{qj}/W_{qj}.$$

For a prespecified feature count s, we define S_q as the set containing the s coordinates with the largest Fisher ratios θ_{qj}. A later separation condition will guarantee that this set is unique. Using the retention weight q_0 allows us to model the confidence screening step directly at the population level. For labeled observations, the constant contribution to q_0 is simply ω/c_ω.

The following proposition establishes the core identity that the posterior-moment estimator uses. It proves that the ideal posterior $\eta_X(X)$ is sufficient for classification, and it confirms that any coordinate independent of $\{C, q_0(X,V)\}$ has a population score of exactly zero.

Proposition 2. *For $r \in \{0,1,2\}$, $1 \le k \le K$, and $1 \le j \le m$,*

$$\mathbb{E}\{q_0(X,V)\eta_{XV,k}(X,V)X_j^r\} = \mathbb{E}\{q_0(X,V)\mathbf{1}(C=k)X_j^r\}.$$

Also,

$$\mathbb{P}\{C = k \mid \eta_X(X)\} = \eta_{X,k}(X) \quad \textit{almost surely}.$$

If X_j is independent of $\{C, q_0(X,V)\}$ and $W_{qj} > 0$, then $\theta_{qj} = 0$.

Proposition 2 shows that training-only measurements can identify the class moments of X through the ideal posterior without ever becoming prediction coordinates themselves. The Fisher ratio θ_{qj} equals zero when the screened class means of X_j coincide. Thus it does not detect a coordinate whose class information lies only in its class-conditional variances or higher-order moments.

To analyze the finite-sample behavior for a fixed class count K, we assume that (X_i, V_i, C_i) for $1 \le i \le N$ are independent observations drawn from prespecified distributions P_i. We define the population quantities under the average distribution $P = N^{-1}\sum_i P_i$, and we include any sampling strata directly in V_i. This setup strictly covers independent sampling with fixed class counts, where identically distributed observations appear as a special case. The screened distribution P_q is derived from P. We define the oracle responsibility vector R_i^0 as either $q_0(X_i,V_i)e_{C_i}$ or $q_0(X_i,V_i)\eta_{XV}(X_i,V_i)$, assigning zero mass to unretained observations. The choice may depend on the sampling type recorded in V_i; we use the hard-label oracle for labeled observations. Proposition 2 guarantees that both choices target exactly the same moments. The first choice serves as the theoretical target for hard generated responsibilities, and the second serves as the target for soft posterior probabilities. We let $\widehat{R}_i \in [0,1]^K$ denote the normalized generated responsibility actually used by the estimator, which may depend on the fitted teacher. In our empirical studies, fixed-data reallocations act strictly as descriptive comparisons; they are not independent draws from a new population for each allocation. By including the sampling type in V_i, we can rigorously analyze a prespecified mixture of labeled and paired unlabeled observations using a relative label weight ω. The normalization constant c_ω leaves every fitted moment mathematically unchanged. We measure the discrepancy between the estimated and oracle responsibilities and define the overall error scale by

$$\epsilon_N = \frac{1}{N}\sum_{i=1}^N \|\widehat{R}_i - R_i^0\|_1\{1+\|X_i\|_\infty^2\}, \qquad d_N(\alpha) = \left\{\frac{\log(6Km^2/\alpha)}{N}\right\}^{1/2} + \frac{\log(6Km^2/\alpha)}{N} + \epsilon_N.$$

From the observations $(X_i, \widehat{R}_i)$, we compute the empirical Fisher ratios $\widehat{\theta}_j$, the class means, and the pooled within-class covariance using (13). Their population counterparts are defined under the screened distribution P_q. For any subset of coordinates S, we denote the population pooled covariance on X_S by $\Sigma_{q,S}$ and its empirical version by $\widehat{\Sigma}_S$.

The following theorem establishes uniform guarantees for moment estimation and exact coordinate selection. We use the Orlicz norm $\|Z\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}\exp(Z^2/c^2) \le 2\}$ to bound the uncentered variable, which explicitly includes its mean. The theorem's premises constrain the population distributions, while the realized errors from the teacher and the retention rule enter the high-probability implication strictly through ϵ_N.

Theorem 7. *Suppose $\sup_{i,j}\|X_{ij}\|_{\psi_2} \le C_X$, and*

$$\min_k \frac{1}{N}\sum_{i=1}^N \mathbb{E}R_{ik}^0 \ge c_0, \qquad \min_j W_{qj} \ge w_0$$

for constants $c_0, w_0 > 0$. There are constants $C, c > 0$ such that, with probability at least $1-\alpha$, the implication $d_N(\alpha) \le c$ gives

$$\max_{1\le j\le m} |\widehat{\theta}_j - \theta_{qj}| \le C d_N(\alpha),$$

and, for every fixed S of size s,

$$\max_k \|\widehat{\mu}_{k,S} - \mu_{qk,S}\|_2 + \|\widehat{\Sigma}_S - \Sigma_{q,S}\|_{\mathrm{op}} \le Cs\, d_N(\alpha).$$

If $\min_{j\in S_q} \theta_{qj} - \max_{j\notin S_q} \theta_{qj} > 2Cd_N(\alpha)$, then the s largest empirical scores select S_q.

Theorem 7 cleanly separates the sampling error from the responsibility estimation error. Positive class masses and within-class variances keep the population moments and Fisher ratios well-defined. Related semi-supervised discriminant methods use moment constraints or estimated class labels (Loog, 2014; Wang et al., 2016). Exact labeled responsibilities and exact posterior probabilities on paired unlabeled observations under the fixed retention rule give $\epsilon_N = 0$ for this mixed oracle. If we use hard responsibilities under the same retention rule and the hard-label oracle, ϵ_N equals twice the confidence-weighted pseudo-label error, including the covariate factor in its definition. When the screened classes inherently overlap, this hard-assignment error need not converge to zero. Exact selection succeeds on the stated event when $d_N(\alpha) \le c$ and the population Fisher gap strictly exceeds $2Cd_N(\alpha)$, whereas using all coordinates would include any nuisance variation present.

We can also incorporate fitted preprocessing steps without assuming that the transformed observations remain independent. Let $Z \in \mathbb{R}^{d_Z}$ denote the raw deployable vector. We define the oracle responsibilities on the raw complete observations using either $q_0 e_C$ or $q_0 \eta_{ZV}$, where $\eta_{ZV} = \mathbb{E}(e_C \mid Z, V)$. Because the population moments apply the identical transformation to the observations, we can track the error bounds through any affine map T. If T has a linear part $A \in \mathbb{R}^{m\times d_Z}$, we define its scale as $a_T = \max_j \|A_{j\cdot}\|_1$. Let $d_N^Z(\alpha)$ be the error scale evaluated using the raw dimension and raw coordinates. Assuming the same class-mass bound and a uniform sub-Gaussian bound $\sup_{i,j} \|Z_{ij}\|_{\psi_2} \le C_Z$, the maximum-norm errors for the class means and covariances are bounded by $Ca_T d_N^Z(\alpha)$ and $Ca_T^2 d_N^Z(\alpha)$, respectively. This holds with probability at least $1-\alpha$, simultaneously over all such maps T, provided $d_N^Z(\alpha) \le c$.

To enforce boundedness, we choose a fixed $B_c > 0$ and clip each coordinate of the affine map to $[-B_c, B_c]$. The conclusions of Theorem 7 hold simultaneously for all such clipped maps T that satisfy the population conditions, provided we replace $d_N(\alpha)$ with

$$\left\{ \frac{(d_Z+1)\log(2N) + \log(6Km^2/\alpha)}{N} \right\}^{1/2} + \epsilon_N,$$

where ϵ_N evaluates the error using the transformed covariates $X_i = T(Z_i)$ alongside the fixed raw oracle responsibilities. The constants in this bound may depend on B_c and the population bounds. Because this conclusion applies uniformly over T, it allows data-dependent transformations like principal component analysis, standardization, and clipping to be fitted directly on the labeled observations. When evaluating classifiers and affinities, we compare the estimated moments to the population moments generated by the same realized map, which means this result does not assert that the final score graph itself is strictly invariant.

On the selected active set S_q, we define the population common-covariance scores as

$$g_k(x) = x_{S_q}^\top \Sigma_{q,S_q}^{-1} \mu_{qk,S_q} - \frac{1}{2} \mu_{qk,S_q}^\top \Sigma_{q,S_q}^{-1} \mu_{qk,S_q}, \qquad 1 \le k \le K.$$

The empirical scores $\widehat{g}_k$ estimate these quantities using the empirical set $\widehat{S}_q$, the estimated means, and the regularized covariance $\widehat{\Sigma}_{\widehat{S}_q} + \lambda_N I_s$. We let $g_{(1)}(x) \ge g_{(2)}(x)$ denote the two largest population scores for a given x.

The next corollary connects these moment estimates to disagreement with the population classifier. To bound this disagreement, we use a margin condition on the population discriminant scores, analogous to the binary posterior margin condition of Audibert & Tsybakov (2007). Its constants are $A > 0$ and $\beta \ge 0$, and the bound explicitly tracks the impact of the covariance ridge. The constant C_M in the bound may depend on the truncation level M, the population constants, and the maximum mean norm $\max_k \|\mu_{qk,S_q}\|_2$. Consequently, this result does not yield a uniform rate if the class-mean norms grow with the feature dimension s.

Corollary 4. *Suppose the assumptions of Theorem 7 hold and $\lambda_{\min}(\Sigma_{q,S_q}) \ge \lambda_0 > 0$. If*

$$\mathbb{P}_q\{g_{(1)}(X) - g_{(2)}(X) \le t\} \le At^\beta, \qquad t > 0,$$

then, with probability at least $1-\alpha$, the conclusions below hold whenever $d_N(\alpha) \le c$ and the selection inequality in Theorem 7 holds. For every $M > 0$,

$$\mathbb{P}_q\{\arg\max_k \widehat{g}_k(X) \neq \arg\max_k g_k(X)\} \le \mathbb{P}_q(\|X_{S_q}\|_2 > M) + A\{C_M s(d_N(\alpha) + \lambda_N)\}^\beta.$$

If the screened class-conditional distributions of X_{S_q} are Gaussian with common covariance and equal class weights, $\arg\max_k g_k$ is the Bayes classifier based on X_{S_q} under P_q.

Corollary 4 provides a prediction bound based directly on the class moments used to construct the affinity graph. The first term in the bound isolates the error from observations outside a fixed bounded region, which safely accommodates Gaussian covariates. To connect the screened distribution to the deployment setting, we define the expected retention weight $\bar{q}_0(x) = \mathbb{E}\{q_0(X,V) \mid X = x\}$. If this weight is strictly bounded below by $\underline{q} > 0$, we can multiply the bound by $c_q/\underline{q}$ to control the disagreement under the original average sampling distribution P. This directly yields a deployment guarantee when P represents the deployment distribution. Specifically, any event $\mathcal{A}$ determined purely by X satisfies the inequality

$$\mathbb{P}_q(\mathcal{A}) = c_q^{-1}\mathbb{E}\{\bar{q}_0(X)\mathbf{1}(\mathcal{A})\} \ge \underline{q}\,\mathbb{P}(\mathcal{A})/c_q.$$

When the labeled sampling is completely independent of X, selecting a fixed positive labeled fraction with a positive weight $\omega > 0$ satisfies this lower bound condition. However, confidence screening may still alter the class-conditional distributions. To extend the screened Bayes conclusion to deployment, it is sufficient that the screened and deployment decision rules agree almost surely. For any x where $\{\bar{q}_0(x) > 0\}$, conditional expectation yields

$$\mathbb{P}_q(C = k \mid X = x) = \eta_{X,k}(x)\frac{\mathbb{E}\{q_0(X,V) \mid X = x, C = k\}}{\bar{q}_0(x)}.$$

Applying this identical logic with X_{S_q} in place of X proves that if retention is independent of the class label conditional on X_{S_q}, the posterior required by the selected-coordinate classifier remains perfectly preserved. Screening based solely on X_{S_q} guarantees this condition, whereas screening based on other coordinates or an auxiliary view need not. Because matching the class means and covariances does not imply that the full posteriors match, a sufficient condition for extending the Bayes conclusion to all of X is that the selected coordinates X_{S_q} preserve this true class posterior.

When we apply a fixed diagonal shrinkage, the disagreement bound in Corollary 4 continues to hold if we replace both the empirical and population covariance matrices with their respective diagonally shrunken versions. Consequently, the margin condition must be evaluated using the population scores generated by this shrunken covariance. Under the Gaussian assumptions in the corollary, the Bayes optimality conclusion holds either when there is no shrinkage ($\tau_{\text{cov}} = 0$) or when the true covariance Σ_{q,S_q} is strictly diagonal. Applying a fixed positive ridge explicitly leaves the λ_N term in the bound. Because of this, the bound does not guarantee consistency for the unregularized population classifier.

Write $g(x) = (g_1(x), \ldots, g_K(x))^\top$. For a fixed bandwidth $h_g > 0$, we define the population and empirical score-based affinities as

$$A_g(x,x') = \exp\{-\|g(x) - g(x')\|_2^2/(2h_g^2)\}, \qquad \widehat{A}_g(x,x') = \exp\{-\|\widehat{g}(x) - \widehat{g}(x')\|_2^2/(2h_g^2)\}.$$

Given a set $\mathcal{X} \subseteq \mathbb{R}^m$, constants $a, G > 0$, and n points in $\mathcal{X}$, we construct the affinity matrices A and $\widehat{A}$ by evaluating these functions and strictly setting their diagonals to zero. We define their degree matrices as $D = \operatorname{diag}(A\mathbf{1})$ and $\widehat{D} = \operatorname{diag}(\widehat{A}\mathbf{1})$, and we form the symmetrically normalized affinity matrices and the affinity error bound as

$$P = D^{-1/2}AD^{-1/2}, \qquad \widehat{P} = \widehat{D}^{-1/2}\widehat{A}\widehat{D}^{-1/2}, \qquad \Delta_A = (4Ga + 2a^2)/h_g^2.$$

For a selected number d of eigenvectors, we define U_d and $\widehat{U}_d$ as the $n \times d$ matrices containing the orthonormal eigenvectors that correspond to the exact same consecutive indices in the decreasingly ordered spectra of P and $\widehat{P}$. We let $\Theta(\widehat{U}_d, U_d)$ denote the diagonal matrix of principal angles between these two column spaces, and we set $d_0 = \frac{1}{2}\exp(-2G^2/h_g^2)$. The following proposition rigorously transfers the score error into an error bound for the normalized posterior affinity and its corresponding eigenmap.

Proposition 3. *Suppose*

$$\sup_{x\in\mathcal{X}} \|\widehat{g}(x) - g(x)\|_2 \le a, \qquad \sup_{x\in\mathcal{X}} \|g(x)\|_2 \le G.$$

If $n \ge 2$, then

$$\sup_{x,x'\in\mathcal{X}} |\widehat{A}_g(x,x') - A_g(x,x')| \le \Delta_A.$$

If $\Delta_A \le d_0/2$, then

$$\|\widehat{P} - P\|_{\mathrm{op}} \le C d_0^{-2} \Delta_A.$$

If these eigenvalues of P are separated from its remaining spectrum by $\gamma_n > 0$, then

$$\|\sin\Theta(\widehat{U}_d, U_d)\|_F \le \frac{C\sqrt{d}}{\gamma_n d_0^2}\Delta_A.$$

Proposition 3 confirms that the classification score and the resulting eigenmap depend exclusively on the exact same estimated object. On the stated event of Corollary 4, with exact selection and $Cs\, d_N(\alpha) \le \lambda_0/2$, the score calculation in its proof allows us to take $a \le C\sqrt{K}\, s\{d_N(\alpha) + \lambda_N\}$ uniformly over bounded sets. This establishes that the responsibility error affects both the direct classifier and the normalized eigenmap through the same moment rate, with additional dependence on the margin, bandwidth, and spectral gap in the respective bounds.

By utilizing the properties of the Gaussian affinity, we can also derive a relative perturbation bound. Importantly, this alternative bound does not require a uniform positive lower bound on the node degrees, nor does it require the restriction $\Delta_A \le d_0/2$.

Corollary 5. *Under the score bounds of Proposition 3, let $n \ge 2$ and use its indexed eigenspaces with $\gamma_n > 0$. Then*

$$\|\widehat{P} - P\|_{\mathrm{op}} \le \min\{2, \exp(2\Delta_A) - 1\}, \quad \|\sin\Theta(\widehat{U}_d, U_d)\|_F \le \frac{2\sqrt{d}}{\gamma_n}\min\{2, \exp(2\Delta_A) - 1\}.$$

Corollary 5 gives an eigenspace error bound of order $\sqrt{d}\,\Delta_A/\gamma_n$ as $\Delta_A \to 0$. Small node degrees do not introduce any additional factors. However, a small spectral gap or a narrow affinity bandwidth can still substantially increase the error.

Because the normalized affinity identifies an eigenspace rather than a specific eigenvector basis, the weighted ridge prediction in (15) must remain perfectly invariant under any rotation of that basis. To formalize this, we define $H = \mathcal{L} \cup \mathcal{U}$, let T_H denote the matrix containing the target rows t_i, and define the weight matrix $W_H = \mathrm{diag}(v_i : i \in H)$. We apply weighted centering via $\mathsf{C}_H Z = Z - \mathbf{1}(\mathbf{1}^\top W_H Z)/(\mathbf{1}^\top W_H \mathbf{1})$, and we measure the target scale as $M_T = \|W_H^{1/2}\mathsf{C}_H T_H\|_F$. We define the maximum row norm as $\|Z\|_{2,\infty} = \max_i \|Z_{i\cdot}\|_2$, and we let $\widehat{f}_U(i)$ denote the fitted prediction score vector at row i. The following proposition rigorously verifies that the spectral predictor is well-defined, and it transfers the eigenspace error directly into a bound on these fitted scores.

Proposition 4. *Let $U, \widetilde{U} \in \mathbb{R}^{n\times d_g}$ have orthonormal columns, $0 < v_i \le v_+$, and $\alpha \ge \alpha_0 > 0$. For every orthogonal $O \in \mathbb{R}^{d_g\times d_g}$, the fitted scores in* (15) *satisfy*

$$\widehat{f}_{UO}(i) = \widehat{f}_U(i), \qquad 1 \le i \le n.$$

If O minimizes $\|\widetilde{U} - UO\|_F$, $\varepsilon_F = \|\widetilde{U} - UO\|_F \le 1$, and $\varepsilon_\infty = \|\widetilde{U} - UO\|_{2,\infty}$, then

$$\max_{1\le i\le n} \|\widehat{f}_{\widetilde{U}}(i) - \widehat{f}_U(i)\|_2 \le C M_T(\varepsilon_F + \varepsilon_\infty),$$

where C depends only on d_g, v_+, and α_0. If also $\|t_i\|_2 \le 1$ for $i \in H$, then

$$\Big\{\frac{1}{n}\sum_{i=1}^n \|\widehat{f}_{\widetilde{U}}(i) - \widehat{f}_U(i)\|_2^2\Big\}^{1/2} \le C\varepsilon_F.$$

Proposition 4 establishes that signs and arbitrary rotations of the eigenvectors cannot alter the final spectral predictions. When the targets are valid probability vectors, the root mean squared bound securely transfers the Frobenius eigenspace error without introducing any factor that grows with the sample size. Proposition 3 bounds this eigenspace error after orthogonal alignment for the normalized posterior affinity. Because this comparison strictly concerns the predictions derived from the estimated and ideal graphs fitted to the identical targets, converting this stability bound into a disagreement bound still requires a margin condition on the reference spectral scores.

A.5 Posterior response and generated scores

The population posterior responsibilities precisely identify the class moments of the deployable features under the screened distribution. The following proofs establish this identity and bound the errors in both moment estimation and coordinate selection. Importantly, the same moment error enters the downstream bounds for classification, the affinity graph, and the spectral predictions under their respective conditions.

Proof of Proposition 2. Because $q_0(X,V)X_j^r$ is measurable with respect to (X,V),

$$\mathbb{E}\{q_0(X,V)\eta_{XV,k}(X,V)X_j^r\} = \mathbb{E}\left[q_0(X,V)X_j^r\mathbb{E}\{\mathbf{1}(C=k)\mid X,V\}\right] = \mathbb{E}\{q_0(X,V)\mathbf{1}(C=k)X_j^r\}.$$

The tower property also gives

$$\begin{aligned}\mathbb{P}\{C=k\mid \eta_X(X)\} &= \mathbb{E}\{\mathbf{1}(C=k)\mid \eta_X(X)\}\\ &= \mathbb{E}\left[\mathbb{E}\{\mathbf{1}(C=k)\mid X\}\mid \eta_X(X)\right] = \mathbb{E}\{\eta_{X,k}(X)\mid \eta_X(X)\} = \eta_{X,k}(X).\end{aligned}$$

If X_j is independent of $\{C, q_0(X,V)\}$, weighting by q_0 preserves independence of X_j and C. Hence $\mu_{qkj} = \mu_{qj}$ for every k, which gives $B_{qj} = 0$ and $\theta_{qj} = 0$. This completes the proof. □

While the algorithm computes the confidence fraction empirically, q_0 is defined as a fixed population weight. The following proposition bounds the difference between these two retention rules for soft posterior responsibilities. To establish this, we take N independent and identically distributed paired observations, denote the true posterior as $p_i = \eta_{XV}(X_i,V_i)$, and let $\widehat{p}_i$ be arbitrary fitted probability vectors. We fix a target fraction $0<\tau<1$ and a corresponding population threshold t_0, and we define the oracle retention weight as $q_0(X_i,V_i) = \mathbf{1}\{\max_k p_{ik} \ge t_0\}$. We then define $\widehat{t}_N$ as the empirical $(1-\tau)$-quantile of the maximum fitted scores $\max_k \widehat{p}_{ik}$. Using this empirical threshold, we set the empirical retention indicators $\widehat{q}_i = \mathbf{1}\{\max_k \widehat{p}_{ik} \ge \widehat{t}_N\}$ and record the actual retained fraction as $\widehat{\tau}_N = N^{-1}\sum_i \widehat{q}_i$, specifically keeping all ties at the threshold. We measure the error in the fitted posterior probabilities using the empirical root-mean-square norm $\epsilon_{\text{post}} = \left\{\frac{1}{N}\sum_{i=1}^N \|\widehat{p}_i - p_i\|_1^2\right\}^{1/2}$. Finally, we substitute the oracle target $R_i^0 = q_0(X_i,V_i)p_i$ and the generated responsibility $\widehat{R}_i = \widehat{q}_i\widehat{p}_i$ into the definition of ϵ_N.

Proposition 5. *Suppose, for constants $A,\beta>0$,*

$$\mathbb{P}\{\max_k p_{1k} \ge t_0\} = \tau, \qquad \mathbb{P}\{|\max_k p_{1k} - t_0| \le t\} \le At^\beta, \quad 0<t\le 1.$$

With probability at least $1-\alpha$, simultaneously for all fitted probability vectors,

$$\frac{1}{N}\sum_{i=1}^N |\widehat{q}_i - q_0(X_i,V_i)| \le C\left\{\epsilon_{\text{post}}^{2\beta/(\beta+2)} + \sqrt{\frac{\log(2/\alpha)}{N}}\right\} + |\widehat{\tau}_N - \tau|.$$

The constant C depends only on A and β.

Proposition 5 explicitly allows the teacher and the empirical threshold to be computed on the exact same observations. Its population conditions fix the oracle selected fraction at τ and bound the probability within distance t of the threshold by At^β for $0<t\le 1$. When the fitted confidences are distinct, empirical quantile thresholding automatically guarantees $|\widehat{\tau}_N - \tau| \le N^{-1}$. When ties exist, the bound gracefully adjusts by using the actual selected fraction. This retention bound subsequently controls the overall responsibility error via the inequality

$$\epsilon_N \le \{1 + \max_i \|X_i\|_\infty^2\}\left\{\epsilon_{\text{post}} + \frac{1}{N}\sum_{i=1}^N |\widehat{q}_i - q_0(X_i,V_i)|\right\}.$$

Under the sub-Gaussian bound in Theorem 7, the prefactor is at most $1 + C_X^2 \log(4Nm/\alpha)$ with probability at least $1 - \alpha/2$. Applying the proposition at level $\alpha/2$ gives an explicit high-probability bound for ϵ_N. If fitted preprocessing bounds every resulting coordinate by B_c, the corresponding covariate factor is at most $1 + B_c^2$. The confidence statement still uses the original independent pairs and their population responsibilities; it does not require independence of the transformed observations. Exact labeled responsibilities contribute zero to this discrepancy, and division by $c_\omega \geq 1$ cannot increase it. The same calculation bounds the paired-unlabeled contribution in a prespecified labeled and unlabeled design. Hard pseudo-labels still contribute their classification error; the proposition does not remove that error.

Proof of Proposition 5. Write $q_i = q_0(X_i, V_i)$, $N_0 = \sum_i q_i$, and $\delta_i = \|\widehat{p}_i - p_i\|_1$. For $u > 0$, put an index in B_u when $|\max_k p_{ik} - t_0| \leq 2u$ or $\delta_i > u$. Outside B_u, an index with $q_i = 1$ has fitted confidence above $t_0 + u$, while an index with $q_i = 0$ has fitted confidence below $t_0 - u$. Retention cannot reverse this ordering. Thus the smaller error count is at most $|B_u|$, and the two counts differ by $N\widehat{\tau}_N - N_0$. Hence

$$\sum_i |\widehat{q}_i - q_i| \leq 2|B_u| + |N\widehat{\tau}_N - N_0|, \qquad \frac{|B_u|}{N} \leq \frac{1}{N}\sum_i \mathbf{1}\{|\max_k p_{ik} - t_0| \leq 2u\} + \frac{\epsilon_{\text{post}}^2}{u^2}.$$

Let F and F_N be the population and empirical distribution functions of the oracle confidence $\max_k p_{ik}$. By the Dvoretzky–Kiefer–Wolfowitz inequality with the constant of Massart (1990), with probability at least $1 - \alpha$, $\sup_t |F_N(t) - F(t)| \leq d_\alpha$, where $d_\alpha = \{\log(2/\alpha)/(2N)\}^{1/2}$. On this single event, for every $0 < u \leq 1/2$ and every collection of fitted scores,

$$\frac{1}{N}\sum_i |\widehat{q}_i - q_i| \leq 2^{\beta+1} A u^\beta + \frac{2\epsilon_{\text{post}}^2}{u^2} + 5d_\alpha + |\widehat{\tau}_N - \tau|.$$

For $0 < \epsilon_{\text{post}} \leq 1$, take $u = \epsilon_{\text{post}}^{2/(\beta+2)}/2$. For $\epsilon_{\text{post}} = 0$, let $u \downarrow 0$. For $\epsilon_{\text{post}} > 1$, use the bound one for the disagreement fraction. This bound permits sample-dependent scores. Finally, $\|\widehat{q}_i\widehat{p}_i - q_i p_i\|_1 \leq \delta_i + |\widehat{q}_i - q_i|$. Cauchy–Schwarz bounds the average of δ_i by ϵ_{post} and proves the responsibility bound after multiplication by the largest covariate weight. This completes the proof. □

Proof of Theorem 7. For $r = 0, 1, 2$, put

$$M_{kjr} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\{R_{ik}^0 X_{ij}^r\}, \qquad \widetilde{M}_{kjr} = \frac{1}{N}\sum_{i=1}^N R_{ik}^0 X_{ij}^r, \qquad \widehat{M}_{kjr} = \frac{1}{N}\sum_{i=1}^N \widehat{R}_{ik} X_{ij}^r,$$

where $X_{ij}^0 = 1$. The variables $R_{ik}^0 X_{ij}$ and $R_{ik}^0 X_{ij}^2$ have uniformly bounded sub-Gaussian and sub-exponential norms, respectively. Bernstein's inequality and one union bound over all zeroth moments, first moments, and second cross moments give, on a common event of probability at least $1 - \alpha$,

$$\max_{k,j,r} |\widetilde{M}_{kjr} - M_{kjr}| \leq C_1 \left[\left\{ \frac{\log(6Km^2/\alpha)}{N} \right\}^{1/2} + \frac{\log(6Km^2/\alpha)}{N} \right].$$

For every $r \leq 2$,

$$|\widehat{M}_{kjr} - \widetilde{M}_{kjr}| \leq \frac{1}{N}\sum_{i=1}^N \|\widehat{R}_i - R_i^0\|_1 \{1 + \|X_i\|_\infty^2\} \leq \epsilon_N.$$

On the same event, the second cross-moment bounds and the responsibility error give

$$\max_{k,j,\ell} \left| \frac{1}{N}\sum_{i=1}^N \widehat{R}_{ik} X_{ij} X_{i\ell} - \frac{1}{N}\sum_{i=1}^N \mathbb{E}(R_{ik}^0 X_{ij} X_{i\ell}) \right| \leq C_2 d_N(\alpha).$$

Under either oracle choice, Proposition 2 gives $M_{kjr} = \mathbb{E}_P\{q_0(X, V)\mathbf{1}(C = k)X_j^r\}$. Write $M_{k0} = M_{kj0}$, $M_{\cdot 0} = \sum_k M_{k0}$, and define the corresponding sums for the empirical moments. The population quantities satisfy

$$\pi_{qk} = \frac{M_{k0}}{M_{\cdot 0}}, \qquad \mu_{qkj} = \frac{M_{kj1}}{M_{k0}}, \qquad \mu_{qj} = \frac{\sum_k M_{kj1}}{M_{\cdot 0}},$$

$$B_{qj} = \sum_k \pi_{qk}\mu_{qkj}^2 - \mu_{qj}^2, \qquad W_{qj} = \frac{\sum_k M_{kj2}}{M_{\cdot 0}} - \sum_k \pi_{qk}\mu_{qkj}^2.$$

The empirical formulas are identical. Because $M_{k0} \geq c_0$, $d_N(\alpha) \leq c$ gives $\widehat{M}_{k0} \geq c_0/2$ for small c. Bounded moments make all five maps Lipschitz, with constants depending only on K, c_0 and the sub-Gaussian parameter. Hence

$$\max_j \{|\widehat{B}_j - B_{qj}| + |\widehat{W}_j - W_{qj}|\} \leq C_3 d_N(\alpha).$$

Because $W_{qj} \geq w_0$, decreasing c gives $\widehat{W}_j \geq w_0/2$. The ratio identity

$$\frac{\widehat{B}_j}{\widehat{W}_j} - \frac{B_{qj}}{W_{qj}} = \frac{\widehat{B}_j - B_{qj}}{\widehat{W}_j} + \frac{B_{qj}(W_{qj} - \widehat{W}_j)}{\widehat{W}_j W_{qj}}$$

proves the first conclusion.

For a matrix A, write $\|A\|_{\max} = \max_{i,j} |A_{ij}|$. For fixed S, the first-moment bound and $\|v\|_2 \leq \sqrt{s}\|v\|_\infty$ control the means. Expanding the pooled covariance in cross moments and mean outer products, and using $\|A\|_{\mathrm{op}} \leq s\|A\|_{\max}$, gives

$$\max_k \|\widehat{\mu}_{k,S} - \mu_{qk,S}\|_2 + \|\widehat{\Sigma}_S - \Sigma_{q,S}\|_{\mathrm{op}} \leq C_4 s\, d_N(\alpha).$$

On $\max_j |\widehat{\theta}_j - \theta_{qj}| \leq C d_N(\alpha)$, the score gap separates S_q from its complement. The s largest scores select S_q.

For an affine map T with linear part A, use superscripts Z and T for the raw and transformed moments. Direct expansion gives

$$\widehat{\mu}_k^T - \mu_{qk}^T = A(\widehat{\mu}_k^Z - \mu_{qk}^Z), \qquad \widehat{\Sigma}^T - \Sigma_q^T = A(\widehat{\Sigma}^Z - \Sigma_q^Z)A^\top.$$

The row ℓ_1 bounds apply to every realized map when $d_N^Z(\alpha) \leq c$. For positive diagonal D, an affine map with linear part D^{-1} restricted to s coordinates gives mean error at most $C\|D^{-1}\|_{\mathrm{op}}\sqrt{s}\, d_N^Z(\alpha)$ and covariance error at most $C\|D^{-1}\|_{\mathrm{op}}^2 s\, d_N^Z(\alpha)$. Apply the same transformation to the population moments. Diagonal shrinkage contracts the transformed covariance error, as proved below.

For clipping, consider $f(z) = (-B_c) \vee \{(a^\top z + b) \wedge B_c\}$, where $a \in \mathbb{R}^{d_Z}$ and $b \in \mathbb{R}$. For $-B_c < t < B_c$, $f(z) > t$ is equivalent to $a^\top z + b > t$, a halfspace condition in $\mathbb{R}^{d_Z+1}$. On $2N$ observations, rounding function values to a grid of mesh B_c/N gives at most $(CN^2)^{C(d_Z+1)}$ distinct vectors by the halfspace growth bound. Pairwise products have approximation error CB_c^2/N and the same logarithmic count. Fixed oracle responsibilities in $[0,1]$ increase neither bound. Symmetrization, conditional Hoeffding bounds for the Rademacher sums, and bounded differences give the stated uniform sampling term; see van der Vaart (1998, Chapter 19) for the empirical-process inequalities. Center each summand under P_i.

Their average expectation is under P. The fixed oracle on the raw observation preserves the posterior identity under every transformation. Estimated responsibilities add at most ϵ_N to each moment. The ratio, inverse-covariance, and score calculations then apply uniformly, with the transformed population quantities. This completes the proof. □

Proof of Corollary 4. If $\beta = 0$, letting $t \to \infty$ in the margin condition gives $A \geq 1$, and the disagreement bound is immediate. Suppose $\beta > 0$. On the stated high-probability event, Theorem 7 gives $\widehat{S}_q = S_q$ and

$$\max_k \|\widehat{\mu}_{k,S_q} - \mu_{qk,S_q}\|_2 + \|\widehat{\Sigma}_{S_q} - \Sigma_{q,S_q}\|_{\mathrm{op}} \leq Cs\, d_N(\alpha).$$

When $Cs\, d_N(\alpha) \leq \lambda_0/2$, the smallest eigenvalue of $\widehat{\Sigma}_{S_q} + \lambda_N I_s$ is at least $\lambda_0/2 + \lambda_N$. The resolvent identity gives

$$\|(\widehat{\Sigma}_{S_q} + \lambda_N I_s)^{-1} - \Sigma_{q,S_q}^{-1}\|_{\mathrm{op}} \leq C_{\lambda_0} s\{d_N(\alpha) + \lambda_N\}.$$

The first-moment calculation also gives $\max_k \|\widehat{\mu}_{k,S_q} - \mu_{qk,S_q}\|_2 \le C\sqrt{s}\, d_N(\alpha)$. Put $L = \max_k \|\mu_{qk,S_q}\|_2$. Expanding the linear and quadratic score terms bounds their error by $C\{ML + L^2 + M + L + 1\}s\{d_N(\alpha) + \lambda_N\}$. With this dependence included in C_M, the score formula implies

$$\sup_{\{x:\|x_{S_q}\|_2 \le M\}} \max_k |\widehat{g}_k(x) - g_k(x)| \le C_M s\{d_N(\alpha) + \lambda_N\} =: a_N.$$

On $\{\|X_{S_q}\|_2 \le M\}$, class disagreement requires a population score gap at most $2a_N$. The margin condition gives

$$\mathbb{P}_q\{\arg\max_k \widehat{g}_k(X) \ne \arg\max_k g_k(X)\} \le \mathbb{P}_q(\|X_{S_q}\|_2 > M) + A(2a_N)^\beta,$$

and the factor 2 is absorbed into C_M. If $Cs\, d_N(\alpha) > \lambda_0/2$, increasing C_M makes the claimed probability bound at least one, which proves the complementary case. Under the stated Gaussian model, the log class density of X_{S_q} differs from g_k by a term common to all classes. Equal class weights hence make $\arg\max_k g_k$ Bayes optimal among classifiers based on X_{S_q}.

For the shrinkage version, put $E = \widehat{\Sigma}_{S_q} - \Sigma_{q,S_q}$. Symmetry gives $\|\operatorname{diag}(E)\|_{\mathrm{op}} = \max_j |E_{jj}| \le \|E\|_{\mathrm{op}}$, and hence $\|(1-\tau_{\mathrm{cov}})E + \tau_{\mathrm{cov}} \operatorname{diag}(E)\|_{\mathrm{op}} \le \|E\|_{\mathrm{op}}$. Both Σ_{q,S_q} and $\operatorname{diag}(\Sigma_{q,S_q})$, and hence their convex combination, are bounded below by $\lambda_0 I_s$. The same resolvent, score, and margin bounds apply. If $\tau_{\mathrm{cov}} = 0$ or Σ_{q,S_q} is diagonal, the population covariance is unchanged, and the Gaussian conclusion follows. This completes the proof. □

Proof of Proposition 3. Put $u = g(x) - g(x')$, $\widehat{u} = \widehat{g}(x) - \widehat{g}(x')$, and $e = \widehat{u} - u$. The premises give $\|u\|_2 \le 2G$ and $\|e\|_2 \le 2a$. Hence

$$\left|\|\widehat{u}\|_2^2 - \|u\|_2^2\right| = |2u^\top e + \|e\|_2^2| \le 8Ga + 4a^2.$$

The derivative of $z \mapsto \exp\{-z/(2h_g^2)\}$ has absolute value at most $(2h_g^2)^{-1}$ for $z \ge 0$. The mean value theorem proves the affinity bound.

Let $\overline{A} = A/n$, $\widehat{\overline{A}} = \widehat{A}/n$, $\overline{D} = D/n$, and $\widehat{\overline{D}} = \widehat{D}/n$. The entrywise bound gives

$$\|\widehat{\overline{A}} - \overline{A}\|_{\mathrm{op}} \le \|\widehat{\overline{A}} - \overline{A}\|_F \le \Delta_A, \qquad \|\widehat{\overline{D}} - \overline{D}\|_{\mathrm{op}} \le \Delta_A.$$

For $i \ne j$, the score bound gives $A_{ij} \ge \exp(-2G^2/h_g^2)$. Because $n \ge 2$, it follows that $\min_i D_{ii}/n \ge \{(n-1)/n\}\exp(-2G^2/h_g^2) \ge d_0$. This bound and $\Delta_A \le d_0/2$ imply

$$\|\overline{D}^{-1/2}\|_{\mathrm{op}} \le d_0^{-1/2}, \qquad \|\widehat{\overline{D}}^{-1/2}\|_{\mathrm{op}} \le (2/d_0)^{1/2}.$$

The scalar mean value theorem gives

$$\|\widehat{\overline{D}}^{-1/2} - \overline{D}^{-1/2}\|_{\mathrm{op}} \le C d_0^{-3/2}\Delta_A.$$

Because every affinity is at most one, $\|\overline{A}\|_{\mathrm{op}} \le 1$. Expanding

$$\widehat{P} - P = \widehat{\overline{D}}^{-1/2}(\widehat{\overline{A}} - \overline{A})\widehat{\overline{D}}^{-1/2} + (\widehat{\overline{D}}^{-1/2} - \overline{D}^{-1/2})\overline{A}\widehat{\overline{D}}^{-1/2} + \overline{D}^{-1/2}\overline{A}(\widehat{\overline{D}}^{-1/2} - \overline{D}^{-1/2})$$

gives $\|\widehat{P} - P\|_{\mathrm{op}} \le C d_0^{-2}\Delta_A$. If $\|\widehat{P} - P\|_{\mathrm{op}} \ge \gamma_n/2$, use $\|\sin\Theta(\widehat{U}_d, U_d)\|_F \le \sqrt{d}$. If $\|\widehat{P} - P\|_{\mathrm{op}} < \gamma_n/2$, Weyl's inequality and the Davis–Kahan theorem give the same bound:

$$\|\sin\Theta(\widehat{U}_d, U_d)\|_F \le \frac{2\sqrt{d}}{\gamma_n}\|\widehat{P} - P\|_{\mathrm{op}},$$

which proves the eigenspace bound. This completes the proof. □

Proof of Corollary 5. The squared-distance calculation in Proposition 3 gives

$$\left|\log\frac{\widehat{A}_{ij}}{A_{ij}}\right| \le \Delta_A, \qquad i \ne j.$$

Thus the affinity entries and their row sums satisfy

$$e^{-\Delta_A} A_{ij} \le \widehat{A}_{ij} \le e^{\Delta_A} A_{ij}, \qquad e^{-\Delta_A} D_{ii} \le \widehat{D}_{ii} \le e^{\Delta_A} D_{ii}.$$

All degrees are positive. Normalization gives the entrywise inequalities

$$e^{-2\Delta_A} P_{ij} \le \widehat{P}_{ij} \le e^{2\Delta_A} P_{ij}, \qquad |\widehat{P}_{ij} - P_{ij}| \le \{e^{2\Delta_A} - 1\} P_{ij}.$$

The symmetric matrix P has nonnegative entries and operator norm one. For $v \in \mathbb{R}^n$, let $|v|$ denote componentwise absolute values. Then

$$|v^\top (\widehat{P} - P) v| \le \{e^{2\Delta_A} - 1\} |v|^\top P |v| \le \{e^{2\Delta_A} - 1\} \|v\|_2^2.$$

Also $\|\widehat{P} - P\|_{\mathrm{op}} \le 2$. If this perturbation norm is at least $\gamma_n/2$, the asserted eigenspace bound follows from $\|\sin\Theta\|_F \le \sqrt{d}$. If the perturbation norm is less than $\gamma_n/2$, Weyl's inequality separates the fitted cluster from the complementary ideal eigenvalues by at least $\gamma_n/2$. The Davis–Kahan bound for this separation gives the second assertion. This completes the proof. □

Proof of Proposition 4. Let

$$\overline{U}_H = \frac{\mathbf{1}^\top W_H U_H}{\mathbf{1}^\top W_H \mathbf{1}}, \qquad \overline{T}_H = \frac{\mathbf{1}^\top W_H T_H}{\mathbf{1}^\top W_H \mathbf{1}},$$

and put $Z = W_H^{1/2} \mathsf{C}_H U_H$ and $Y = W_H^{1/2} \mathsf{C}_H T_H$. The normal equations for (15) give

$$\widehat{B}_U = (Z^\top Z + \alpha I)^{-1} Z^\top Y, \qquad \widehat{b}_U = \overline{T}_H - \overline{U}_H \widehat{B}_U.$$

For orthogonal O, the centered design is ZO, and

$$\widehat{B}_{UO} = O^\top \widehat{B}_U, \qquad \widehat{b}_{UO} = \widehat{b}_U.$$

This proves $\widehat{f}_{UO}(i) = \widehat{f}_U(i)$.

By this invariance, replace UO by U and write $E = \widetilde{U} - U$. If $\widetilde{Z} = W_H^{1/2} \mathsf{C}_H \widetilde{U}_H$, weighted centering is a contraction in the W_H norm. Orthonormality of U gives

$$\|\widetilde{Z} - Z\|_F \le v_+^{1/2} \varepsilon_F, \qquad \|Z\|_F \le v_+^{1/2} \sqrt{d_g}, \qquad \|Y\|_F = M_T.$$

Put $A = Z^\top Z + \alpha I$ and $\widetilde{A} = \widetilde{Z}^\top \widetilde{Z} + \alpha I$. Then

$$\|\widetilde{A} - A\|_{\mathrm{op}} \le 2\|Z\|_F \|\widetilde{Z} - Z\|_F + \|\widetilde{Z} - Z\|_F^2 \le C_1(\varepsilon_F + \varepsilon_F^2), \quad \|\widetilde{Z}^\top Y - Z^\top Y\|_F \le v_+^{1/2} M_T \varepsilon_F.$$

Both A and $\widetilde{A}$ have minimum eigenvalue at least α_0. The resolvent identity and $\varepsilon_F \le 1$ yield

$$\|\widehat{B}_{\widetilde{U}} - \widehat{B}_U\|_F \le C_2 M_T \varepsilon_F, \qquad \|\widehat{B}_{\widetilde{U}}\|_F \le C_3 M_T,$$

where the constants depend only on d_g, v_+, and α_0.

Let $\overline{E}_H = (\mathbf{1}^\top W_H E_H)/(\mathbf{1}^\top W_H \mathbf{1})$. Positive weights give $\|\overline{E}_H\|_2 \le \varepsilon_\infty$. Every row of an orthonormal-column matrix has norm at most one, and hence

$$\|\widehat{f}_{\widetilde{U}}(i) - \widehat{f}_U(i)\|_2 \le \|E_i - \overline{E}_H\|_2 \|\widehat{B}_{\widetilde{U}}\|_F + \|U_i - \overline{U}_H\|_2 \|\widehat{B}_{\widetilde{U}} - \widehat{B}_U\|_F \le C M_T (\varepsilon_\infty + \varepsilon_F).$$

Taking the maximum over i proves the rowwise bound.

For the mean squared bound, put $S_v = \sum_{i \in H} v_i$. Cauchy–Schwarz and $\sum_{i \in H} v_i^2 \le v_+ S_v$ give

$$\|\overline{E}_H\|_2 \le \sqrt{v_+/S_v}\, \varepsilon_F, \qquad \|\overline{U}_H\|_2 \le \sqrt{v_+/S_v}.$$

Thus,

$$\|E - \mathbf{1}\overline{E}_H\|_F \le \{1 + \sqrt{n v_+/S_v}\} \varepsilon_F, \qquad \|U - \mathbf{1}\overline{U}_H\|_{\mathrm{op}} \le 1 + \sqrt{n v_+/S_v}.$$

Apply the coefficient bounds above to all prediction rows. The common response intercept cancels, and hence

$$\Big\{ \frac{1}{n} \sum_{i=1}^n \|\widehat{f}_{\widetilde{U}}(i) - \widehat{f}_U(i)\|_2^2 \Big\}^{1/2} \le C M_T \{ n^{-1/2} + \sqrt{v_+/S_v} \} \varepsilon_F.$$

Weighted centering gives $M_T^2 \le \sum_{i \in H} v_i \|t_i\|_2^2 \le S_v$, and $S_v \le n v_+$. Substitution gives the asserted bound. This completes the proof. □

A.6 CONTINUOUS PREDICTIVE REPRESENTATION

Prediction from a finite representation depends on how well its eigenfunctions approximate the conditional mean. Proposition 6 separates conditional-mean approximation from metric and spectral estimation. Proposition 7 bounds the approximation term through population prediction bias. The following theorem gives a continuous model in which the first few coordinates represent the entire conditional mean.

Let $X = (X^{(1)}, \dots, X^{(K)})$, where the $K \geq 2$ blocks are independent and uniform on the sphere of radius $\sqrt{s}$ in $\mathbb{R}^s$, and let $m = Ks$. The prescribed response is $V = X^{(1)} + \sigma\varepsilon$, with independent Rademacher coordinates in ε, independent of X. Use $h^2 = m\tau^2$, $\rho_m = (s\tau^2 m)^{-1}$, and the certified calculation in Proposition 8, with screening cutoff $(s\tau^2)^{-1}$. Put $\theta = (a - \delta\mathbf{1})/\{m(1-\delta)\}$, and write θ_0 and $\widehat{\theta}$ for this transformation of a_0 and $\widehat{a}$. For positive sequences, $a_n \asymp b_n$ means that a_n/b_n is bounded above and away from zero. Put $N = Rb$, and $e_{R,b} = R^{-1/2} + \{\log(2b)/b\}^{1/2} + \eta_R^{1/2}$. Fit least squares to V using an intercept and the first s learned coordinates on the metric observations. Denote its predictions on additional observations by $\widehat{f}_i$, $i \in I_{\mathrm{ev}}$; their covariates may enter the graph.

Theorem 8. *Let $s \geq 3$, $\sigma \geq 0$, and $\delta \in (0,1)$ be fixed. For fixed sufficiently large τ, the unique population minimizer satisfies*

$$a_{0j} = \delta + \frac{m(1-\delta)}{s} \quad (j \leq s), \qquad a_{0j} = \delta \quad (j > s).$$

If $R, b \to \infty$, $\log(2m)/R + \log(2mb)/b \to 0$, $\eta_R \to 0$, $N \asymp |I_{\mathrm{ev}}| \asymp n$, and k_n satisfies Theorem 19, then, for fixed $0 < t_N < 1 - \delta$,

$$\mathbb{P}(\widehat{S} = \{1, \dots, s\}) \to 1, \qquad \|\widehat{\theta} - \theta_0\|_2 = O_{\mathbb{P}}(e_{R,b}),$$

and

$$\frac{1}{s|I_{\mathrm{ev}}|} \sum_{i \in I_{\mathrm{ev}}} \|\widehat{f}_i - X_i^{(1)}\|_2^2 = O_{\mathbb{P}}\{(k_n/n)^{2/(s-1)} + e_{R,b}^2 + N^{-1}\}.$$

Theorem 8 derives the population metric and its predictive eigenspace from the distribution. The first s eigenfunctions may be chosen as the coordinates of $X^{(1)}$, and the stated bound controls prediction of this conditional mean. If error is measured against an independent response V, the normalized mean squared prediction error approaches σ^2. Dependence within the other sphere blocks does not prevent their removal. A response-independent representation that treats these blocks symmetrically has a strictly larger population prediction lower bound in (18). The comparison does not apply to methods given the response view. Corollary 8 extends this prediction result to nonlinear responses on the same sphere and permits estimated coordinate scales.

Proof of Theorem 8. Put $S_0 = \{1, \dots, s\}$ and $\bar{\rho} = (s\tau^2)^{-1}$. In Lemma 2, take $B_X = \sqrt{s}$, $B_V = \sqrt{s} + \sigma$, and any fixed $b_0 \geq 3$. Its constants satisfy $C_{\mathrm{inv}} = O(1)$ and $C_{H,2} = O(\tau^{-4})$ as $\tau \to \infty$. Choose τ such that the lemma applies and

$$\bar{\rho} > sC_{H,2}, \qquad \frac{2}{s\tau^2} - C_{H,2} - (4C_{H,2} + \bar{\rho})(1-\delta) > 0.$$

The constants are independent of K. For the numerical values $s = 3$, $\sigma = 0.4$, $\delta = 0.02$, $\tau = 300$, and $b_0 = 64$, the curvature lower bound exceeds 2.50×10^{-6} and the second left side exceeds 1.97×10^{-6}.

By Proposition 1, the population prediction operator applied to $f_0(X) = X^{(1)}$ is independent of the metric on the other blocks. Their unpenalized risk derivatives are exactly zero. For the following derivative calculation only, extend the metric domain to $a \geq 0$ with $\mathbf{1}^\top a \leq m$, as permitted by Lemma 2. At $a = 0$, the normalized kernel is one. Centering the log-kernel derivative with its two scaling derivatives gives

$$\partial_{a_j} W_a(x, x')\big|_{a=0} = \frac{x_j x_j'}{m\tau^2}.$$

Every observed coordinate has mean zero and variance one, and $\mathrm{cov}(V, X_j)$ is the jth coordinate vector for $j \leq s$ and zero for $j > s$. Thus

$$m\partial_{a_j} \mathcal{R}_{\infty,0}(a, h)\big|_{a=0} = -\frac{2}{s\tau^2} \mathbf{1}_{\{j \leq s\}}.$$

Integrate the Hessian bound along $a = t\mathbf{1}$, $0 \le t \le 1$. The ridge gradient equals zero at $\mathbf{1}$, and hence

$$\max_{j \le s} m\partial_{a_j} \mathcal{R}_{\infty,\rho_m}(\mathbf{1}, h) \le -\frac{2}{s\tau^2} + C_{H,2}, \qquad \partial_{a_j} \mathcal{R}_{\infty,\rho_m}(\mathbf{1}, h) = 0 \quad (j > s).$$

The parameter inequalities give population $U(S_0) < 0$. By Proposition 8, every minimizer lies on S_0, where the Hessian is bounded below by $\kappa = (\bar{\rho} - sC_{H,2})(1-\delta)^2 > 0$. Invariance under simultaneous permutations of the first s observed and response coordinates and uniqueness give $\theta_{0j} = 1/s$ on S_0.

Apply Lemma 3 at $a = \mathbf{1}$ with $d_0 = m$. Its error tends to zero under the two growth conditions. Because $C_{H,2} < 1/(s\tau^2)$, screening at cutoff $(s\tau^2)^{-1}$ selects exactly S_0 with probability tending to one. The empirical exchange bound is strictly negative with probability tending to one, and the curvature is at least κ. Thus certification precedes restriction to the s coordinates.

On this face, apply Lemma 3 at a_0 with $d_0 = s$. The population gradient has equal components on S_0. For the returned point with objective gap at most η_R, strong convexity gives, with $v = \widehat{\theta} - \theta_0$,

$$\frac{\kappa}{2}\|v\|_2^2 \le \left\| \nabla_{\theta,S_0} \widehat{\mathcal{R}}_{\rho_m}(a_0, h) - \nabla_{\theta,S_0} \mathcal{R}_{\infty,\rho_m}(a_0, h) \right\|_2 \|v\|_2 + \eta_R.$$

The score norm is $O_{\mathbb{P}}\{R^{-1/2} + \sqrt{\log(2b)/b}\}$. Solving this quadratic inequality proves $\|v\|_2 = O_{\mathbb{P}}(e_{R,b})$. The computed point lies on the certified face, with active coordinates converging to $1/s$ in θ. Hence thresholding selects S_0. Also

$$\widehat{a}_j^{\mathrm{rep}} - 1 = \frac{sm(1-\delta)}{s\delta + m(1-\delta)}(\widehat{\theta}_j - 1/s), \qquad j \in S_0.$$

Thus $\|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2 = O_{\mathbb{P}}(e_{R,b})$, where the population representation metric equals one on S_0 and zero elsewhere.

The selected manifold is the sphere of radius $\sqrt{s}$ with uniform density. Its operator in (11) is a positive constant times the negative Laplace–Beltrami operator. For a homogeneous harmonic polynomial of degree ℓ in $\mathbb{R}^s$, the polar-coordinate formula for the Euclidean Laplacian gives the spherical eigenvalue $\ell(\ell+s-2)/s$. Harmonic decomposition and density of polynomial restrictions give completeness. Degrees zero, one, and two have eigenvalues proportional to 0, $(s-1)/s$, and 2. Thus the first s nonconstant eigenfunctions are the orthonormal coordinates of $X^{(1)}$, with a positive external eigengap.

Condition 3 now holds with $d = s - 1$. The uniform spectral argument in Theorem 19 gives

$$\inf_{O^\top O = I_s} \|\widehat{Y} - n^{-1/2} X^{(1)} O\|_F = O_{\mathbb{P}}\{(k_n/n)^{1/(s-1)} + e_{R,b}\}.$$

The conditional mean belongs to this span. In Proposition 6, take $\varphi = (1, X^{(1)\top})^\top$, $u = 0$, and replace r_n by the preceding spectral rate. Its zero-mean score has order $N^{-1/2}$, and both terms involving the learned-coordinate error have order $(k_n/n)^{1/(s-1)} + e_{R,b}$. Squaring gives the stated prediction bound. This completes the proof. □

B Prediction from learned spectral coordinates

A spectral approximation properly controls the coordinates of a learned representation. However, any subsequent prediction also depends on the portion of the conditional mean that falls outside the span of these coordinates. In this section, we first separate these two sources of error. We then relate the resulting approximation error to the predictive criterion. The continuous model of Theorem 8 has zero approximation error for its stated representation. Spectral series regression relies on this identical distinction between approximation and estimation errors (Lee & Izbicki, 2016). In our framework, the metric estimation and the final regression coordinates may depend on the exact same responses.

B.1 Prediction error after metric estimation

Let $f_0(X) = \mathbb{E}(V \mid X)$ denote the true conditional mean, where $V \in \mathbb{R}^{q_V}$ and $\|V\|_\infty \le B_V$. For a fixed population metric, we stack the population spectral coordinates into $\varphi(Z) =$

$(1, \psi_2(Z), \ldots, \psi_{p+1}(Z))^\top$. Using this vector, we define the optimal projection coefficients and the corresponding approximation error as

$$B_p = \mathbb{E}\{\varphi(Z) f_0(X)^\top\}, \qquad A_p = q_V^{-1}\mathbb{E}\|f_0(X) - B_p^\top \varphi(Z)\|_2^2.$$

Because the functions ψ_j are orthonormal in $L^2(P)$ and strictly orthogonal to constants, A_p measures the conditional-mean error of the best linear prediction derived from the population spectral coordinates, including an intercept. Consequently, A_p captures the information lost through both feature selection and spectral truncation.

To evaluate this representation, let I_{tr} index the N metric observations used for training, and let I_{ev} index the additional independent observations used for evaluation. The graph construction may incorporate the covariates from both of these sets. For the training observations $i \in I_{\mathrm{tr}}$, we fit a least-squares model to V_i using an intercept and the first p learned spectral coordinates. We let $\widehat{f}_i$ denote the resulting predictions on the evaluation set I_{ev}. If the least-squares design matrix is singular, we apply the Moore–Penrose inverse, though under the conditions below, this convention remains inactive with probability tending to one. We define the combined spectral error rate as

$$r_n = \frac{(k_n/n)^{1/d} + \omega_{R,J,b}}{\gamma_p}.$$

Proposition 6. *Assume the conditions of Corollary 1 for the prescribed response view. Let m, p, q_V, and the population representation metric be fixed, and suppose $N \asymp |I_{\mathrm{ev}}| \asymp n$ and $r_n \to 0$. Then*

$$\frac{1}{q_V |I_{\mathrm{ev}}|} \sum_{i \in I_{\mathrm{ev}}} \|\widehat{f}_i - f_0(X_i)\|_2^2 = O_{\mathbb{P}}\{A_p + r_n^2 + N^{-1}\}.$$

Proposition 6 bounds downstream prediction error using the eigenspace error. Specifically, the square of this spectral error controls the additional squared error introduced by learning the representation. However, the inherent approximation term A_p requires separate control. Crucially, this bound never requires the fitted coordinates to be independent of the training response errors. Similar requirements for compact support, bounded positive kernels, and spectral separation appear in the spectral-series conditions of Lee & Izbicki (2016, Section 5). While their analysis assumes independent samples for basis estimation and coefficient estimation to simplify the proofs, the calculation below formally allows the exact same observations to be reused for both tasks, and it permits repeated eigenvalues within the selected subspace.

Proof of Proposition 6. Write $u_i = f_0(X_i) - B_p^\top \varphi(Z_i)$ and $\varepsilon_i = V_i - f_0(X_i)$. The finite collection of population eigenfunctions is bounded on the compact manifold. Hence

$$\mathbb{E}(\varphi\varphi^\top) = I_{p+1}, \qquad \mathbb{E}(\varphi u^\top) = 0, \qquad \mathbb{E}(\varepsilon \mid X) = 0, \qquad \|B_p\|_F < \infty.$$

Let F_0 have rows $\varphi(Z_i)^\top$. Rotate the learned coordinates and set $\widehat{F} = (\mathbf{1}, \sqrt{n}\widehat{Y}O)$ and $D = \widehat{F} - F_0$. The rotation does not change least-squares predictions. Corollary 1 gives

$$n^{-1/2}\|D\|_F = O_{\mathbb{P}}(r_n), \qquad N^{-1}F_{0,\mathrm{tr}}^\top F_{0,\mathrm{tr}} = I_{p+1} + O_{\mathbb{P}}(N^{-1/2}).$$

Because $N \asymp n$, the same Gram matrix with $\widehat{F}_{\mathrm{tr}}$ equals $I_{p+1} + o_{\mathbb{P}}(1)$. On its invertibility set, the fitted coefficient matrix satisfies

$$\widehat{B} - B_p = \Big(\frac{\widehat{F}_{\mathrm{tr}}^\top \widehat{F}_{\mathrm{tr}}}{N}\Big)^{-1}\Big\{\frac{F_{0,\mathrm{tr}}^\top(u_{\mathrm{tr}} + \varepsilon_{\mathrm{tr}})}{N} + \frac{D_{\mathrm{tr}}^\top(u_{\mathrm{tr}} + \varepsilon_{\mathrm{tr}})}{N} - \frac{\widehat{F}_{\mathrm{tr}}^\top D_{\mathrm{tr}} B_p}{N}\Big\}.$$

The first term in braces is $O_{\mathbb{P}}(N^{-1/2})$ by its zero mean and bounded second moment. Cauchy–Schwarz bounds the remaining terms by $O_{\mathbb{P}}(r_n)$, allowing dependence between D and $\varepsilon_{\mathrm{tr}}$. Thus $\|\widehat{B} - B_p\|_F = O_{\mathbb{P}}(N^{-1/2} + r_n)$. On the evaluation observations,

$$\widehat{f} - f_0 = \widehat{F}_{\mathrm{ev}}(\widehat{B} - B_p) + D_{\mathrm{ev}} B_p - u_{\mathrm{ev}}.$$

The first two terms have squared Frobenius norm divided by $|I_{\mathrm{ev}}|$ of order $O_{\mathbb{P}}(N^{-1} + r_n^2)$. Also

$$\mathbb{E}\left\{\frac{\|u_{\mathrm{ev}}\|_F^2}{q_V |I_{\mathrm{ev}}|}\right\} = A_p.$$

Markov's inequality gives $\|u_{\mathrm{ev}}\|_F^2/(q_V|I_{\mathrm{ev}}|) = O_{\mathbb{P}}(A_p)$, with equality to zero almost surely when $A_p = 0$. This completes the proof. □

We can also use a learned metric to predict a completely new response that was not used during metric estimation. Suppose a new response $Y \in \mathbb{R}^{q_Y}$ satisfies the finite variance condition $\mathbb{E}\|Y\|_2^2 < \infty$, and assume the complete observations, now including Y, remain independent copies. We denote its true conditional mean as $g_0(X) = \mathbb{E}(Y \mid X)$. Using the same population spectral coordinates φ defined previously, we establish the new optimal projection coefficients and the corresponding approximation error for this downstream task as

$$B_p^Y = \mathbb{E}\{\varphi(Z)g_0(X)^\top\}, \qquad A_p^Y = q_Y^{-1}\mathbb{E}\|g_0(X) - (B_p^Y)^\top\varphi(Z)\|_2^2.$$

To evaluate downstream prediction, we let I_{lab} index a subset of M observations selected strictly independently of the observed values. This labeled set may overlap with the observations originally used to estimate the metric. For these M observations, we center and standardize the learned coordinates, and we apply diagonal spectral penalty weights restricted to $[0, 1]$. When viewed in the aligned coordinate system of Proposition 6, these operations construct a positive semidefinite coefficient penalty matrix P_M that safely leaves the intercept unpenalized. We let Y_{lab} and $\widehat{F}_{\mathrm{lab}}$ represent the downstream response matrix and the aligned design matrix evaluated on I_{lab}, respectively, and we let $\widehat{g}_i$ denote the final ridge predictions.

Fixing a maximum penalty $\bar{\alpha} < \infty$, the following corollary formally separates the original response used to estimate the metric from this new response predicted exclusively after the representation is formed.

Corollary 6. *Assume the conditions of Corollary 1 for the metric response. Let m, p, q_Y, and the population representation metric be fixed, and suppose $N \asymp M \asymp |I_{\mathrm{ev}}| \asymp n$ and $r_n \to 0$. If $0 \le \alpha_M \le \bar{\alpha} < \infty$ and $\widehat{B}_\alpha$ minimizes*

$$\|Y_{\mathrm{lab}} - \widehat{F}_{\mathrm{lab}}B\|_F^2 + \alpha_M \operatorname{tr}(B^\top P_M B),$$

then

$$\frac{1}{q_Y|I_{\mathrm{ev}}|}\sum_{i\in I_{\mathrm{ev}}}\|\widehat{g}_i - g_0(X_i)\|_2^2 = O_{\mathbb{P}}\{A_p^Y + r_n^2 + M^{-1}\}.$$

Corollary 6 allows a fixed, positive fraction of the original graph observations to possess these downstream responses. The bound systematically covers standard empirical operations, including coordinate standardization, the application of diagonal spectral penalties, and the data-dependent selection of the penalty α_M from any fixed finite set. Crucially, the approximation term A_p^Y remains completely specific to this downstream conditional mean. Furthermore, while the original response used to estimate the metric still requires the strict boundedness condition, this new downstream response requires only a finite second moment.

Proof of Corollary 6. Write $u_i = g_0(X_i) - (B_p^Y)^\top\varphi(Z_i)$ and $\varepsilon_i = Y_i - g_0(X_i)$. Use F_0, $\widehat{F}$, and D as in the proof of Proposition 6. Corollary 1 and the value-independent selection of I_{lab} give

$$M^{-1}F_{0,\mathrm{lab}}^\top F_{0,\mathrm{lab}} = I_{p+1} + O_{\mathbb{P}}(M^{-1/2}), \qquad M^{-1/2}\|D_{\mathrm{lab}}\|_F = O_{\mathbb{P}}(r_n).$$

The first relation and the bounded spectral penalty weights imply $\|P_M\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$. The ridge normal equation gives

$$\widehat{B}_\alpha - B_p^Y = \left(\frac{\widehat{F}_{\mathrm{lab}}^\top\widehat{F}_{\mathrm{lab}}}{M} + \frac{\alpha_M P_M}{M}\right)^{-1} \times \left\{\frac{F_{0,\mathrm{lab}}^\top(u_{\mathrm{lab}} + \varepsilon_{\mathrm{lab}})}{M} + \frac{D_{\mathrm{lab}}^\top(u_{\mathrm{lab}} + \varepsilon_{\mathrm{lab}})}{M} - \frac{\widehat{F}_{\mathrm{lab}}^\top D_{\mathrm{lab}}B_p^Y}{M} - \frac{\alpha_M P_M B_p^Y}{M}\right\}.$$

The inverse is bounded in probability. Bounded eigenfunctions, Jensen's inequality, and $\mathbb{E}\|Y\|_2^2 < \infty$ give

$$\mathbb{E}\|u_i + \varepsilon_i\|_2^2 < \infty, \qquad \mathbb{E}\{\varphi(Z_i)(u_i + \varepsilon_i)^\top\} = 0.$$

The second identity follows from B_p^Y and conditional centering. Independence and Chebyshev's inequality bound the first term by $O_{\mathbb{P}}(M^{-1/2})$. Markov's inequality gives $M^{-1/2}\|u_{\mathrm{lab}} + \varepsilon_{\mathrm{lab}}\|_F = O_{\mathbb{P}}(1)$. Cauchy–Schwarz then bounds the next two terms by $O_{\mathbb{P}}(r_n)$, without independence between

the learned coordinates and the downstream response errors. The final term is $O_{\mathbb{P}}(M^{-1})$. Thus $\|\widehat{B}_\alpha - B_p^Y\|_F = O_{\mathbb{P}}(M^{-1/2} + r_n)$. On the evaluation observations,

$$\widehat{g} - g_0 = \widehat{F}_{\mathrm{ev}}(\widehat{B}_\alpha - B_p^Y) + D_{\mathrm{ev}} B_p^Y - u_{\mathrm{ev}}.$$

The first two terms contribute $O_{\mathbb{P}}(M^{-1} + r_n^2)$ to average squared error. Markov's inequality gives $\|u_{\mathrm{ev}}\|_F^2/(q_Y|I_{\mathrm{ev}}|) = O_{\mathbb{P}}(A_p^Y)$. This completes the proof. □

This identical result directly controls conditional excess classification risk. If the downstream task is to predict a class label C, we define the indicator variables $Y_c = \mathbf{1}\{C = c\}$ for $c = 1, \ldots, q_Y$, and we write the true class probabilities as $\eta_c(x) = \mathbb{P}(C = c \mid X = x)$. We then let $c^*(x)$ denote the true Bayes-optimal class that maximizes $\eta_c(x)$, and we let $\widehat{c}_i$ denote the predicted class that maximizes the fitted scores $\widehat{g}_{ic}$.

Corollary 7. *Assume the conditions of Corollary 6. Then*

$$\frac{1}{|I_{\mathrm{ev}}|} \sum_{i \in I_{\mathrm{ev}}} \{\eta_{c^*(X_i)}(X_i) - \eta_{\widehat{c}_i}(X_i)\} = O_{\mathbb{P}}\{A_p^Y + r_n^2 + M^{-1}\}^{1/2}.$$

Corollary 7 rigorously converts the class-probability approximation bound directly into a bound on conditional excess classification risk averaged over the evaluation covariates, notably achieving this without requiring any additional classification margin condition.

Proof of Corollary 7. For every $i \in I_{\mathrm{ev}}$, the definition of $\widehat{c}_i$ gives

$$0 \le \eta_{c^*(X_i)}(X_i) - \eta_{\widehat{c}_i}(X_i) \le 2\|\widehat{g}_i - \eta(X_i)\|_\infty \le 2\|\widehat{g}_i - \eta(X_i)\|_2.$$

Cauchy–Schwarz and Corollary 6 prove the assertion. This completes the proof. □

When we use a fixed number of data partitions to estimate the metric, averaging their respective fitted class-score vectors safely preserves the conclusion of Corollary 7. Because convexity ensures that the squared error of the average is bounded by the average of the squared errors, the preceding argument applies to each partition under the stated conditions.

B.2 How predictive risk controls spectral approximation

To establish the local-bandwidth result, we fix a support manifold and write the coordinates as $X = \phi(Z)$. Using a single prescribed response V, we define the conditional mean as $f(Z) = \mathbb{E}(V \mid Z)$. Let g_a be the metric induced by $\mathrm{diag}(a)^{1/2}\phi$ and p_a the density with respect to its volume measure. Write $\mathcal{B}_a f = -(2p_a)^{-1} \operatorname{div}_{g_a}(p_a \nabla_{g_a} f)$, the prescribed-view operator of Theorem 9. We define the roughness penalty $\mathcal{J}(a) = q_V^{-1}\|\mathcal{B}_a f\|^2_{L^2(P)}$ and the irreducible variance $\mathcal{V} = q_V^{-1}\mathbb{E}\|V - f(Z)\|_2^2$. We let $A_p(a)$ denote the projection error defined earlier, but evaluated specifically for the continuous operator $\mathcal{L}_a$ in (11). Because these two differential operators incorporate distinct density terms, the following result explicitly accounts for their difference.

Proposition 7. *Assume Condition 3 on this support and $f \in C^6(\mathcal{M})$ for the prescribed response view. Uniformly over $a \in \mathcal{A}_\delta$,*

$$A_p(a) \le \frac{C\mathcal{J}(a)}{\lambda_{p+2}(a)^2} \le \frac{C}{\lambda_{p+2}(a)^2}\left[h^{-4}\{\mathcal{R}_{\infty,0}(a,h) - \mathcal{V}\} + h^2\right].$$

If a_h minimizes $\mathcal{R}_{\infty,\rho_h}(a,h)$, then

$$A_p(a_h) \le \frac{C}{\lambda_{p+2}(a_h)^2}\left[\inf_{a \in \mathcal{A}_\delta}\left\{\mathcal{J}(a) + \frac{\rho_h}{2mh^4}\|a - \mathbf{1}\|_2^2\right\} + h^2\right].$$

Proposition 7 successfully bounds the response-mean variation that the finite representation fails to capture. By taking the unregularized population predictive loss, subtracting the irreducible variance, and dividing the remainder by h^4, we control the roughness penalty in this bound up to an $O(h^2)$ remainder. We emphasize that this assertion provides an upper bound on the approximation error; it does not claim that the chosen metric directly minimizes $A_p(a)$ itself. Furthermore, this analysis utilizes a shrinking prediction bandwidth, which behaves differently from the fixed scaled bandwidth analyzed in Theorem 8.

Proof of Proposition 7. For a scalar v, integration by parts gives

$$\langle v, \mathcal{B}_a v\rangle_{L^2(P)} = \frac{1}{2}\|\nabla_{g_a} v\|^2_{L^2(P)}.$$

The compact metric family has a uniform Poincaré constant. Applying its inequality to $v - \mathbb{E}v$ yields

$$\|\nabla_{g_a} v\|_{L^2(P)} \le C\|\mathcal{B}_a v\|_{L^2(P)}.$$

The definitions of the two operators give

$$\mathcal{L}_a v = \sigma_\varphi p_a^{-2/d}\left\{\mathcal{B}_a v + \frac{1}{d}\langle \nabla_{g_a}\log p_a, \nabla_{g_a} v\rangle_{g_a}\right\}.$$

The density bounds and the preceding gradient inequality imply $\|\mathcal{L}_a v\|_{L^2(P)} \le C\|\mathcal{B}_a v\|_{L^2(P)}$, uniformly in a. Let $\Pi_p(a)$ project onto constants and the first p nonconstant eigenfunctions. Parseval's identity gives

$$\|(I - \Pi_p(a))v\|^2_{L^2(P)} \le \lambda_{p+2}(a)^{-2}\|\mathcal{L}_a v\|^2_{L^2(P)}.$$

Apply this inequality to each component of f and sum. The uniform expansion in the proof of Theorem 9 gives

$$h^{-4}\{\mathcal{R}_{\infty,0}(a,h) - \mathcal{V}\} = \mathcal{J}(a) + O(h^2).$$

Finally, minimizing the penalized risk and dropping its nonnegative penalty at a_h gives

$$\mathcal{J}(a_h) \le \inf_{a\in\mathcal{A}_\delta}\left\{\mathcal{J}(a) + \frac{\rho_h}{2mh^4}\|a - \mathbf{1}\|_2^2\right\} + O(h^2).$$

These inequalities prove both assertions. This completes the proof. □

C IDENTIFICATION AND COMPUTATION IN THE CONTINUOUS RESPONSE MODEL

The model in Theorem 8 permits increasing observed dimension while the predictive state remains a fixed-dimensional manifold. Its population metric, support, spectral coordinates, and computational accuracy can all be derived from the stated distribution. The calculation below also permits dependence among the excluded coordinates within each sphere block.

C.1 AN EXPLICIT BOUND FOR SCALING DERIVATIVES

We bound the scaling derivatives uniformly over the feasible metric set. Let B_X and B_V bound observed and response coordinates. Fix a minimum block size $b_0 \ge 3$ and $h^2 = m\tau^2$. Define

$$k_{\min} = e^{-2B_X^2/\tau^2}, \qquad r_u = \frac{b_0 - 2}{k_{\min}(b_0 - 1) - 1}, \qquad c_b = \frac{b_0 - 1}{b_0 - 2} + \frac{1}{2},$$

and, when the denominators are positive, put

$$C_{\text{inv}} = \frac{c_b}{1 - c_b(r_u^2/k_{\min} - 1)}, \qquad C_{H,2} = \frac{4B_X^4 B_V^2(6 + 8C_{\text{inv}})}{\tau^4}.$$

For matrices, $\|A\|_\infty = \max_i \sum_j |A_{ij}|$. At the fixed bandwidth h, write $W_a = W(a, h; \mathbf{1})$ and $\ell_b^V(a,h) = \|(I_b - W_a)V\|_F^2/(bq_V)$ for the prescribed-response affinity and loss. These constants are deterministic functions of the coordinate bounds, bandwidth, and minimum block size. They give a quantitative version of the derivative control in Lemma 4.

Lemma 2. *Suppose $k_{\min}(b_0 - 1) > 1$ and $c_b(r_u^2/k_{\min} - 1) < 1$. For every $b \ge b_0$, every prescribed-response block, and every $a \ge 0$ with $\mathbf{1}^\top a \le m$,*

$$\|(I_b + W_a)^{-1}\|_\infty \le C_{\text{inv}}, \qquad \max_{j,\ell} |\partial_{a_j}\partial_{a_\ell}\ell_b^V(a,h)| \le C_{H,2}m^{-2}.$$

The same derivative bound holds for the population prediction risk without its ridge penalty.

Lemma 2 controls the entire feasible set. Thus its bound can justify restriction to selected coordinates before optimization. It is not an empirical Hessian evaluated only at a computed solution.

Proof of Lemma 2. Write $W_{ij} = u_i K_{ij} u_j$, $i \neq j$, with $k_{\min} \leq K_{ij} \leq 1$. Let $M = \max_i u_i$, $s_u = \min_i u_i$, and let T sum the remaining $b-2$ scaling factors. The row equations at these two coordinates imply

$$k_{\min} M(s_u + T) \leq 1 \leq s_u(M + T).$$

Because $T/s_u \geq b-2$,

$$\frac{M}{s_u} \leq \frac{b-2}{k_{\min}(b-1)-1} \leq r_u.$$

The row equations also bound W_{ij} between $k_{\min}\{r_u^2(b-1)\}^{-1}$ and $r_u^2\{k_{\min}(b-1)\}^{-1}$. For $P_b = (\mathbf{1}\mathbf{1}^\top - I_b)/(b-1)$,

$$\|W - P_b\|_\infty \leq r_u^2/k_{\min} - 1, \qquad \|(I_b + P_b)^{-1}\|_\infty = \frac{b-1}{b-2} + \frac{1}{2} \leq c_b.$$

The Neumann series proves the inverse bound.

For a parameter direction v, write $C_v = \partial_v \log K$ and $z_v = \partial_v \log u$. Then

$$\|C_v\|_{\max} \leq c_v := \frac{2B_X^2 \|v\|_1}{m\tau^2}, \qquad (I_b + W)z_v = -(W \circ C_v)\mathbf{1}.$$

Put $G_{v,ij} = C_{v,ij} + z_{v,i} + z_{v,j}$. The row equations give

$$\partial_v W = W \circ G_v, \qquad \partial_v W \mathbf{1} = 0, \qquad \|G_v\|_{\max} \leq (1 + 2C_{\text{inv}})c_v.$$

Under $\langle A, B\rangle_W = \sum_{i\neq j} W_{ij} A_{ij} B_{ij}$, additive row and column correction projects orthogonally onto matrices with zero W-weighted row sums. Hence

$$\|G_v\|_W \leq \|C_v\|_W \leq \sqrt{b}\, c_v, \qquad \left\| \frac{\partial_v \partial_w W}{W} \right\|_W \leq \|G_v G_w\|_W \leq \sqrt{b}(1 + 2C_{\text{inv}})c_v c_w.$$

The quotient is entrywise off the diagonal. Weighted Cauchy–Schwarz now gives

$$\|\partial_v W V\|_F \leq B_V \sqrt{bq_V}\, c_v, \qquad \|\partial_v \partial_w W V\|_F \leq B_V \sqrt{bq_V}(1 + 2C_{\text{inv}})c_v c_w.$$

Differentiate $\ell_b^V = \|(I-W)V\|_F^2/(bq_V)$ twice and use $\|(I-W)V\|_F \leq 2B_V\sqrt{bq_V}$. The absolute mixed derivative is at most $B_V^2(6 + 8C_{\text{inv}})c_v c_w$. Taking coordinate directions proves the stated bound.

For the population scaling, let $S_u = \int u\, dP$. The scaling equation gives $1 \leq S_u^2 \leq k_{\min}^{-1}$ and $\sqrt{k_{\min}} \leq u \leq k_{\min}^{-1}$. Its normalized kernel lies between $k_{\min}^2$ and $k_{\min}^{-2}$. The projection onto constants has inverse bound $3/2$ for $I + \Pi$. Because $r_u \geq k_{\min}^{-1}$, the same Neumann bound C_{inv} applies to $I + T$. Replacing sums by integrals in the preceding derivative calculation proves the population assertion. This completes the proof. □

C.2 A GLOBAL CERTIFICATE BEFORE RESTRICTED OPTIMIZATION

We use gradient contrasts to certify that every global minimizer lies on a selected face. Use $a = \delta\mathbf{1} + m(1-\delta)\theta$ and $\rho_m = \bar{\rho}/m$, with $\Delta_m = \{\theta \geq 0 : \mathbf{1}^\top\theta = 1\}$. Let $F(\theta) = \widehat{\mathcal{R}}_{\rho_m}(a, h)$ and $\theta^{\text{unif}} = m^{-1}\mathbf{1}$. For a nonempty proper coordinate set S, define

$$U(S) = \max_{\ell\in S,\, j\notin S} \{\partial_{\theta_\ell} F(\theta^{\text{unif}}) - \partial_{\theta_j} F(\theta^{\text{unif}})\} + (4C_{H,2} + \bar{\rho})(1-\delta)^2,$$

and $\kappa(S) = (\bar{\rho} - |S|C_{H,2})(1-\delta)^2$. For a feasible point on the face supported by S, let $r(\theta)$ be the distance of $-\nabla F(\theta)$ from the normal cone of that face, restricted to its coordinates. Let $\widetilde{r}(\theta)$ be the same normal-cone distance computed with the numerical gradient. These quantities concern the observed objective and are computed after the data are given.

Proposition 8. *Under Lemma 2, if $U(S) < 0$ and $\kappa(S) > 0$, every global minimizer is supported on S, and*

$$0 \leq F(\theta) - \min_{u\in\Delta_m} F(u) \leq \frac{r(\theta)^2}{2\kappa(S)} \qquad \textit{for } \operatorname{supp}(\theta) \subset S.$$

If objective and θ-gradient evaluations have certified errors ϵ_f and ϵ_g in absolute value and maximum norm, a valid objective interval has width at most

$$2\epsilon_f + \frac{\{\widetilde{r}(\theta) + \sqrt{|S|}\epsilon_g\}^2}{2\kappa(S)}.$$

Proposition 8 gives a global conclusion only after the exchange and curvature bounds succeed. The computed contrast adds twice the maximum-norm gradient evaluation error. The computation in Section I.6 records both inequalities and includes evaluation error in the final objective gap. This avoids an exponential search over excluded coordinates.

Proof of Proposition 8. The data part of the θ-Hessian has every entry bounded by $C_{H,2}(1-\delta)^2$. Because $\|\theta - \theta^{\mathrm{unif}}\|_1 \le 2$, its coordinate-gradient contrasts change by at most $4C_{H,2}(1-\delta)^2$. The ridge contrast is $\bar{\rho}(1-\delta)^2(\theta_\ell - \theta_j) \le \bar{\rho}(1-\delta)^2$. Thus $U(S) < 0$ makes movement of any positive mass from $j \notin S$ to $\ell \in S$ decrease F. A global minimizer cannot assign positive mass outside S.

For v supported on S, $\|v\|_1^2 \le |S|\|v\|_2^2$. Lemma 2 and the ridge Hessian give

$$v^\top \nabla^2 F(\theta) v \ge \kappa(S)\|v\|_2^2.$$

Let θ^* minimize F on the face and choose a face-normal vector z with $\|\nabla F(\theta) + z\|_2 = r(\theta)$. Strong convexity and $\langle z, \theta^* - \theta\rangle \le 0$ imply

$$F(\theta) - F(\theta^*) \le r(\theta)\|\theta - \theta^*\|_2 - \frac{\kappa(S)}{2}\|\theta - \theta^*\|_2^2 \le \frac{r(\theta)^2}{2\kappa(S)}.$$

The distance to a closed convex set is 1-Lipschitz. A gradient error of maximum norm ϵ_g changes this distance by at most $\sqrt{|S|}\epsilon_g$. Subtracting and adding the objective evaluation error proves the final interval bound. This completes the proof. □

Redistributing excluded mass across the selected coordinates permits larger ridge penalties. For a nonempty proper set S, the conclusions of Proposition 8 hold if $\kappa(S) > 0$ and

$$\frac{1}{|S|}\sum_{\ell \in S} \partial_{\theta_\ell} F(\theta^{\mathrm{unif}}) - \min_{j \notin S} \partial_{\theta_j} F(\theta^{\mathrm{unif}}) + \left(4C_{H,2} + \frac{\bar{\rho}}{|S|}\right)(1-\delta)^2 < 0. \tag{16}$$

To prove this assertion, put $t = \sum_{j \notin S} \theta_j > 0$, $v_\ell = t/|S|$ for $\ell \in S$, and $v_j = -\theta_j$ for $j \notin S$. The segment $\theta + uv$, $0 \le u \le 1$, is feasible. Because $\|v\|_1 = 2t$ and $\|\theta - \theta^{\mathrm{unif}}\|_1 \le 2$, the Hessian bound in Lemma 2 changes the data directional derivative by at most $4tC_{H,2}(1-\delta)^2$. The ridge contribution divided by t equals

$$\bar{\rho}(1-\delta)^2\Big\{\frac{1-t}{|S|} - \frac{1}{t}\sum_{j \notin S}\theta_j^2\Big\} \le \frac{\bar{\rho}}{|S|}(1-\delta)^2.$$

Inequality (16) makes the full directional derivative negative. Hence every minimizer is supported on S. The restricted curvature and normal-cone argument in the proof of Proposition 8 give the stated objective bounds. A maximum-norm error in the initial gradient contributes twice that error to (16).

Optimization can also be certified on the complete simplex. If $\bar{\rho} > mC_{H,2}$, the same Hessian bound gives global strong convexity with constant $(\bar{\rho} - mC_{H,2})(1-\delta)^2$. Let $N_{\Delta_m}(\theta)$ denote the normal cone of Δ_m at a feasible θ, and suppose $\|\widetilde{g} - \nabla F(\theta)\|_\infty \le \epsilon_g$. Then, for any $z \in N_{\Delta_m}(\theta)$,

$$F(\theta) - \min_{u \in \Delta_m} F(u) \le \frac{\{\|\widetilde{g} + z\|_2 + \sqrt{m}\,\epsilon_g\}^2}{2(\bar{\rho} - mC_{H,2})(1-\delta)^2}. \tag{17}$$

Indeed, the data Hessian has operator norm at most $mC_{H,2}(1-\delta)^2$. Add the ridge Hessian and apply the strong-convexity calculation in Proposition 8. The gradient error has Euclidean norm at most $\sqrt{m}\,\epsilon_g$. This proves (17) without a support assumption. Projection after a gradient step of length $\{\bar{\rho}(1-\delta)^2\}^{-1}$ contracts by at most $mC_{H,2}/\bar{\rho}$ in exact arithmetic: projection is nonexpansive, and the eigenvalues of the gradient-step derivative have absolute value at most this ratio.

Equation (17) concerns the exact optimization error. An interval containing both $F(\theta)$ and its minimum adds $2\epsilon_f$ to the gap bound when the objective evaluation error is at most ϵ_f. The objective interval can thus be wider than a certified optimization error. This distinction permits accurate metric optimization when the prediction kernel is close to constant.

C.3 Numerical error in the objective certificate

Scaling and differentiation errors enter the bound in Proposition 8. We give their calculation explicitly. Let $\widetilde{W}$ be computed from approximate log-scaling factors $\widetilde{z}$, put $r = \|\widetilde{W}\mathbf{1} - \mathbf{1}\|_\infty$, and let $M = \operatorname{diag}(\widetilde{W}\mathbf{1}) + \widetilde{W}$. An inverse bound $C_z \geq \|M^{-1}\|_\infty$ follows by comparison with $I + P_b$. Set $t = 2C_z r$. For $\|v\|_\infty \leq t$, the row-equation remainder is at most $2\|\widetilde{W}\mathbf{1}\|_\infty e^{2t}t^2$. Thus

$$4C_z\|\widetilde{W}\mathbf{1}\|_\infty e^{2t}t \leq 1 \quad \Longrightarrow \quad \|z - \widetilde{z}\|_\infty \leq t.$$

Indeed, the map obtained by subtracting the row equation multiplied by M^{-1} maps this ball into itself. A fixed point is a scaling root, and strict convexity gives uniqueness. If the computed log kernel has error at most e_K, use in this numerical-error calculation the value of C_{inv} obtained by replacing $k_{\min}$ with $k_{\min}e^{-2e_K}$ in its defining formulas, and verify the corresponding Neumann condition. A common scalar factor in a kernel does not change its bistochastic normalization. Every kernel on the log-linear path from the exact kernel to the rounded kernel can be rescaled to have entries in $[k_{\min}e^{-2e_K}, 1]$. The inverse bound is valid along this entire path. Integration of the scaling derivative bound gives

$$\|W - \widetilde{W}\|_\infty \leq e_W, \qquad e_W = \|\widetilde{W}\mathbf{1}\|_\infty(e^{2t} - 1) + e^{(1+2C_{\mathrm{inv}})e_K} - 1.$$

In particular, the absolute block-loss error is at most $B_V^2(4e_W + e_W^2)$ before arithmetic error in the squared loss is included.

For the adjoint derivative, let

$$H = -\frac{(V - WV)V^\top + V(V - WV)^\top}{bq_V}, \qquad z_H = (I + W)^{-1}(W \circ H)\mathbf{1}.$$

Write $\widetilde{H}$ and $\widetilde{z}_H$ for their computed versions and let r_H bound the residual in the computed adjoint system. With arithmetic errors added to the right sides, valid bounds are

$$e_H = \frac{2B_V^2 e_W}{b}, \qquad e_z = C_{\mathrm{inv}}\{e_W\|\widetilde{H}\|_{\max} + e_H + 2e_W\|\widetilde{z}_H\|_\infty + r_H\},$$

and

$$e_g = \frac{2B_X^2 b}{h^2}\left[e_W\{\|\widetilde{H}\|_{\max} + 2\|\widetilde{z}_H\|_\infty\} + e_H + 2e_z\right].$$

The last quantity bounds the maximum-norm error in the a-gradient of one block. The θ-gradient error multiplies it by $m(1 - \delta)$. To obtain the bound, subtract the two adjoint equations, use $\|M - (I + W)\|_\infty \leq 2e_W$, and then subtract $W \circ (H - z_H\mathbf{1}^\top - \mathbf{1}z_H^\top)$. Summing the absolute entries of that difference and using $|\partial_{a_j} \log K_{i\ell}| \leq 2B_X^2/h^2$ gives e_g.

The implementation enlarges row, kernel, adjoint, and summation residuals by floating-point error bounds before using these inequalities. Blockwise errors are averaged with the same weights as the empirical loss. The sign of the exchange certificate and the positivity of the curvature bound are checked with these errors included. These checks are separate from an optimizer's stopping message.

C.4 Score approximation with increasing observed dimension

For the prescribed response, write ℓ_b^V for one block loss without its ridge penalty and $\mathcal{R}_{\infty,0}$ for its population counterpart. The next bound concerns scores at specified population parameters. Only a finite set of score coordinates is needed; no covering of an m-dimensional parameter set is required.

Lemma 3. *Under Lemma 2, let a be deterministic and let $J_0 \subset \{1, \ldots, m\}$ have size d_0. For R independent blocks of size b,*

$$\max_{j \in J_0}\left|m\partial_{a_j}\{\widehat{\mathcal{R}}_{\rho_m}(a, h) - \mathcal{R}_{\infty,\rho_m}(a, h)\}\right| = O_{\mathbb{P}}\left\{\sqrt{\frac{\log(2d_0)}{R}} + \sqrt{\frac{\log(2bd_0)}{b}}\right\}.$$

The constants do not depend on m, b, or R.

Lemma 3 separates variation across independent blocks from their approximation to a population prediction operator. Screening uses $d_0 = m$. Estimation on the recovered support uses fixed $d_0 = s$, which removes the logarithm of the observed dimension from its rate.

Proof of Lemma 3. Use the rescaled coordinate derivative $m\partial_{a_j}$. The log-kernel derivative, the population log-scaling derivative, and the prediction derivative are uniformly bounded by the inverse bound in Lemma 2. Let $u_a(x)$ denote population scaling. For one block, set

$$\overline{W}_{i\ell} = \frac{u_a(X_i)K_a(X_i, X_\ell)u_a(X_\ell)}{b-1}\mathbf{1}_{\{i\neq\ell\}}.$$

Conditional on X_i, each row sum averages $b-1$ independent bounded variables, with mean one and derivative mean zero. The corresponding means for $\sum_{\ell\neq i}\overline{W}_{i\ell}V_\ell$ and its derivative are $T_a f_0(X_i)$ and its derivative. Hoeffding's inequality over rows, response coordinates, and $j \in J_0$ gives, with probability at least $1 - b^{-3}$,

$$\max_i |\overline{W}\mathbf{1} - \mathbf{1}|_i + \max_{1\le i\le b}\max_{j\in J_0}|m\partial_{a_j}(\overline{W}\mathbf{1})|_i \le C\epsilon_b, \qquad \epsilon_b = \sqrt{\frac{\log(2bd_0)}{b}},$$

and the two prediction averages differ from their conditional means by at most $C\epsilon_b$. The response dimension is fixed.

Put $P_b = (\mathbf{1}\mathbf{1}^\top - I_b)/(b-1)$ and $M_b = \mathrm{diag}(\overline{W}\mathbf{1}) + \overline{W}$. The population scaling bounds give $\|\overline{W} - P_b\|_\infty \le k_{\min}^{-2} - 1 \le r_u^2/k_{\min} - 1$. The row residual gives $\|M_b - (I_b + P_b)\|_\infty \le r_u^2/k_{\min} - 1 + C\epsilon_b$. The Neumann bound in Lemma 2 yields $\|M_b^{-1}\|_\infty \le 2C_{\mathrm{inv}}$ for sufficiently small ϵ_b. Multiplying the row equation by this fixed inverse and subtracting it from the log-scaling vector maps the ball of radius $4C_{\mathrm{inv}}C\epsilon_b$ about $\log u_a(X_i) - \frac{1}{2}\log(b-1)$ into itself. The quadratic remainder is $O(\epsilon_b^2)$ on this ball. A fixed point solves the row equation, and strict convexity gives the unique scaling root, at distance at most $C\epsilon_b$.

Differentiate the row equation and subtract the sampled population derivative. The inverse bound for $I_b + W$, the row-derivative residual, and the root bound give derivative error at most $C\epsilon_b$. Writing

$$g_j(X, V) = -\frac{2}{q_V}\{V - T_a f_0(X)\}^\top m\partial_{a_j} T_a f_0(X),$$

we obtain

$$\max_{j\in J_0}\left|m\partial_{a_j}\ell_b^V - \frac{1}{b}\sum_{i=1}^b g_j(X_i, V_i)\right| \le C\epsilon_b.$$

Bounded scores give the same bound for large ϵ_b and control the complement of probability b^{-3}. Taking expectations gives bias at most $C\epsilon_b$. Hoeffding's inequality for the R independent bounded scores over d_0 coordinates adds $O_{\mathbb{P}}\{\sqrt{\log(2d_0)/R}\}$. The deterministic ridge derivatives cancel. This completes the proof. □

C.5 Prediction of a different response from the learned coordinates

A response view can identify coordinates that remain useful for predicting another measurement. The next result considers this question in the sphere model. The response used for metric estimation is unchanged, but the later prediction target can vary nonlinearly over the same sphere. We also allow the coordinate standardization used in the numerical study.

Use the distribution in Theorem 8 with $s = 3$. For the N metric observations, let $\widehat{d}_j$ be their sample standard deviation in coordinate j, truncated below at $1/2$. Divide every observed coordinate by $\widehat{d}_j$ before metric estimation and graph construction. Use $\rho_m = c/(3\tau^2 m)$ and the certificate in (16), with screening cutoff $(3\tau^2)^{-1}$. Let H_L contain the real spherical harmonics of degrees $1, \dots, L$, orthonormal under the uniform probability distribution on the sphere of radius $\sqrt{3}$. Its dimension is $q_L = L(L+2)$. The additional response is $Y = H_L(X^{(1)}) + \xi$, where $\mathbb{E}\|\xi\|_2^2 < \infty$, $\mathbb{E}(\xi \mid X, V) = 0$, and the observations are independent copies of (X, V, ξ).

Observe Y on M of the metric observations, with indices fixed independently of their values. Use an intercept and the first $p = J(J+2)$ learned coordinates, where $J \ge L$. Center and standardize

each regression coordinate on these M observations. Fit ridge regression with penalty $\lambda_M \geq 0$ multiplying the squared coefficient norm in the unnormalized least-squares criterion. Denote its predictions by $\widehat{g}_i$. Clipping to the coordinate bounds of H_L is permitted. Use $e_{R,b}$ as defined before Theorem 8.

Corollary 8. *Assume the growth conditions of Theorem 8, $M \asymp N$, and fixed $L \leq J$. If $0 < c < 6/(1-\delta)$, $\lambda_M/M \to 0$, and τ is fixed and sufficiently large, then*

$$\mathbb{P}(\widehat{S} = \{1,2,3\}) \to 1, \qquad \frac{1}{q_L |I_{\mathrm{ev}}|} \sum_{i \in I_{\mathrm{ev}}} \|\widehat{g}_i - H_L(X_i^{(1)})\|_2^2 = O_{\mathbb{P}}\left\{ \frac{k_n}{n} + e_{R,b}^2 + M^{-1} + \frac{\lambda_M^2}{M^2} \right\}.$$

Corollary 8 separates identification of the predictive coordinates from estimation of a later response. The response view identifies the first sphere, and its spectral coordinates represent each fixed collection of spherical harmonics without population approximation error. The assertion permits a smaller fixed fraction of labeled observations for the later response. It does not cover $M/N \to 0$ or compare prediction risk with other supervised methods.

Proof of Corollary 8. Put $S_0 = \{1,2,3\}$ and $\bar{\rho} = c/(3\tau^2)$. For sufficiently large τ, Lemma 2 gives $C_{H,2} = O(\tau^{-4})$ and

$$\bar{\rho} > 3C_{H,2}, \qquad \frac{2}{3\tau^2} - C_{H,2} - \left(4C_{H,2} + \frac{\bar{\rho}}{3}\right)(1-\delta) > 0.$$

The second inequality is possible because $c < 6/(1-\delta)$. The population derivative calculation in the proof of Theorem 8, followed by (16), confines every population minimizer to S_0. Restricted strong convexity and permutation invariance give $\theta_{0j} = 1/3$ on S_0 and zero elsewhere.

Bernstein's inequality for the bounded coordinates and their squares gives

$$\max_{j \leq m} |\widehat{d}_j^2 - 1| = O_{\mathbb{P}}\left\{ \sqrt{\frac{\log(2m)}{N}} \right\} = o_{\mathbb{P}}(R^{-1/2}).$$

The last relation follows from $N = Rb$ and $\log(2m)/b \to 0$. The lower truncation is inactive with probability tending to one. Write $Q(a)$ for the observed prediction loss without its ridge penalty and $Q^\circ(a)$ for this loss after standardization. Pairwise distances give the exact identity $Q^\circ(a) = Q(a_1/\widehat{d}_1^2, \ldots, a_m/\widehat{d}_m^2)$. Lemma 2 applies on the enlarged set $\sum_j a_j \leq 4m$ by replacing a by $a/4$ and τ by $\tau/2$. Its proof also bounds each first derivative by C/m. On $\max_j |\widehat{d}_j^2 - 1| \leq 1/2$, the chain rule and integration of the second-derivative bound imply

$$\sup_{a \in \mathcal{A}_\delta} m \|\nabla Q^\circ(a) - \nabla Q(a)\|_\infty \leq C \max_{j \leq m} |\widehat{d}_j^2 - 1| = o_{\mathbb{P}}(R^{-1/2}).$$

The constant is independent of m, and the ridge is unchanged. The proof of Theorem 8 then gives certification, support S_0, and $\|\widehat{\theta} - \theta_0\|_2 = O_{\mathbb{P}}(e_{R,b})$.

Standardization gives the original-observation graph with weights $\widehat{a}_j^{\mathrm{rep}}/\widehat{d}_j^2$. Normalize them to sum to three; common scaling preserves all distance-to-radius ratios. The preceding bound and $\max_{j \in S_0} |\widehat{d}_j^2 - 1| = O_{\mathbb{P}}(N^{-1/2})$ give distance $O_{\mathbb{P}}(e_{R,b})$ from the three unit weights.

The degree-ℓ harmonic polynomials in three variables have dimension $\binom{\ell+2}{2} - \binom{\ell}{2} = 2\ell + 1$. Degrees up to J form a complete eigenspace with $p = J(J+2)$ nonconstant coordinates and a positive external eigengap. Theorem 19 gives Frobenius error $O_{\mathbb{P}}(\sqrt{n}\{(k_n/n)^{1/2} + e_{R,b}\})$ after rotation and multiplication of the learned coordinates by $\sqrt{n}$. Because $M \asymp N \asymp n$, the same bound applies on the regression sample, whose population-coordinate Gram matrix converges to the identity. Bounded spherical harmonics and $\mathbb{E}\|\xi\|_2^2 < \infty$ give a response score of order $M^{-1/2}$ by Chebyshev's inequality. Markov's inequality also gives $M^{-1/2}\|\xi_{\mathrm{lab}}\|_F = O_{\mathbb{P}}(1)$. Centering and standardization give an invertible coordinate transformation with bounded norm and bounded inverse with probability tending to one.

Apply Proposition 6 to the first L degrees, with zero approximation error. The bounded transformed penalty and fixed population coefficients give ridge error λ_M/M. The coefficient and evaluation-coordinate errors have order $M^{-1/2} + (k_n/n)^{1/2} + e_{R,b} + \lambda_M/M$. Squaring proves the assertion. Coordinatewise clipping cannot increase squared error relative to the bounded conditional mean. This completes the proof. □

C.6 A COMPARISON WHEN THE RESPONSE VIEW IS UNAVAILABLE

In the sphere model, the observed blocks have the same distribution. A procedure that treats them symmetrically and does not use the response cannot identify which block determines the target. This restriction applies to the comparison below, but not to response-informed PLS, kernel regression, or neural regression.

Let $\mathcal{E}$ be a possibly random p-dimensional subspace of centered functions in $L^2(P_X)$. Its distribution is equivariant under permutations of the K observed blocks, and its construction uses no responses. Denote the orthogonal projection by $\Pi_{\mathcal{E}}$ and let $c \in \mathbb{R}^s$ be an intercept. The $m = Ks$ observed coordinate functions are orthonormal. Bessel's inequality gives

$$\sum_{r=1}^{K}\sum_{j=1}^{s} \|\Pi_{\mathcal{E}} X_j^{(r)}\|^2_{L^2(P_X)} \le p.$$

Exchangeability makes the expected contribution of each block equal. Thus the expected population risk of the best linear prediction from this representation satisfies

$$\mathbb{E} \inf_{c,\, g_1,\dots,g_s \in \mathcal{E}} \frac{1}{s}\mathbb{E}\Big\{ \sum_{j=1}^{s} (V_j - c_j - g_j(X))^2 \;\mid\; \mathcal{E} \Big\} \ge \sigma^2 + 1 - \frac{p}{Ks}. \tag{18}$$

The inner expectation uses a fresh observation independent of the fitted subspace. For the population PALE representation in Theorem 8, $p = s$ and its coordinate span contains $\mathbb{E}(V \mid X)$. Its corresponding risk is exactly σ^2. This difference concerns information provided by the response view; it is not a lower bound for methods given that view.

D FURTHER RESULTS ON GRAPH DEGENERACY

In this section, we explicitly relate the minimum eigenmap loss to the graph eigenvalues, and we then analytically construct a specific Gaussian affinity family for which this loss tends to zero.

Let $x_1, \dots, x_n \in \mathbb{R}^m$ denote standardized observations, and let $W \in \mathbb{R}^{n\times n}$ define a symmetric nonnegative affinity matrix. We write its degree matrix as $D = \operatorname{diag}(W\mathbf{1})$ and define its unnormalized Laplacian as $L = D - W$. For any fixed connected graph, computing a p-dimensional Laplacian Eigenmap corresponds to solving

$$\min_{Y \in \mathbb{R}^{n\times p}} \operatorname{tr}(Y^\top L Y) \quad \text{subject to} \quad Y^\top D Y = I_p, \qquad Y^\top D\mathbf{1} = 0. \tag{19}$$

These explicit constraints strictly fix the overall scale and forcefully exclude the trivial constant eigenvector. An arbitrary orthogonal rotation of a minimizing basis gives another minimizer. One solution consists of generalized eigenvectors associated directly with the p smallest strictly positive generalized eigenvalues.

If the affinity W is exactly bistochastic, then the degree matrix simplifies to $D = I_n$ and the Laplacian becomes $L = I_n - W$. We denote the ordered eigenvalues of this Laplacian by $0 = \lambda_1(L) \le \lambda_2(L) \le \cdots \le \lambda_n(L)$. Under these conditions, the standard variational characterization directly yields the profile loss specified in the following proposition.

Proposition 9. *Let $1 \le p \le n-1$. If $W = W^\top \ge 0$, $W\mathbf{1} = \mathbf{1}$, and the graph is connected, then the minimum in* (19) *equals*

$$\Phi_p(W) = \sum_{r=2}^{p+1} \lambda_r(I_n - W).$$

Every minimizer spans the eigenvectors associated with these eigenvalues.

Every valid minimizer spans a subspace generated by eigenvectors associated with these eigenvalues. This subspace is unique when the selected eigenvalues are separated from the remaining spectrum.

However, Proposition 9 takes on a fundamentally different meaning when the affinity W itself must be estimated. Directly minimizing $\Phi_p(W)$ simply encourages p additional graph eigenvalues to

collapse toward zero. These small eigenvalues describe weak connections between groups and do not by themselves establish the accuracy of the local geometry. Theorem 1 provides an explicit construction with this limit.

Importantly, Theorem 1 also explicitly permits row entropy lower bounds that fall strictly below $\log(s-1)$. Enforcing lower bounds close to $\log(n-1)$ would forcefully require nearly uniform rows, which reduces the locality of the representation.

This loss also tends to zero inside a standard Gaussian metric family when a single observed coordinate is constant within $p+1$ equally sized groups of at least three observations and has distinct values across groups. The following proposition makes this precise parametric scope fully explicit.

Proposition 10. *Let $n = (p+1)s$ and suppose a scalar feature z_i is constant within each of $p+1$ groups of size $s \geq 3$ and has distinct values across groups. For $i \neq j$, let*

$$K_{ij}(t) = \exp\{-t(z_i - z_j)^2\}, \qquad t > 0,$$

and let $W(t)$ be its symmetric bistochastic scaling with a zero diagonal. As $t \to \infty$,

$$\Phi_p\{W(t)\} \longrightarrow 0 \quad \textit{and} \quad -\sum_{j \neq i} W_{ij}(t) \log W_{ij}(t) \longrightarrow \log(s-1).$$

Proposition 10 formally establishes this structural failure entirely within a simple one-parameter Gaussian family. While imposing compact metric domains and prediction bandwidths bounded away from zero prevents this limit for fixed observations, these rigid constraints alone cannot identify which specific affinity actually represents reproducible predictive information.

Remark 1. Consider the bilinear score $A_{ij}(M) = \exp(x_i^\top M x_j)$ evaluated with a strictly deleted diagonal. Setting the metric $M = 0$ generates the perfectly uniform complete graph. Its corresponding bistochastic affinity assigns exactly $(n-1)^{-1}$ to every off-diagonal entry, which yields the profile loss $\Phi_p = pn/(n-1)$. Consequently, $M = 0$ explicitly does not produce the zero-loss graph. Instead, the true small-loss sequence corresponds to a family of affinities forming nearly separated blocks. Therefore, a penalty that merely keeps M bounded strictly away from zero fundamentally addresses the wrong failure mode.

Corollary 22 formally shows that any degree barrier remains strictly constant over all valid bistochastic graphs. While operations like Sparsemax (Martins & Astudillo, 2016) and nearest-neighbor truncation correctly produce exact zeros, neither operation mathematically excludes these separated blocks. Applying a rigid connectivity constraint such as $\lambda_2(I_n - W) \geq c > 0$ successfully excludes this pathological failure sequence, but it can also exclude genuine narrow connections inside the true manifold. Similarly, introducing a log pseudo-determinant barrier correctly penalizes small nonzero eigenvalues, yet it fundamentally cannot identify which connected affinity is actually supported by the observed features. Instead, predicting measurements explicitly excluded from the neighbor-selection process directly supplies a statistical target. Under the identification, geometric, and tuning conditions of Corollary 1, the learned spectral coordinates converge, up to rotation, to those of the selected population predictive metric.

E Further properties of the method

The specific metric parameterization, the randomized mask distribution, and the induced symmetry properties jointly determine the internal prediction tasks and the final representation. The theoretical results below establish their concrete finite-sample properties.

The maximum possible weight and the exact simplex parameterization defining the feasible set $\mathcal{A}_\delta$ are given by

$$A_m = \delta + m(1-\delta), \qquad a = \delta \mathbf{1} + m(1-\delta)\theta, \qquad \theta_j \geq 0, \quad \mathbf{1}^\top \theta = 1.$$

Because of the trace constraint, the active metric weights can attain values of order m. Under Condition 1 and $\pi \geq \pi_- > 0$, the bandwidth scaling $h^2 = m\tau^2$ gives an $O(1)$ contribution to the kernel exponent from each such coordinate. Each boundary coordinate has the fixed weight δ and an $O(m^{-1})$ contribution. The total contribution of $O(m)$ boundary coordinates need not tend to zero.

A fixed-count mask gives an exact balance between observed and held-out coordinates. Let a mask select k coordinates uniformly without replacement and put $\pi = k/m$. The following proposition gives its distance variance and feature-pair coverage.

Proposition 11. *For $z_\ell = a_\ell(x_{i\ell} - x_{j\ell})^2$ and $\bar{z} = m^{-1}\sum_{\ell=1}^m z_\ell$,*

$$\mathbb{E}_\xi d^2_{a,\xi}(x_i, x_j) = \sum_{\ell=1}^m z_\ell, \qquad \mathrm{var}_\xi\{d^2_{a,\xi}(x_i, x_j)\} = \frac{m^2(1-k/m)}{k(m-1)} \sum_{\ell=1}^m (z_\ell - \bar{z})^2.$$

For two distinct features r and s,

$$\mathbb{P}(\xi_r = 1, \xi_s = 0) = \frac{k(m-k)}{m(m-1)}.$$

With J independent masks, the probability that some ordered feature pair is never separated is at most $m(m-1)\exp\left\{-\frac{Jk(m-k)}{m(m-1)}\right\}$.

Proposition 11 gives two reasons to use k near $m/2$. This choice maximizes the chance that one feature constructs the graph while another feature assesses it, and it leaves enough coordinates on both sides of each prediction task. A larger k reduces distance variation but leaves fewer prediction targets. We use balanced masks unless the scientific views prescribe a different split.

The coverage bound also gives a direct mask count. To separate every ordered feature pair with probability at least $1 - \alpha$, it is sufficient to take

$$J \geq \frac{m(m-1)}{k(m-k)} \log\left\{\frac{m(m-1)}{\alpha}\right\}.$$

For a balanced split, this order is logarithmic in m. More masks may still reduce the Monte Carlo term in Theorem 11, but pair coverage alone does not require a mask count proportional to the number of features.

The unrestricted fixed-count distribution remains the default when no scientific grouping is available. Its population criterion is the mask average in (10), and Proposition 25 gives identification conditions for this aggregate risk. Corollary 10 verifies these conditions in an independent repeated-pair model. A paired balanced design gives a second primitive verification when repeated measurements are known. For an even number of remaining coordinates, it selects one member of every repeated-measurement pair with an independent fair orientation and selects half of the remaining coordinates. Every coordinate then has inclusion probability $1/2$, and each mask observes exactly half of all features. Proposition 26 shows that this random design places independent nuisance coordinates at the lower boundary, while Theorem 10 gives its local geometric target.

The matrix in (3) has positive off-diagonal entries. This support gives a unique symmetric scaling and an invertible derivative equation.

Proposition 12. *Suppose $n \geq 3$ and $K = K^\top$ has zero diagonal and positive off-diagonal entries. There is a unique $u \in \mathbb{R}^n_+$ such that $\mathrm{diag}(u)K\,\mathrm{diag}(u)$ is bistochastic. The resulting matrix W is irreducible and aperiodic. In particular, $I_n + W$ is positive definite.*

Proposition 12 gives a well-defined undirected affinity for every feasible metric and bandwidth. Bistochasticity normalizes neighborhood averages for prediction. The penalized prediction criterion selects the metric.

A prespecified response view can be used when scientific measurements distinguish the coordinates that define distance from those that assess it. In this case, X_{ξ^c} in (5) is replaced by an $n \times q_V$ response matrix V, $|\xi^c|$ is replaced by q_V, and none of the columns of V enters (2). The metric dimension remains m. A deterministic full observed view takes $\xi = \mathbf{1}$ and $\pi = 1$; the positive response dimension replaces the requirement that a metric coordinate be omitted. Every subsequent occurrence of X_{ξ^c}, $|\xi^c|$, and m_ξ is read as V, q_V, and $\mathbb{E}(V \mid X_\xi)$ under this convention. All scaling, optimization, and representation steps are unchanged. The sparse increasing-dimensional result uses this version with fixed q_V.

The prediction bandwidth is selected by minimizing the same criterion over $\mathcal{H}$ for a fixed ridge value. When bandwidth and ridge are both compared, the candidates are fitted on training blocks and

selected by the independent feature-prediction loss in Supplementary Section O. Repeated masks on the same block reduce the approximation error for the mask expectation and do not increase the number of independent observations.

The block loss deletes every self-connection but uses all rows in the symmetric scaling. The block size is allowed to increase in the population-target result. This regime connects the finite graph predictor in (5) to a population bistochastic smoother and removes the fixed-block approximation gap. Supplementary Section O gives an observation-level validation criterion for comparing a finite collection of fitted candidates.

We distinguish two estimands before tuning. The stable prediction mode uses $h = \sqrt{m}\,\tau$ on a fixed grid and targets the exact finite-bandwidth risk in (10). The geometry mode uses a shrinking candidate set whose largest bandwidth tends to zero and whose ridge ratios ρ/h^4 approach a common $\rho_\star$. Corollary 24 shows that validation within this set achieves the geometric rate and attains the oracle validation risk. The lower bound δ is fixed before fitting because it changes the parameter space and the population target. Section U.2 compares three lower bounds in a nonlinear response model. The representation threshold is chosen at an order larger than the simultaneous metric error and smaller than the active-weight separation.

The complete estimator is computed in five steps.

Step 1: Standardize the features using prespecified scales or a pilot sample.

Step 2: Partition the metric sample into disjoint blocks and generate the prespecified masks.

Step 3: For each $h \in \mathcal{H}$, minimize (6) over $\mathcal{A}_\delta$ and record an optimization certificate η_R.

Step 4: Apply (8) and construct the symmetric k_n-nearest-neighbor affinity on the representation observations.

Step 5: Solve (9) once and use the out-of-sample formula in Supplementary Section P for observations outside the representation sample.

The labels used in a later scientific analysis do not enter any step. Feature scales, mask seeds, the prediction-bandwidth set, t_N, the neighborhood order, and the representation dimension are fixed before evaluation. An independent representation sample gives a direct conditional analysis. The simultaneous spectral result in Theorem 19 also covers a full-data representation computed from the observations used to estimate the metric.

The representation dimension may be fixed by the scientific objective. When no dimension is specified, Supplementary Section Q gives an eigengap rule and proves its consistency under a unique oracle gap. This rule selects a stable spectral boundary; it does not assert that every data set has one intrinsic dimension.

The construction depends on pairwise distances and averages of prediction errors. Under a row permutation, each observation has the same block assignment and the mask attached to that block. Pilot and evaluation subsets are relabeled in the same way. The following proposition gives the resulting permutation identities. A unique selected eigenspace gives the corresponding identity for every fitted basis.

Proposition 13. *Let P and S be permutation matrices of orders n and m. For $\widetilde{X} = PXS$, $\widetilde{a} = S^\top a$, and $\widetilde{\xi} = S^\top \xi$,*

$$\widetilde{W} = PWP^\top, \qquad \ell_n(\widetilde{a}, h; \widetilde{\xi}) = \ell_n(a, h; \xi).$$

If $(\widehat{a}, \widehat{h})$ satisfies (7) for X, then $(S^\top \widehat{a}, \widehat{h})$ satisfies the same inequality for $\widetilde{X}$, the relabeled blocks, and the permuted masks. For these corresponding metrics, the minimizer sets in (9) are mapped by $Y \mapsto PY$. If the selected spectral subspace is unique, every pair of fitted orthonormal bases satisfies $\widehat{\widetilde{Y}} = P\widehat{Y}O$ for an orthogonal matrix O.

Proposition 13 gives invariance under consistent relabeling of observations and features. A new partition of the observations can change the objective. The thresholded support and normalized representation metric in (8) have the same feature-permutation equivariance. The diagonal parameterization and coordinate masks are not invariant to a general rotation of the feature space. The full symmetric metric in Supplementary Section M permits cross-coordinate distance terms within each observed view, but the definition of the views remains basis-dependent.

Remark 2. The predictive criterion uses repeated information across features. If a scientifically relevant coordinate is independent of every other observed coordinate, no unlabeled feature-prediction loss can identify its relevance. Natural feature groups can replace random masks when views are known. For example, one assay can define an affinity that predicts another assay measured on the same subjects. Random masks are appropriate when information is distributed across many pixels, wavelengths, genes, or related measurements.

Remark 2 states the identification boundary of the method. It also separates the role of a mask from arbitrary data augmentation. A mask is useful only when the held-out coordinates contain information about the same latent state.

F Optimization and exact differentiation

The predictive objective does not require an eigendecomposition during metric estimation. Its main calculation is the derivative of the symmetric scaling. This derivative can be obtained from one positive definite linear system per mask.

Write $z = \log u$ and $\eta_{ij} = \log K_{ij}$ for $i \neq j$. For a scalar perturbation, let $\dot{\eta}$ denote the derivative of the log kernel and let $\dot{z}$ denote the derivative of the log scaling. Differentiating (4) gives

$$(I_n + W)\dot{z} = -r, \qquad r_i = \sum_{j=1}^{n} W_{ij}\dot{\eta}_{ij}. \tag{20}$$

Proposition 12 makes the coefficient matrix positive definite. Once (20) is solved,

$$\dot{W}_{ij} = W_{ij}(\dot{z}_i + \dot{z}_j + \dot{\eta}_{ij}). \tag{21}$$

Equations (20) and (21) account for the dependence of every scaling factor on every metric parameter. Treating the scaling factors as fixed gives an incorrect gradient.

The next proposition gives a form that computes all feature derivatives without solving one system per feature. Let $Z = X_{\xi^c}$, $E = Z - WZ$, and

$$G = -\frac{EZ^\top + ZE^\top}{n|\xi^c|}.$$

Define $c = (G \circ W)\mathbf{1}$ and solve $(I_n + W)b = c$.

For the loss in (5), put

$$C = W \circ \{G - b\mathbf{1}^\top - \mathbf{1}b^\top\}.$$

Proposition 14. *For $\ell = 1, \dots, m$,*

$$\frac{\partial \ell_n(a, h; \xi)}{\partial a_\ell} = -\frac{\xi_\ell}{2\pi h^2} \sum_{i,j=1}^{n} C_{ij}(x_{i\ell} - x_{j\ell})^2.$$

The derivative with respect to $\log h$ *is*

$$\frac{\partial \ell_n(a, h; \xi)}{\partial \log h} = \frac{1}{h^2} \sum_{i,j=1}^{n} C_{ij} d^2_{a,\xi}(x_i, x_j).$$

Proposition 14 reduces the derivative cost to the construction of W, matrix products with the held-out features, and one solve with $I_n + W$. The main implementation selects h from a grid and uses only the first derivative in the proposition.

When m is large and the predictive metric is sparse at the lower boundary, one gradient evaluation gives a candidate set for a lower-dimensional certified optimization. Let $\widehat{\mathcal{R}}_0$ denote (6) without the ridge term and define

$$\widehat{d}_j(h) = -m\frac{\partial}{\partial a_j}\widehat{\mathcal{R}}_0(\mathbf{1}, h), \qquad d_j(h) = -m\frac{\partial}{\partial a_j}\mathcal{R}_{\infty,0}(\mathbf{1}, h).$$

For a prespecified $\lambda_R > 0$, put

$$\widehat{S}_{\mathrm{scr}}(h) = \{j : \widehat{d}_j(h) > \lambda_R\}.$$

Let $\widehat{S}_{\mathrm{cand}}(h) = \widehat{S}_{\mathrm{scr}}(h)$ when its restricted polytope is nonempty. If it is empty, add coordinates in decreasing score order until it becomes nonempty and call the augmented set $\widehat{S}_{\mathrm{cand}}(h)$. The screening calculation uses the same blocks as the final criterion and requires no additional fit.

The score screen proposes a restricted problem. A second calculation certifies that the restriction does not change the empirical global minimum. For a candidate set S whose restricted polytope is nonempty, define, at a fixed h,

$$\widehat{\Gamma}_j(S,h) = \sup_{\substack{a\in\mathcal{A}_\delta \\ a_j>\delta}} \min_{\substack{\ell\in S \\ a_\ell<A_m}} \left\{ \frac{\partial\widehat{\mathcal{R}}_\rho(a,h)}{\partial a_\ell} - \frac{\partial\widehat{\mathcal{R}}_\rho(a,h)}{\partial a_j} \right\}, \qquad j \notin S.$$

The inner set is nonempty at every feasible point in the supremum. Interval bounds on the analytic gradient give a computable upper bound for $\widehat{\Gamma}_j(S,h)$ over the branch-and-bound cells. Let e_j denote the jth standard basis vector. An inexpensive sufficient check uses

$$\widehat{H}_{\ell j}(h) = \sup_{a\in\mathcal{A}_\delta} \|(\mathsf{e}_\ell - \mathsf{e}_j)^\top \nabla_a^2 \widehat{\mathcal{R}}_\rho(a,h)\|_2.$$

Let C_H be the constant for second coordinate derivatives of the unpenalized block loss in Supplementary Lemma 4. The constants for first and third derivatives are defined separately. The Hessian bound gives the deterministic enclosure

$$\widehat{H}_{\ell j}(h) \le H_m, \qquad H_m = 2C_H m^{-3/2} + \sqrt{2}\,\rho/m.$$

The mean value theorem implies $\widehat{\Gamma}_j(S,h) < 0$ whenever

$$\max_{\ell\in S}\left[\frac{\partial\widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_\ell} - \frac{\partial\widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_j} + \widehat{H}_{\ell j}(h)\,\mathrm{diam}(\mathcal{A}_\delta)\right] < 0.$$

A single interval-Hessian bound can verify this inequality. Cellwise gradient bounds give a less conservative certificate when the single-box check fails.

The following proposition controls the candidate set obtained from the scores at the Euclidean metric.

Proposition 15. *Suppose Condition 1 holds and $m\log(mb)/b \le c_0$. With probability at least $1-\alpha$,*

$$\max_{h\in\mathcal{H}}\max_{1\le j\le m} |\widehat{d}_j(h) - d_j(h)| \le C\left[\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}} + \sqrt{\frac{m\log(mb)}{b}}\right].$$

Let the right side be at most λ_R. If, for a fixed h and a set S_0,

$$\min_{j\in S_0} d_j(h) > 2\lambda_R, \qquad \max_{j\notin S_0} d_j(h) \le 0,$$

then $\widehat{S}_{\mathrm{scr}}(h) = S_0$.

Proposition 15 identifies the population support from one empirical gradient under a score margin. It does not by itself justify restricting a nonconvex empirical objective away from the Euclidean metric.

The next proposition gives that global justification. Let L_R denote a certified Lipschitz constant of $\widehat{\mathcal{R}}_\rho$ on $\mathcal{A}_\delta$, uniformly over $\mathcal{H}$. Its population exchange condition is verified by Propositions 29 and 30.

Proposition 16. *Let $S \subset \{1,\dots,m\}$ have size s and suppose its restricted polytope is nonempty. If*

$$\max_{j\notin S} \widehat{\Gamma}_j(S,h) < 0,$$

then every global minimizer at bandwidth h has $a_j = \delta$ for $j \notin S$. A regular subdivision gives a global objective certificate at most η for the original fixed-bandwidth problem after at most

$C_s \left(\frac{L_R \operatorname{diam}(\mathcal{A}_\delta)}{\eta} \right)^{s-1}$ *cell evaluations. Suppose, for some* $\chi > 0$*, every* $a \in \mathcal{A}_\delta$ *with* $a_j > \delta$ *and* $j \notin S$ *admits an* $\ell \in S$ *with* $a_\ell < A_m$ *such that*

$$\frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_j} \le -\chi.$$

If $\sup_{a \in \mathcal{A}_\delta} \|\nabla_a \widehat{\mathcal{R}}_\rho(a,h) - \nabla_a \mathcal{R}_{\infty,\rho}(a,h)\|_\infty < \chi/4$*, then* $\max_{j \notin S} \widehat{\Gamma}_j(S,h) \le -\chi/2$.

Proposition 16 moves the exponential term from the observed feature dimension to the candidate-set size only after the global exchange calculation succeeds. Because $\operatorname{diam}(\mathcal{A}_\delta) \le \sqrt{2} m(1-\delta)$, the restricted covering factor is polynomial in m when s is fixed. The uniform derivative argument used in Proposition 31 makes the last gradient error converge to zero for fixed m. When either certificate fails, the unscreened optimization remains the valid procedure.

The certification cost includes the iterations needed for symmetric scaling. Let $0 < \epsilon \le \min(1/2, \eta/8)$ bound the absolute error in each block loss and in each scaled score $m\partial_{a_j}\ell_b$. The proof below gives a deterministic calculation with cost at most $C \log(b/\epsilon)$ times $C_{\mathrm{ev}}(b,m) = C(b^3 + b^2 m)$ arithmetic operations per block. The constants depend on the measurement and prediction-scale bounds. The analytic bound H_m requires no Hessian evaluation. Numerical intervals must verify the sign of the one-box inequality. For a prescribed response, each evaluation bound has the additional factor $1 + q_V/m$, accounting for response matrix products. This factor is bounded when q_V is fixed. For the asymptotic screening assertion, take $\epsilon \to 0$.

Proposition 17. *Suppose numerical intervals verify the one-box inequality above with* H_m *substituted for every* $\widehat{H}_{\ell j}(h)$ *and for every* $h \in \mathcal{H}$*, and let* s *be the largest candidate-set size. For sufficiently small* $\eta > 0$*, a certificate with objective tolerance* η *requires at most* $C \log(b/\epsilon)$ *times*

$$O\{|\mathcal{H}| R(b^3 + b^2 m) + |\mathcal{H}| m^2\} + C_s |\mathcal{H}| R(b^3 + b^2 m) \left\{ \frac{L_R \operatorname{diam}(\mathcal{A}_\delta)}{\eta} \right\}^{s-1}$$

arithmetic operations. If, for some $\chi > 0$*,*

$$\max_{j \notin S} \max_{\ell \in S} \left[\frac{\partial \mathcal{R}_{\infty,\rho}(\mathbf{1},h)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho}(\mathbf{1},h)}{\partial a_j} + H_m \operatorname{diam}(\mathcal{A}_\delta) \right] \le -\chi,$$

then the one-box inequality holds with probability tending to one whenever the candidate set equals S *with probability tending to one and the uniform gradient error is* $o_{\mathbb{P}}(1)$.

Proposition 17 counts the full-dimensional certification calculation and the restricted global search. The exponential term depends on s only under the stated one-box margin. If this check fails and a cellwise certificate over $\mathcal{A}_\delta$ is used, its full-dimensional cell evaluations must be added to the cost, and no reduced-dimensional complexity claim is made.

The Euclidean Hessian enclosure above is uniform but loses a factor $m^{1/2}$. A coordinatewise enclosure is sharper when the active dimension is fixed. Put

$$H_m^{(1)} = 2C_H m^{-2} + \rho/m.$$

The coordinate derivative bounds in Supplementary Lemma 4 give

$$\max_{1 \le r \le m} \sup_{a \in \mathcal{A}_\delta} \left| \frac{\partial}{\partial a_r} \left(\frac{\partial \widehat{\mathcal{R}}_\rho}{\partial a_\ell} - \frac{\partial \widehat{\mathcal{R}}_\rho}{\partial a_j} \right) \right| \le H_m^{(1)}.$$

Because $\sup_{a \in \mathcal{A}_\delta} \|a - \mathbf{1}\|_1 \le 2m(1-\delta)$, this bound yields a dimension-adapted certificate.

Proposition 18. *Let* S *have size* s *and suppose its restricted polytope is nonempty. If*

$$\max_{j \notin S} \max_{\ell \in S} \left[\frac{\partial \widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_\ell} - \frac{\partial \widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_j} + 2m(1-\delta) H_m^{(1)} \right] < 0,$$

then $\max_{j \notin S} \widehat{\Gamma}_j(S,h) < 0$*. Suppose* $\rho_m = \bar{\rho}/m$ *and, uniformly over* $h \in \mathcal{H}$*,*

$$\min_{j \in S} d_j(h) \ge c_d, \qquad \max_{j \notin S} d_j(h) \le 0,$$

and

$$\max_{j\notin S}\max_{\ell\in S}\left[m\left\{\frac{\partial\mathcal{R}_{\infty,\rho_m}(\mathbf{1},h)}{\partial a_\ell}-\frac{\partial\mathcal{R}_{\infty,\rho_m}(\mathbf{1},h)}{\partial a_j}\right\}+(4C_H+2\bar{\rho})(1-\delta)\right]\le-\chi$$

for fixed $c_d,\chi>0$. If $\lambda_R\to 0$ and the bound in Proposition 15 is $o(\lambda_R)$, then, with probability tending to one, $\widehat{S}_{\mathrm{cand}}(h)=S$ and the preceding certificate holds for every $h\in\mathcal{H}$. The operation bound in Proposition 17 then holds with an exponential term of dimension $s-1$.

Proposition 18 gives a verifiable regime in which screening and global certification succeed before the restricted optimization begins. Its population margin is on the scale of the screening scores and remains nondegenerate when active weights have order m.

After the one-box check succeeds, interval evaluation of the tangent Hessian is repeated on the restricted θ-simplex. A positive lower bound certifies strong convexity and makes projected gradient descent a global algorithm for that bandwidth. Theorem 18 proves that both the one-box and restricted-curvature checks succeed with probability tending to one in the sparse prespecified-view regime.

For a fixed bandwidth, projected gradient descent updates

$$a^{(t+1)}=\Pi_{\mathcal{A}_\delta}\{a^{(t)}-\gamma_t\nabla\widehat{\mathcal{R}}(a^{(t)},h)\},$$

where $\widehat{\mathcal{R}}$ denotes the criterion in (6), $\gamma_t>0$ is the step size, and $\Pi_{\mathcal{A}_\delta}$ is Euclidean projection onto $\mathcal{A}_\delta$. The projection has coordinates

$$\{\Pi_{\mathcal{A}_\delta}(v)\}_j=\max(\delta,v_j-t),$$

where t is chosen to make the coordinates sum to m. The criterion uses the disjoint blocks fixed before optimization. Hence the objective evaluated by the algorithm is exactly the objective used in the statistical analysis.

The nonconvex criterion requires an explicit optimization error. We use adaptive branch and bound on the compact polytope $\mathcal{A}_\delta$. Each cell is the intersection of $\mathcal{A}_\delta$ and a coordinate box $[\ell,u]$, and a selected cell is divided at the midpoint of a widest coordinate interval. At an iteration, let $\mathfrak{P}_t$ be a finite partition into cells. For every nonempty cell C, choose a feasible point c_C, compute $r_C=\sup_{a\in C}\|a-c_C\|_2$, and evaluate the criterion at c_C. If U_t is the smallest value found and

$$B_t=\min_{C\in\mathfrak{P}_t,\,h\in\mathcal{H}}\{\widehat{\mathcal{R}}_\rho(c_C,h)-L_Rr_C\},$$

then U_t-B_t is a computable global objective certificate. Cells whose lower bound exceeds U_t are discarded, and a cell attaining B_t is bisected. Projected gradient steps from evaluated points can decrease U_t without changing the lower bound. For a second certificate, define the normal-cone residual

$$s_R(a,h)=\inf_{v\in N_{\mathcal{A}_\delta}(a)}\|\nabla_a\widehat{\mathcal{R}}_\rho(a,h)+v\|_2.$$

Let Q have orthonormal columns spanning $\mathbf{1}^\perp$ and put

$$M_R=\max_{h\in\mathcal{H}}\sup_{a\in\mathcal{A}_\delta}\left[-\lambda_{\min}\{Q^\top\nabla_a^2\widehat{\mathcal{R}}_0(a,h)Q\}\right]_+.$$

Proposition 19. *Suppose the coordinates used for distance and assessment are bounded. Every coordinate mask includes and omits at least one coordinate, or a prespecified response view has $q_V\ge 1$. For each $h\in\mathcal{H}$, the criterion in* (6) *has a Lipschitz gradient and is L_R-Lipschitz on $\mathcal{A}_\delta$. At every branch-and-bound iteration,*

$$0\le U_t-\inf_{a\in\mathcal{A}_\delta,\,h\in\mathcal{H}}\widehat{\mathcal{R}}_\rho(a,h)\le U_t-B_t.$$

If $\overline{r}_t=\max_{C\in\mathfrak{P}_t}r_C$ over the remaining cells, then

$$U_t-B_t\le 2L_R\overline{r}_t.$$

Midpoint bisection of a widest coordinate interval in a cell attaining the smallest lower bound gives $U_t-B_t\to 0$. If the tangent Hessian for a fixed h is bounded below by $\kappa_R>0$ on $\mathcal{A}_\delta$, then

$$\widehat{\mathcal{R}}_\rho(a,h)-\inf_{a'\in\mathcal{A}_\delta}\widehat{\mathcal{R}}_\rho(a',h)\le\frac{s_R(a,h)^2}{2\kappa_R}.$$

In particular, $\rho/m > M_R$ gives $\kappa_R = \rho/m - M_R$. If L_g is a gradient Lipschitz constant, projected gradient descent with step L_g^{-1} then satisfies

$$\widehat{\mathcal{R}}_\rho(a^{(t)}, h) - \widehat{\mathcal{R}}_\rho(a_h^*, h) \le (1 - \kappa_R/L_g)^t \{\widehat{\mathcal{R}}_\rho(a^{(0)}, h) - \widehat{\mathcal{R}}_\rho(a_h^*, h)\},$$

where a_h^* is the global minimizer for the fixed bandwidth. Without the curvature condition, projected gradient descent with an Armijo line search decreases the criterion, and every accumulation point is stationary.

Proposition 19 links a computed solution to the statistical estimator by taking $\eta_R = U_t - B_t$. The partition is refined only where its lower bound remains competitive and need not evaluate a fixed exhaustive net. Interval bounds on the gradient give valid cellwise Lipschitz constants, while interval bounds on the tangent Hessian give an upper bound for M_R. When $\rho/m > M_R$, each bandwidth requires one strongly convex fit and the discrete bandwidth values are compared directly. Without a global objective or curvature certificate, stationarity is reported but is not treated as a global objective bound.

A computable loss bound avoids unknown derivative constants in the global certificate. For a block response matrix V, subtract the midpoint of each column range and let B_V bound the absolute centered entries. This translation does not change the residual because every row of W sums to one. Write ℓ_b^V for the resulting prediction loss. The following proposition gives a finite-sample bound from the observed measurements.

Proposition 20. *For every $a, c \in \mathcal{A}_\delta$ and fixed h, ξ,*

$$|\ell_b^V(a, h; \xi) - \ell_b^V(c, h; \xi)| \le 4B_V^2 \max_{i \ne j} |\log K_{ij}(a, h; \xi) - \log K_{ij}(c, h; \xi)|.$$

Proposition 20 gives a lower bound on every parameter cell using linear optimization of the log kernel over that cell. On a box intersected with the trace constraint, each linear optimization is a continuous knapsack problem. The cell bound is the average of the nonnegative block lower bounds plus the minimum ridge penalty on the cell. Numerical errors in the evaluated block losses are subtracted from these lower bounds and added to candidate upper bounds. A finite computational budget produces an achieved objective gap; it does not imply that a requested tolerance has been met.

Proof of Proposition 20. Put $a_t = c + t(a - c)$ and $C_{ij} = \partial_t \log K_{ij}(a_t, h; \xi)$ for $i \ne j$, with $C_{ii} = 0$. The matrix C is constant in t. With $z_i = \log u_i$, implicit differentiation gives

$$\dot{W}_{ij} = W_{ij}(C_{ij} + \dot{z}_i + \dot{z}_j), \qquad \dot{W}\mathbf{1} = 0.$$

The zero row sums imply

$$\sum_{i \ne j} \frac{\dot{W}_{ij}^2}{W_{ij}} = \sum_{i \ne j} \dot{W}_{ij} C_{ij} \le \Big(\sum_{i \ne j} \frac{\dot{W}_{ij}^2}{W_{ij}} \Big)^{1/2} \Big(\sum_{i \ne j} W_{ij} C_{ij}^2 \Big)^{1/2}.$$

Thus the first sum is at most $b\|C\|_{\max}^2$, where $\|C\|_{\max} = \max_{i \ne j} |C_{ij}|$. If q_V is the number of response columns, weighted Cauchy–Schwarz gives

$$\|\dot{W}V\|_F^2 \le q_V B_V^2 \sum_{i \ne j} \frac{\dot{W}_{ij}^2}{W_{ij}} \le b q_V B_V^2 \|C\|_{\max}^2.$$

Because $\|(I - W)V\|_F \le 2B_V \sqrt{b q_V}$,

$$|\partial_t \ell_b^V(a_t, h; \xi)| \le \frac{2}{b q_V} \|(I - W)V\|_F \|\dot{W}V\|_F \le 4B_V^2 \|C\|_{\max}.$$

Integration over $0 \le t \le 1$ proves the assertion. This completes the proof. □

The cost for a batch of size b is $O(b^2 m)$ for pairwise distances, $O(b^2)$ per symmetric scaling iteration, and the cost of solving a system with $I_b + W$. A direct solve costs $O(b^3)$; conjugate gradients use matrix-vector products of cost $O(b^2)$. For positive neighbor radii, the final affinity stores

$O(nk_n)$ edges, and a Lanczos iteration uses a matrix-vector product of the same order. Section R.1 proves this edge bound and gives exact block computation without a dense distance matrix. Exact neighbor search can still cost $O(n^2m)$; fixed-dimensional spatial indexing reduces this search cost without changing the graph definition. Supplementary Section R gives stable log-domain scaling, the branch-and-bound bounds, and stopping rules. Supplementary Section N gives joint metric and bandwidth estimation when a continuous prediction bandwidth is needed.

G Further statistical and spectral results

Metric estimation and spectral representation have different statistical roles, but they need not use disjoint observations. The block criterion estimates a regularized predictive metric, and the representation graph uses the selected coordinates of this metric. We first define the population target, quantify its geometric interpretation, and then connect metric estimation, support selection, and spectral convergence.

This risk defines the regularized predictive metric. A fixed ρ is part of this target, and the target need not equal a scientific metric specified outside the observed distribution. Write $\mathcal{R}_{\infty,0}$ for (10) with $\rho = 0$, and let $(a^\dagger, h^\dagger)$ minimize this unpenalized risk. Every minimizer (a_0, h_0) of (10) satisfies

$$\mathcal{R}_{\infty,0}(a_0, h_0) - \mathcal{R}_{\infty,0}(a^\dagger, h^\dagger) \le \frac{\rho}{2m}\|a^\dagger - \mathbf{1}\|_2^2.$$

Thus the regularization contribution is explicit. If the unpenalized minimizer is unique, compactness gives $(a_0, h_0) \to (a^\dagger, h^\dagger)$ as $\rho \downarrow 0$.

The following proposition identifies the object approximated by the graph smoother.

Proposition 21. *For every square-integrable $g(X_\xi)$,*

$$\mathbb{E}\|X_{\xi^c} - g(X_\xi)\|_2^2 = \mathbb{E}\|X_{\xi^c} - m_\xi(X_\xi)\|_2^2 + \mathbb{E}\|g(X_\xi) - m_\xi(X_\xi)\|_2^2.$$

The population graph predictor based on X_ξ equals $T_{a,h,\xi} m_\xi(X_\xi)$.

Proposition 21 separates irreducible conditional variance from approximation of the conditional mean. The second assertion is specific to the graph smoother and connects the population operator to (10).

The conditional means have a geometric interpretation under repeated measurements. Let Z lie on a compact connected d-dimensional manifold $\mathcal{M}$. Suppose $X_\xi = \phi_\xi(Z)$ and there is a continuously differentiable map ψ_ξ such that

$$\mathbb{E}(X_{\xi^c} \mid X_\xi) = \psi_\xi(Z) \quad \text{almost surely.}$$

For mask probabilities p_ξ, define the aggregate map

$$\Psi(z) = \left\{ p_\xi^{1/2} |\xi^c|^{-1/2} \psi_\xi(z) : p_\xi > 0 \right\}.$$

Let $d_{\mathcal{M}}$ be the geodesic distance for a fixed smooth Riemannian metric on $\mathcal{M}$. Write $T_z\mathcal{M}$ for its tangent space at z and $D\psi_\xi(z)$ for the differential of ψ_ξ. The next proposition states when prediction distinguishes the latent state.

Proposition 22. *Suppose each ψ_ξ is continuously differentiable and Ψ is a one-to-one immersion. There are constants $0 < c_\Psi < C_\Psi < \infty$ such that*

$$c_\Psi d_{\mathcal{M}}(z, z') \le \|\Psi(z) - \Psi(z')\|_2 \le C_\Psi d_{\mathcal{M}}(z, z')$$

for all $z, z' \in \mathcal{M}$. The matrix

$$G_{\mathrm{pred}}(z) = \sum_{\xi : p_\xi > 0} \frac{p_\xi}{|\xi^c|} D\psi_\xi(z)^\top D\psi_\xi(z)$$

is positive definite on every tangent space.

Proposition 22 shows that repeated conditional means define a Riemannian geometry on the latent state. Stack the coordinate maps as $\phi = (\phi_1, \dots, \phi_m)^\top$. A diagonal affinity metric a induces the pullback metric

$$g_a(z) = D\phi(z)^\top \operatorname{diag}(a) D\phi(z).$$

Write d_g for the geodesic distance induced by a Riemannian metric g. The following proposition first gives a global relation between these two geometries.

Proposition 23. *Suppose the assumptions of Proposition 22 hold and ϕ is a one-to-one immersion. There are constants $0 < c_G < C_G < \infty$ such that, for every $a \in \mathcal{A}_\delta$, $z \in \mathcal{M}$, and $v \in T_z\mathcal{M}$,*

$$c_G v^\top G_{\mathrm{pred}}(z) v \le v^\top g_a(z) v \le C_G v^\top G_{\mathrm{pred}}(z) v.$$

The geodesic distances induced by g_a and G_{pred} satisfy the corresponding uniform bi-Lipschitz bounds.

Proposition 23 shows that every feasible affinity metric preserves the predictive manifold up to uniform distortion. Selection of a determines which predictive directions are smoother. The next result makes this selection explicit for a local prediction bandwidth.

For every mask of positive probability, suppose ϕ_ξ is an embedding. Let $g_{a,\xi}$ be the metric induced by $\mathrm{diag}(a_\xi/\pi)^{1/2}\phi_\xi$, and let $p_{a,\xi}$ be the density with respect to its Riemannian volume measure. Write $\Delta_g = \mathrm{div}_g \nabla_g$ for the Laplace–Beltrami operator. Define

$$\mathcal{B}_{a,\xi} f = -\frac{1}{2}\Big\{\Delta_{g_{a,\xi}} f + \langle \nabla_{g_{a,\xi}} \log p_{a,\xi}, \nabla_{g_{a,\xi}} f\rangle_{g_{a,\xi}}\Big\}, \quad \mathcal{J}(a) = \mathbb{E}_\xi\left[\frac{1}{|\xi^c|}\int_{\mathcal{M}} \|\mathcal{B}_{a,\xi}\psi_\xi(z)\|_2^2 \, dP(z)\right].$$

Let

$$\mathcal{V} = \mathbb{E}_\xi\left[\frac{1}{|\xi^c|}\mathbb{E}\{\|X_{\xi^c} - m_\xi(X_\xi)\|_2^2\}\right].$$

Theorem 9. *Suppose m is fixed, $\mathcal{M}$ has no boundary, every ϕ_ξ is an embedding, and ϕ_ξ and ψ_ξ are C^6, uniformly over the finite mask support. Suppose the density of P with respect to each $dV_{g_{\mathbf{1},\xi}}$ is C^6 and bounded above and away from zero. Then*

$$\sup_{a\in\mathcal{A}_\delta} \left|h^{-4}\{\mathcal{R}_{\infty,0}(a,h) - \mathcal{V}\} - \mathcal{J}(a)\right| \longrightarrow 0 \qquad \text{as } h \downarrow 0.$$

If $\rho_h/h^4 \to \rho_\star$ and a_h minimizes $\mathcal{R}_{\infty,\rho_h}(a,h)$ over $\mathcal{A}_\delta$, every accumulation point of a_h minimizes

$$\mathcal{J}(a) + \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2.$$

If this minimizer is unique, a_h converges to it.

Theorem 9 identifies a general geometric target without equating two different metric tensors. The limiting criterion minimizes the squared weighted-Laplacian roughness of the excluded conditional means plus the limiting ridge penalty. The following corollary gives an identification rate for the resulting pullback metric.

Let

$$\mathcal{J}_{\rho_\star}(a) = \mathcal{J}(a) + \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2.$$

Let $\kappa_J > 0$ be a curvature constant.

Corollary 9. *Suppose the assumptions of Theorem 9 hold and $a_\star$ is the unique minimizer of $\mathcal{J}_{\rho_\star}$. If*

$$\mathcal{J}_{\rho_\star}(a) - \mathcal{J}_{\rho_\star}(a_\star) \ge \kappa_J \|a - a_\star\|_2^2, \qquad a \in \mathcal{A}_\delta,$$

then

$$\|a_h - a_\star\|_2 = O\left\{h + \left|\frac{\rho_h}{h^4} - \rho_\star\right|^{1/2}\right\}.$$

Corollary 9 identifies $g_{a_\star}$ as the predictive geometry that minimizes $\mathcal{J}_{\rho_\star}$. The following proposition translates the parameter bound into the metric tensor and its geodesic distance. This is the relation used when a fixed positive prediction bandwidth is chosen.

For $h > 0$ and $\rho \ge 0$, let $a(h,\rho)$ minimize $\mathcal{R}_{\infty,\rho}(a,h)$ over $\mathcal{A}_\delta$.

Proposition 24. *Suppose the assumptions of Corollary 9 hold. For all sufficiently small h,*

$$\|a(h,\rho) - a_\star\|_2 \le C\left\{h^2 + \left|\frac{\rho}{h^4} - \rho_\star\right|\right\}^{1/2}.$$

There is a constant $C_g < \infty$ such that

$$\sup_{z\in\mathcal{M}} \|g_{a(h,\rho)}(z) - g_{a_\star}(z)\|_{\mathrm{op}} \le C_g\|a(h,\rho) - a_\star\|_2, \quad \sup_{z\ne z'}\left|\frac{d_{g_{a(h,\rho)}}(z,z')}{d_{g_{a_\star}}(z,z')} - 1\right| \le C_g\|a(h,\rho) - a_\star\|_2.$$

Proposition 24 gives a quantitative claim about the metric estimated by the method. It does not identify $g_{a_\star}$ with G_{pred}. The conditional-mean tensor describes which latent states are distinguishable, while $g_{a_\star}$ is the diagonal pullback metric that minimizes predictive roughness plus the limiting ridge penalty. In the geometry-calibrated version of the method, h and ρ/h^4 are chosen to make the bound in Proposition 24 smaller than a prespecified tolerance.

Proposition 1 identifies the nuisance contribution for a product view. We next use that identity for the full random-mask criterion.

Partition the feature indices into a signal set $\mathcal{S}$ and a nuisance set $\mathcal{N}$, with cardinalities s and q. Suppose

$$P_X = P_{X_\mathcal{S}} \otimes \bigotimes_{j\in\mathcal{N}} P_j.$$

For a mask ξ, let $m_{\xi,\mathcal{S}}$ contain the conditional means of the held-out signal coordinates given $X_{\xi\cap\mathcal{S}}$, and let $T^{\mathcal{S}}_{a,h,\xi}$ be the bistochastic operator formed from the observed signal coordinates. Put

$$Q_h(u) = \mathbb{E}_\xi\left[\frac{1}{|\xi^c|}\|m_{\xi,\mathcal{S}} - T^{\mathcal{S}}_{u,h,\xi} m_{\xi,\mathcal{S}}\|^2_{L^2(P)}\right].$$

The term for a mask with no held-out signal coordinate is zero. Define

$$F_h(u) = Q_h(u) + \frac{\rho}{2m}\|u - \mathbf{1}_s\|_2^2, \qquad \mathcal{A}_\mathcal{S} = \{u \in [\delta, A_m]^s : \mathbf{1}_s^\top u = m - q\delta\}.$$

Let u° be an interior point of $\mathcal{A}_\mathcal{S}$ and let λ° satisfy

$$\nabla F_h(u^\circ) + \lambda^\circ \mathbf{1}_s = 0.$$

The following proposition gives a direct identification result for the mask distribution used by the estimator.

Proposition 25. *Suppose the mask support is finite and every observed coordinate kernel admits a positive symmetric bistochastic scaling. Then*

$$\mathcal{R}_{\infty,\rho}(a,h) = \mathcal{V} + Q_h(a_\mathcal{S}) + \frac{\rho}{2m}\|a - \mathbf{1}\|_2^2.$$

If, for some $\kappa_\mathcal{S} > 0$,

$$\nabla^2 F_h(u) \succeq \kappa_\mathcal{S} I_s, \quad u \in [\delta, A_m]^s, \qquad \lambda^\circ + \frac{\rho}{m}(\delta - 1) > 0,$$

then the unique minimizer is

$$a_0 = (u^\circ, \delta\mathbf{1}_q).$$

The nuisance constraints are strictly complementary, and the Hessian is positive definite on the critical cone. At $a = \mathbf{1}$, the unpenalized screening scores satisfy

$$d_j(h) = -m\partial_j Q_h(\mathbf{1}_s), \quad j \in \mathcal{S}, \qquad d_j(h) = 0, \quad j \in \mathcal{N}.$$

Proposition 25 concerns the aggregate risk of any finite mask distribution, including uniform fixed-count masks. It does not require the maskwise risks to be separable. The curvature and multiplier inequalities are population properties of the random-mask criterion and can be evaluated by deterministic integration in the models used below.

A paired mask gives a primitive verification of these aggregate conditions. Fix h, let the signal coordinates be (U_ℓ, V_ℓ), $1 \le \ell \le r$, and let $N_1, \dots, N_q$ be nuisance coordinates, where q is even and all pairs and nuisance coordinates are mutually independent. Here $s = 2r$. A paired mask observes exactly one member of each signal pair with an independent fair orientation and selects $q/2$ nuisance coordinates uniformly. Thus $m = 2r + q$, $\pi = 1/2$, and $|\xi^c| = m/2$. For a signal coordinate j, let $\bar{j}$ denote its partner, put $\mu_j(x) = \mathbb{E}(X_{\bar{j}} \mid X_j = x)$, and define

$$r_j(t) = \|\mu_j - T_{j,t}\mu_j\|^2_{L^2(P_j)}, \qquad \omega_j = m^{-1}.$$

Here $T_{j,t}$ is the coordinatewise bistochastic operator with Gaussian kernel $\exp\{-t(x - x')^2/(2\pi h^2)\}$. Both $T_{j,t}$ and r_j use the inclusion probability π of the mask design under consideration. Let

$$F_{\mathrm{pair}}(u) = \sum_{j\in\mathcal{S}} \omega_j r_j(u_j) + \frac{\rho}{2m}\sum_{j\in\mathcal{S}}(u_j - 1)^2$$

on $\mathcal{A}_\mathcal{S}$, and define u° and λ° by the corresponding interior Karush–Kuhn–Tucker equations.

Proposition 26. *Suppose every r_j is twice continuously differentiable,*

$$\min_{j\in\mathcal{S}} \inf_{\delta\le t\le A_m} \left\{\omega_j r_j''(t) + \frac{\rho}{m}\right\} \ge \kappa_{\mathrm{pair}} > 0, \qquad \lambda^\circ + \frac{\rho}{m}(\delta - 1) > 0.$$

Then Proposition 25 holds with

$$Q_h(u) = \sum_{j\in\mathcal{S}} \omega_j r_j(u_j), \qquad a_0 = (u^\circ, \delta \mathbf{1}_q).$$

If $r_j'(1) < 0$ for every $j \in \mathcal{S}$, the screening scores are strictly positive on $\mathcal{S}$ and zero on $\mathcal{N}$. Put $\chi = \lambda^\circ + \rho(\delta - 1)/m$. For every $j \in \mathcal{N}$ and feasible a with $a_j > \delta$, there is an $\ell \in \mathcal{S}$ with $a_\ell < u_\ell^\circ$ such that

$$\frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_j} \le -\chi < 0.$$

Proposition 26 verifies exact nuisance removal, score separation, and the global exchange condition under a random balanced design. Every mask observes $m/2$ coordinates. The two orientations ensure that each repeated measurement constructs a graph in one task and supplies a prediction target in another.

The unrestricted fixed-count mask also has a separable aggregate risk in this model. If it selects k of the m coordinates uniformly, evaluate $T_{j,t}$ and r_j at $\pi = k/m$ and put

$$\omega_k = \frac{k}{m(m-1)}, \qquad F_k(u) = \sum_{j\in\mathcal{S}} \omega_k r_j(u_j) + \frac{\rho}{2m} \sum_{j\in\mathcal{S}} (u_j - 1)^2.$$

Let $u^{\circ,k}$ and $\lambda^{\circ,k}$ satisfy the interior Karush–Kuhn–Tucker equations for F_k on $\mathcal{A}_{\mathcal{S}}$.

Corollary 10. *Suppose*

$$\min_{j\in\mathcal{S}} \inf_{\delta\le t\le A_m} \left\{\omega_k r_j''(t) + \frac{\rho}{m}\right\} > 0, \qquad \lambda^{\circ,k} + \frac{\rho}{m}(\delta - 1) > 0.$$

Then

$$Q_h(u) = \sum_{j\in\mathcal{S}} \omega_k r_j(u_j), \qquad a_0 = (u^{\circ,k}, \delta \mathbf{1}_q).$$

All other conclusions of Proposition 26 hold after replacing ω_j, u°, and λ° by ω_k, $u^{\circ,k}$, and $\lambda^{\circ,k}$.

Corollary 10 supplies primitive identification conditions for the default uniform fixed-count mask. The factor ω_k is its ordered-pair separation probability divided by the fixed number of held-out coordinates.

The preceding results identify the metric at a fixed bandwidth. The following proposition supplies the remaining bandwidth comparison for a finite candidate collection. When the kernel depends on a coordinate weight and the bandwidth through their ratio, write

$$r_j(t,h) = \widetilde{r}_j(t/h^2).$$

Proposition 27. *Suppose $\mathcal{H}$ is finite, $|\mathcal{H}| \ge 2$, $h_0 = \min \mathcal{H}$, and*

$$Q_h(u) = \sum_{j\in\mathcal{S}} \omega_j \widetilde{r}_j(u_j/h^2), \qquad \omega_j > 0.$$

Assume that, for some $c_h > 0$,

$$\inf_{j\in\mathcal{S}} \inf_{\delta/h_{\max}^2 \le v \le A_m/h_0^2} \{-\widetilde{r}_j'(v)\} \ge c_h, \qquad h_{\max} = \max \mathcal{H}.$$

If the curvature and multiplier conditions in Proposition 25 hold at h_0, the joint criterion has a unique minimizer (a_0, h_0). For some $r_0, \kappa > 0$,

$$\mathcal{R}_{\infty,\rho}(a, h_0) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \ge \kappa \|a - a_0\|_2^2, \qquad \|a - a_0\|_2 \le r_0,$$

and the bandwidth gap can be chosen as

$$\Delta_h = c_h \delta \sum_{j\in\mathcal{S}} \omega_j \min_{h\in\mathcal{H},\, h > h_0} \left(h_0^{-2} - h^{-2}\right) > 0.$$

Proposition 27 distinguishes metric identification from bandwidth identification. In this monotone subfamily, the population criterion selects the smallest bandwidth in the fixed candidate set. The derivative inequality and the curvature and multiplier conditions give the positive bandwidth gap. This is a sufficient identification example, not a guarantee of an interior optimal bandwidth or of optimal selection over arbitrary candidate families.

The same product structure connects random masks to the local geometry theory without placing nuisance variables on the manifold. Suppose $X_{\mathcal{S}} = \phi_{\mathcal{S}}(Z)$ for $Z \in \mathcal{M}$, while the nuisance coordinates are mutually independent and independent of Z. For each mask, write $\phi_{\xi,\mathcal{S}}$ for the observed signal map and $\psi_{\xi,\mathcal{S}}$ for the conditional mean of the held-out signal coordinates. Let Ξ_1 contain the masks for which $\psi_{\xi,\mathcal{S}}$ is nonconstant. For $\xi \in \Xi_1$, let $g^{\mathcal{S}}_{a,\xi}$ be the metric induced by $\operatorname{diag}(a_{\xi\cap\mathcal{S}}/\pi)^{1/2}\phi_{\xi,\mathcal{S}}$, let $p^{\mathcal{S}}_{a,\xi}$ be the sampling density with respect to its Riemannian volume measure, and define

$$\mathcal{B}^{\mathcal{S}}_{a,\xi} f = -\frac{1}{2}\left\{\Delta_{g^{\mathcal{S}}_{a,\xi}} f + \langle \nabla_{g^{\mathcal{S}}_{a,\xi}} \log p^{\mathcal{S}}_{a,\xi}, \nabla_{g^{\mathcal{S}}_{a,\xi}} f\rangle_{g^{\mathcal{S}}_{a,\xi}}\right\}.$$

Put

$$\mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) = \sum_{\xi\in\Xi_1} \frac{p_\xi}{|\xi^c|}\int_{\mathcal{M}} \|\mathcal{B}^{\mathcal{S}}_{a,\xi}\psi_{\xi,\mathcal{S}}(z)\|_2^2\, dP(z).$$

Theorem 10. *Suppose m is fixed, $\mathcal{M}$ is compact without boundary, the mask support is finite, and every observed nuisance kernel admits a positive symmetric bistochastic scaling. For every $\xi \in \Xi_1$, suppose $\phi_{\xi,\mathcal{S}}$ is a C^6 embedding, $\psi_{\xi,\mathcal{S}}$ is C^6, and the sampling density is C^6 and bounded above and away from zero with respect to $dV_{g^{\mathcal{S}}_{\mathbf{1},\xi}}$. For $\xi \notin \Xi_1$, suppose $\psi_{\xi,\mathcal{S}}$ is constant. Then*

$$\sup_{a\in\mathcal{A}_\delta} \left|h^{-4}\{\mathcal{R}_{\infty,0}(a,h) - \mathcal{V}\} - \mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}})\right| = O(h^2).$$

If $\rho_h/h^4 \to \rho_\star$, every accumulation point of a minimizer a_h minimizes

$$\mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) + \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2.$$

If this minimizer is unique, a_h converges to it.

Theorem 10 permits independent nuisance coordinates in every observed view. Their kernel factors cancel exactly from the conditional-mean predictor, while held-out nuisance coordinates contribute only to $\mathcal{V}$. If the aggregate map formed from the $\psi_{\xi,\mathcal{S}}$ is a one-to-one immersion, Proposition 22 gives a positive definite predictive tensor on $\mathcal{M}$. The paired balanced masks above satisfy this formulation whenever every oriented signal view is an embedding.

Let $a^{\mathrm{mask}}_\star$ be the unique minimizer in Theorem 10, and define

$$g^{\mathcal{S}}_a(z) = D\phi_{\mathcal{S}}(z)^\top \operatorname{diag}(a_{\mathcal{S}}) D\phi_{\mathcal{S}}(z).$$

For the tensor and geodesic assertions, suppose that $\phi_{\mathcal{S}}$ is an embedding. This follows from the preceding assumptions whenever Ξ_1 is nonempty. The following corollary quantifies the geometry estimated under random masks.

Corollary 11. *Suppose the assumptions of Theorem 10 hold and, for some $\kappa_{\mathrm{mask}} > 0$,*

$$\mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) + \frac{\rho_\star}{2m}\|a-\mathbf{1}\|_2^2 - \mathcal{J}_{\mathrm{mask}}\{(a^{\mathrm{mask}}_\star)_{\mathcal{S}}\} - \frac{\rho_\star}{2m}\|a^{\mathrm{mask}}_\star - \mathbf{1}\|_2^2 \geq \kappa_{\mathrm{mask}}\|a - a^{\mathrm{mask}}_\star\|_2^2, \qquad a \in \mathcal{A}_\delta.$$

Then

$$\|a_h - a^{\mathrm{mask}}_\star\|_2 = O\left\{h + \left|\frac{\rho_h}{h^4} - \rho_\star\right|^{1/2}\right\}.$$

The tensor and geodesic bounds in Proposition 24 hold with g_a replaced by $g^{\mathcal{S}}_a$.

Corollary 11 gives the same finite-bandwidth control for uniform fixed-count and paired balanced masks when every informative observed signal view is an embedding. At every bandwidth and ridge value for which Proposition 26 holds, the nuisance coordinates equal δ before representation thresholding, and the tensor uses only the selected signal map.

Consider a prescribed view split in which the response view never enters the distance. Let Z take values -1 and 1 with equal probability, let N be independent of Z, and let $V = Z + \sigma\varepsilon$ with independent mean-zero unit-variance noise. The observed view is (Z, N), and its product Gaussian kernel has weights $a_1 + a_2 = 2$.

Proposition 28. *In the binary latent model,*

$$T_a V(Z,N) = \tanh(a_1/h^2)Z, \qquad \mathbb{E}\{V - T_a V(Z,N)\}^2 = \{1-\tanh(a_1/h^2)\}^2 + \sigma^2.$$

The unpenalized risk over $\mathcal{A}_\delta$ has the unique minimizer $(a_1, a_2) = (2-\delta, \delta)$.

Proposition 28 verifies boundary identification in a repeated-signal model. Supplementary Proposition 45 gives the same conclusion for noisy Gaussian measurements. These examples motivate a boundary-aware limit theory; an interior normal approximation does not describe their optimal nuisance weights.

The same conclusion holds for a general product distribution. For integers $s, q \geq 1$, let (U_ℓ, V_ℓ), $1 \leq \ell \leq s$, be mutually independent signal pairs, and let $N_1, \ldots, N_q$ be independent nuisance coordinates that are also independent of the signal pairs. Put $m = s + q$ and $\mu_\ell(u) = \mathbb{E}(V_\ell \mid U_\ell = u)$. For a fixed prediction bandwidth, suppose the observed-view kernel factorizes over these m coordinates and each coordinate kernel has a positive bistochastic scaling. Let $T_{\ell,t}$ be the population bistochastic operator for the ℓth signal kernel with weight t, and define

$$r_\ell(t) = \mathbb{E}\{V_\ell - T_{\ell,t}\mu_\ell(U_\ell)\}^2.$$

For fixed constants $\omega_\ell > 0$ with $\sum_{\ell=1}^s \omega_\ell = 1$, put

$$F_S(u) = \sum_{\ell=1}^s \omega_\ell r_\ell(u_\ell) + \frac{\rho}{2m}\sum_{\ell=1}^s (u_\ell - 1)^2, \qquad \mathcal{A}_S = \left\{u \in [\delta, A_m]^s : \mathbf{1}^\top u = m - q\delta\right\}.$$

Let a_S° be an interior minimizer of F_S over $\mathcal{A}_S$, and let λ° satisfy

$$\omega_\ell r_\ell'(a_{S\ell}^\circ) + \frac{\rho}{m}(a_{S\ell}^\circ - 1) + \lambda^\circ = 0, \qquad 1 \leq \ell \leq s.$$

The following proposition gives nonparametric sufficient conditions for exact nuisance removal. It reduces identification to one-dimensional prediction-risk curves and one multiplier inequality.

Proposition 29. *Suppose every r_ℓ is twice continuously differentiable,*

$$\min_{1\leq \ell \leq s} \inf_{\delta \leq t \leq A_m} \left\{\omega_\ell r_\ell''(t) + \frac{\rho}{m}\right\} \geq \kappa_S > 0, \qquad \lambda^\circ + \frac{\rho}{m}(\delta - 1) > 0.$$

Then the regularized population risk has the unique minimizer

$$a_0 = (a_S^\circ, \delta \mathbf{1}_q).$$

The nuisance constraints are strictly complementary, and the Hessian is positive definite on the critical cone. If $r_\ell'(1) < 0$ for every signal coordinate, the unpenalized screening scores at $a = \mathbf{1}$ satisfy

$$d_\ell(h) = -m\omega_\ell r_\ell'(1) > 0, \quad 1 \leq \ell \leq s, \qquad d_\ell(h) = 0, \quad s < \ell \leq m.$$

Put $\chi = \lambda^\circ + \rho(\delta - 1)/m$. For every nuisance coordinate j and every feasible a with $a_j > \delta$, there is a signal coordinate ℓ with $a_\ell < a_{S\ell}^\circ$ such that

$$\frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_j} \leq -\chi < 0.$$

Proposition 29 allows heterogeneous signal distributions, noise levels, and coordinate kernels. The multiplier inequality states that transferring metric mass from a nuisance coordinate to the optimized signal metric decreases the risk. Its uniform exchange conclusion also verifies the global restriction in Proposition 16 once the empirical gradient is uniformly accurate.

We next verify these conditions in a multichannel Gaussian model. Let $Z_1, \ldots, Z_s$ be independent standard Gaussian variables, and observe

$$X_\ell = Z_\ell + \varepsilon_\ell, \qquad V_\ell = Z_\ell + e_\ell, \qquad \ell = 1, \ldots, s,$$

where $\varepsilon_1, \ldots, \varepsilon_s$ are independent $N(0, \sigma_X^2)$ variables, independent of $Z = (Z_1, \ldots, Z_s)$. Write $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_s)$ and $e = (e_1, \ldots, e_s)$. The response errors satisfy $\mathbb{E}(e \mid Z, \varepsilon) = 0$ and $\mathbb{E}e_\ell^2 = \sigma_V^2$; they need not be Gaussian or independent across channels. Let $N_1, \ldots, N_q$ be independent Gaussian nuisance coordinates, independent of every signal variable and error. The observed view is $(X_1, \ldots, X_s, N_1, \ldots, N_q)$ and $V = (V_1, \ldots, V_s)$ is held out. Put $v_X = 1 + \sigma_X^2$ and

$$r(a) = \frac{2(a/h^2)v_X}{1 + \sqrt{1 + 4(a/h^2)^2 v_X^2}}, \qquad f(a) = 1 + \sigma_V^2 - \frac{2r(a) - r(a)^2}{v_X}.$$

The candidate signal weight is $a_S = (m - q\delta)/s$.

Proposition 30. *Suppose $\delta < a_S < A_m$ and*

$$\frac{f'(a_S)}{s} + \frac{\rho}{m}(a_S - \delta) < 0.$$

Then the regularized population risk has the unique minimizer

$$a_{0\ell} = a_S, \quad 1 \le \ell \le s, \qquad a_{0\ell} = \delta, \quad s < \ell \le m.$$

The nuisance constraints are strictly complementary, and the Hessian is positive definite on the critical cone. At $a = \mathbf{1}$, the unpenalized screening scores satisfy

$$d_\ell(h) = -\frac{m}{s} f'(1) > 0, \quad 1 \le \ell \le s, \qquad d_\ell(h) = 0, \quad s < \ell \le m.$$

Proposition 30 gives primitive and directly checkable identification conditions. It verifies uniqueness, local curvature, exact nuisance support, strict complementarity, and the screening separation in one model with $s+q$ observed coordinates. Its inequality is the multiplier condition in Proposition 29 with $\omega_\ell = 1/s$, and hence also gives the uniform exchange margin. The condition allows a positive ridge penalty when its shrinkage toward one is weaker than the predictive gain from transferring metric mass to a signal coordinate.

The expected block criterion $\mathcal{R}_{b,\rho}$ differs from the population risk because its smoother is estimated from b observations. Conditions 1 and 2 concern the population distribution and its predictive metric.

The next proposition closes the gap between the finite block predictor and the population operator.

Proposition 31. *Suppose Condition 1 holds. There is a constant C_{bl} independent of b and m such that*

$$\sup_{a \in \mathcal{A}_\delta,\, h \in \mathcal{H}} |\mathcal{R}_{b,\rho}(a,h) - \mathcal{R}_{\infty,\rho}(a,h)| \le C_{\mathrm{bl}} \sqrt{\frac{m \log(mb)}{b}}.$$

If $m \log(mb)/b \le c_0$, the same order holds for the first two derivatives on an interior neighborhood of a_0. For feasible one-sided derivatives, the assertion remains valid at a boundary point.

Proposition 31 bounds the difference between the finite-block risk $\mathcal{R}_{b,\rho}$ and (10). For fixed m, its bound tends to zero as b increases. Increasing R alone does not remove this approximation term.

The uniform approximation also connects the two minimizers. The following corollary quantifies this relation under population identification.

Corollary 12. *Suppose Conditions 1 and 2 hold and m is fixed. If $(a_{0,b}, h_{0,b})$ minimizes $\mathcal{R}_{b,\rho}$, then, for all sufficiently large b,*

$$h_{0,b} = h_0, \qquad \|a_{0,b} - a_0\|_2 \le C \left\{ \frac{m \log(mb)}{b} \right\}^{1/4}.$$

Corollary 12 shows that the finite-block target approaches the predictive population metric. In the product models above, attenuation of an independent nuisance coordinate is a population conclusion; a fixed block approximates that conclusion only through this limit.

Theorem 2 separates block sampling, finite-block approximation, and computation.

For a fixed feature dimension m and a fixed bandwidth set $\mathcal{H}$, Theorem 2 guarantees population-risk consistency under its stated conditions when $R \to \infty$, $b \to \infty$, and $\eta_R \to 0$ in probability. Crucially, these three fundamentally distinct error terms absolutely cannot be exchanged: merely introducing repeated masks mathematically reduces neither the underlying block sampling error nor the inherent finite-block bias, and establishing a small local gradient residual definitively does not by itself reliably bound the global objective gap η_R for a criterion that may be nonconvex.

Use the empirical and mask-averaged criteria $\widetilde{\mathcal{R}}_{R,J}$ and $\overline{\mathcal{R}}_R$ defined before Corollary 3. The following theorem gives a bound from parameter coverings.

Theorem 11. *Suppose Condition 1 holds. Conditionally on the observation blocks, with probability at least $1 - \alpha$,*

$$\sup_{a \in \mathcal{A}_\delta,\, h \in \mathcal{H}} |\widetilde{\mathcal{R}}_{R,J}(a,h) - \overline{\mathcal{R}}_R(a,h)| \le C_3 \sqrt{\frac{m \log(C_4 m R J) + \log(2|\mathcal{H}|/\alpha)}{RJ}}.$$

Theorem 11 treats masks as Monte Carlo draws conditional on the blocks. It does not count RJb as the number of independent observations.

Corollary 13. *Suppose Conditions 1 and 2 hold and m is fixed. An η_R-approximate minimizer based on J masks per block satisfies*

$$\|\widehat{a} - a_0\|_2 = O_{\mathbb{P}}\left[\left\{\frac{\log(2m|\mathcal{H}|)}{R}\right\}^{1/4} + \left\{\frac{\log(2m|\mathcal{H}|)}{RJ}\right\}^{1/4} + \left\{\frac{m\log(mb)}{b}\right\}^{1/4} + \eta_R^{1/2}\right].$$

Corollary 13 is a global population-target rate derived only from risk curvature. For fixed dimension and fixed mask support, Theorem 6 gives the sharper rate from the score expansion. A nondegenerate local distribution is stated for a fixed block size, whose statistical target is $\mathcal{R}_{b,\rho}$.

The stable-bandwidth results above estimate the exact finite-bandwidth target. A second regime estimates the minimizer $a_\star$ of $\mathcal{J}_{\rho_\star}$ itself. For $h > 0$, define

$$q_b(h) = \sqrt{\frac{\log b}{bh^{5d}}} + \frac{1}{bh^{3d}}.$$

These bounds control the exact symmetric scaling with a deleted diagonal. They follow from local-volume bounds and a monotonicity argument for the scaling equation; no pointwise scaling rate is assumed. The next two propositions compare expected block risks under finite response second moments. Theorem 12 uses bounded responses for concentration over blocks.

Proposition 32. *Suppose the assumptions of Theorem 9 hold and the mask support is finite. For fixed m, $0 < h \le 1$, and $bh^{2d} \ge C\log b$,*

$$\sup_{a\in\mathcal{A}_\delta} |\mathcal{R}_{b,0}(a,h) - \mathcal{R}_{\infty,0}(a,h)| \le C q_b(h).$$

Proposition 32 permits decreasing prediction bandwidths without a global kernel lower bound. The exponent $5d$ is a sufficient bound for the exact zero-diagonal scaling. No optimality of this exponent is asserted.

Population cancellation of an independent nuisance kernel does not remove its finite-sample localization cost. Under Theorem 10, suppose each nuisance coordinate has a C^4 density bounded above and away from zero on a compact interval. For every mask of positive probability, suppose the observed signal distribution has a C^4 density bounded above and away from zero on a compact $d_{\xi,\mathcal{S}}$-dimensional C^4 submanifold, which may have a boundary. Assume uniform local-volume bounds over the finite mask support. Put

$$d_\xi = d_{\xi,\mathcal{S}} + |\xi \cap \mathcal{N}|, \qquad d_{\max} = \max_{\xi: p_\xi > 0} d_\xi,$$

and

$$q_{b,\xi}^{\mathrm{prod}}(h) = \sqrt{\frac{\log b}{bh^{5d_\xi}}} + \frac{1}{bh^{3d_\xi}}, \qquad q_b^{\mathrm{prod}}(h) = \max_{\xi: p_\xi > 0} q_{b,\xi}^{\mathrm{prod}}(h).$$

Proposition 33. *Suppose the preceding product-support conditions hold. For fixed m, $0 < h \le 1$, and $bh^{2d_{\max}} \ge C\log b$,*

$$\sup_{a\in\mathcal{A}_\delta} |\mathcal{R}_{b,0}(a,h) - \mathcal{R}_{\infty,0}(a,h)| \le C \max_{\xi: p_\xi > 0} q_{b,\xi}^{\mathrm{prod}}(h) = C q_b^{\mathrm{prod}}(h).$$

Proposition 33 separates exact population nuisance cancellation from finite-sample kernel estimation. Each mask has its own local sample size bh^{d_ξ}, and the aggregate criterion is controlled by the largest observed-support dimension. No maskwise kernel-mass bound is stated with the dimension of a different mask.

Let $\widehat{a}_{\mathrm{geo}}$ be an η_R-approximate minimizer of the criterion with $\mathcal{H} = \{h_R\}$ and ridge ρ_R. Put

$$\Delta_{R,J,b}(h_R) = \sqrt{\frac{m\log(mR/h_R)}{R}} + \sqrt{\frac{m\log(mRJ/h_R)}{RJ}} + q_b(h_R) + \eta_R.$$

Theorem 12. *Suppose the assumptions of Corollary 9 hold and every response coordinate is bounded in absolute value by B. If*

$$h_R \to 0, \qquad \frac{\rho_R}{h_R^4} \to \rho_\star, \qquad \frac{\Delta_{R,J,b}(h_R)}{h_R^4} \to 0 \quad \textit{in probability},$$

then

$$\|\widehat{a}_{\mathrm{geo}} - a_\star\|_2 = O_{\mathbb{P}}\Big[h_R + \Big|\frac{\rho_R}{h_R^4} - \rho_\star\Big|^{1/2} + \Big\{\frac{\Delta_{R,J,b}(h_R)}{h_R^4}\Big\}^{1/2}\Big] = o_{\mathbb{P}}(1).$$

In addition,

$$\sup_{z\in\mathcal{M}} \|g_{\widehat{a}_{\mathrm{geo}}}(z) - g_{a_\star}(z)\|_{\mathrm{op}} + \sup_{z\neq z'} \Big|\frac{d_{g_{\widehat{a}_{\mathrm{geo}}}}(z,z')}{d_{g_{a_\star}}(z,z')} - 1\Big| = O_{\mathbb{P}}\Big[h_R + \Big|\frac{\rho_R}{h_R^4} - \rho_\star\Big|^{1/2} + \Big\{\frac{\Delta_{R,J,b}(h_R)}{h_R^4}\Big\}^{1/2}\Big].$$

Theorem 12 connects the computed block criterion to the local predictive geometry. Sufficient primitive orders are $Rh_R^8/\log(R/h_R) \to \infty$, $RJh_R^8/\log(RJ/h_R) \to \infty$, $bh_R^{5d+8}/\log b \to \infty$, and $\eta_R = o(h_R^4)$. The stable-bandwidth version remains preferable when m increases; its estimand is the exact risk minimizer a_0 and is not described as G_{pred}.

The random-mask product model has the same decomposition with the effective product-view dimension in Proposition 33. Put

$$\Delta^{\mathrm{prod}}_{R,J,b}(h_R) = \sqrt{\frac{m\log(mR/h_R)}{R}} + \sqrt{\frac{m\log(mRJ/h_R)}{RJ}} + q_b^{\mathrm{prod}}(h_R) + \eta_R$$

and

$$e^{\mathrm{prod}}_{R,J,b} = h_R + \Big|\frac{\rho_R}{h_R^4} - \rho_\star\Big|^{1/2} + \left\{\frac{\Delta^{\mathrm{prod}}_{R,J,b}(h_R)}{h_R^4}\right\}^{1/2}.$$

Corollary 14. *Suppose the assumptions of Corollary 11 and Proposition 33 hold,*

$$h_R \to 0, \qquad \frac{\rho_R}{h_R^4} \to \rho_\star, \qquad \frac{\Delta^{\mathrm{prod}}_{R,J,b}(h_R)}{h_R^4} \to 0 \quad \textit{in probability}.$$

Then

$$\|\widehat{a}_{\mathrm{geo}} - a^{\mathrm{mask}}_\star\|_2 + \sup_z \|g^{\mathcal{S}}_{\widehat{a}_{\mathrm{geo}}}(z) - g^{\mathcal{S}}_{a^{\mathrm{mask}}_\star}(z)\|_{\mathrm{op}} + \sup_{z\neq z'} \Big|\frac{d_{g^{\mathcal{S}}_{\widehat{a}_{\mathrm{geo}}}}(z,z')}{d_{g^{\mathcal{S}}_{a^{\mathrm{mask}}_\star}}(z,z')} - 1\Big| = O_{\mathbb{P}}(e^{\mathrm{prod}}_{R,J,b}).$$

Corollary 14 transfers block, mask-integration, and optimization error to the signal geometry under uniform fixed-count or paired balanced masks. A sufficient block order is $bh_R^{5d_{\max}+8}/\log b \to \infty$.

For fixed b, let $(a_{0,b}, h_{0,b})$ be the unique finite-block minimizer, put $A^-_{0,b} = \{j : a_{0,b,j} = \delta\}$, and define

$$\mathcal{T}_{0,b} = \{v \in \mathbb{R}^m : \mathbf{1}^\top v = 0,\ v_j \geq 0 \text{ for } j \in A^-_{0,b}\}.$$

Let $g_{0,b} = \nabla_a \mathcal{R}_{b,\rho}(a_{0,b}, h_{0,b})$ and define the critical cone

$$\mathcal{C}_{0,b} = \{v \in \mathcal{T}_{0,b} : g_{0,b}^\top v = 0\}.$$

Let

$$H_{0,b} = \nabla_a^2 \mathcal{R}_{b,\rho}(a_{0,b}, h_{0,b}), \qquad \Sigma_{0,b} = \mathrm{var}\left\{\nabla_a \ell_b(a_{0,b}, h_{0,b}; \xi) + \frac{\rho}{m}(a_{0,b} - \mathbf{1})\right\},$$

and let $Z_{0,b} \sim N(0, \Sigma_{0,b})$. The following theorem establishes the limit at an interior or boundary optimum.

Theorem 13. *Suppose Condition 1 holds, m and b are fixed, the finite-block minimizer is unique, and*

$$v^\top H_{0,b} v \geq \kappa \|v\|_2^2, \qquad v \in \mathrm{span}(\mathcal{C}_{0,b}).$$

If its bandwidth gap is positive and $\eta_R = o_{\mathbb{P}}(R^{-1})$, then $\mathbb{P}(\widehat{h} = h_{0,b}) \to 1$ and

$$\sqrt{R}(\widehat{a} - a_{0,b}) \Longrightarrow \operatorname*{arg\,min}_{v\in\mathcal{C}_{0,b}} \left\{Z_{0,b}^\top v + \frac{1}{2} v^\top H_{0,b} v\right\}.$$

Theorem 13 concerns constrained M-estimation (Geyer, 1994) for a finite-block predictive metric. The critical cone removes directions with a strictly positive first-order population loss. A strongly active nuisance weight is fixed at its boundary to first order, while a weakly active weight can have a nonnormal limit determined by the cone.

The curvature condition concerns differences of two critical directions. Positivity only along the cone does not ensure a unique minimizer when the Gaussian score is singular. Under strict complementarity, the critical cone is a linear space, and the condition reduces to curvature on the active signal tangent space. Proposition 25 verifies this case: its critical cone is $\{v : v_{\mathcal{N}} = 0,\ \mathbf{1}_s^\top v_{\mathcal{S}} = 0\}$, and its Hessian bound holds on that space. Proposition 29 gives the same conclusion for a prescribed response view. By Proposition 31, strict complementarity and positive tangent curvature also hold for the corresponding finite-block targets when b is sufficiently large.

When $a_{0,b}$ is interior, let $Q \in \mathbb{R}^{m\times(m-1)}$ have orthonormal columns spanning $\mathbf{1}^\perp$. Put $V_{Q,b} = (Q^\top H_{0,b} Q)^{-1} Q^\top \Sigma_{0,b} Q (Q^\top H_{0,b} Q)^{-1}$. For block r, define the fitted score $\widehat{g}_r$ and its mean $\overline{g}$ by

$$\widehat{g}_r = \nabla_a \ell_{I_r}(\widehat{a}, \widehat{h}; \xi_r) + \frac{\rho}{m}(\widehat{a} - \mathbf{1}), \qquad \overline{g} = R^{-1} \sum_{r=1}^{R} \widehat{g}_r.$$

The empirical Hessian and score covariance are

$$\widehat{H}_{0,b} = \nabla_a^2 \widehat{\mathcal{R}}_\rho(\widehat{a}, \widehat{h}), \qquad \widehat{\Sigma}_{0,b} = \frac{1}{R} \sum_{r=1}^{R} (\widehat{g}_r - \overline{g})(\widehat{g}_r - \overline{g})^\top.$$

Let $\widehat{V}_{Q,b}$ be $V_{Q,b}$ with these empirical matrices substituted. In the inverse of $Q^\top \widehat{H}_{0,b} Q$, replace eigenvalues below R^{-1} by R^{-1}. Positive population curvature makes this truncation inactive with probability tending to one. The following theorem establishes the standardized contrast limit and proves consistency of its variance estimator.

Theorem 14. *Under the assumptions of Theorem 13, suppose $A_{0,b}^-$ is empty. Then $\widehat{V}_{Q,b} \to V_{Q,b}$ in probability, and for every fixed $c \in \mathbf{1}^\perp$ with $c^\top Q V_{Q,b} Q^\top c > 0$,*

$$\frac{\sqrt{R}\, c^\top (\widehat{a} - a_{0,b})}{\{c^\top Q \widehat{V}_{Q,b} Q^\top c\}^{1/2}} \Longrightarrow N(0,1).$$

Theorem 14 is restricted to an interior finite-block metric. We next let the block size increase and center inference at the population target a_0.

For a mask ξ, write

$$r_\xi(a,h) = \frac{1}{|\xi^c|} \mathbb{E} \| X_{\xi^c} - T_{a,h,\xi} m_\xi(X_\xi) \|_2^2.$$

Put $A_0^- = \{j : a_{0j} = \delta\}$ and define

$$\mathcal{T}_0 = \{v : \mathbf{1}^\top v = 0,\ v_j \geq 0 \text{ for } j \in A_0^-\}.$$

Let

$$g_0 = \nabla_a \mathcal{R}_{\infty,\rho}(a_0, h_0), \qquad \mathcal{C}_0 = \{v \in \mathcal{T}_0 : g_0^\top v = 0\},$$

and

$$H_0 = \nabla_a^2 \mathcal{R}_{\infty,\rho}(a_0, h_0), \qquad \Sigma_0 = \mathrm{var}_\xi\{\nabla_a r_\xi(a_0, h_0)\}.$$

The covariance records variation across prediction tasks. It can be singular when the mask is deterministic.

Theorem 15. *Suppose Conditions 1 and 2 hold, m is fixed, $b = b_R \to \infty$, and*

$$\frac{Rm \log(m b_R)}{b_R} \to 0, \qquad \eta_R = o_{\mathbb{P}}(R^{-1}).$$

If $v^\top H_0 v \geq \kappa \|v\|_2^2$ for $v \in \mathrm{span}(\mathcal{C}_0)$, then

$$\mathbb{P}(\widehat{h} = h_0) \to 1, \qquad \sqrt{R}(\widehat{a} - a_0) \Longrightarrow \arg\min_{v \in \mathcal{C}_0} \left\{ Z_0^\top v + \frac{1}{2} v^\top H_0 v \right\},$$

where $Z_0 \sim N(0, \Sigma_0)$.

Theorem 15 gives inference for the regularized population metric and covers bounded-noise instances of Proposition 28. The rate condition makes the first two derivatives of the finite-block risk accurate on the $R^{-1/2}$ scale. Random masks yield a nondegenerate root-R limit when their population scores vary. With a deterministic view split, Σ_0 can equal zero and a separate observation-level normalization is needed.

For an interior a_0, put

$$V_Q = (Q^\top H_0 Q)^{-1} Q^\top \Sigma_0 Q (Q^\top H_0 Q)^{-1}.$$

Let $\widehat{V}_Q = \widehat{V}_{Q,b_R}$ use the empirical block scores and Hessian defined above. The increasing block size changes their population limits, but not their construction.

Corollary 15. *Under Theorem 15, suppose a_0 is interior. Then $\widehat{V}_Q \to V_Q$ in probability, and for every fixed $c \in \mathbf{1}^\perp$ with $c^\top Q V_Q Q^\top c > 0$,*

$$\frac{\sqrt{R}\, c^\top (\widehat{a} - a_0)}{\{c^\top Q \widehat{V}_Q Q^\top c\}^{1/2}} \Longrightarrow N(0,1).$$

Corollary 15 centers the contrast at a_0, not at a finite-block surrogate. At a boundary point, the plug-in simulation in Supplementary Materials consistently estimates the cone distribution in Theorem 15.

The preceding limit uses one random mask per block. We now average the masks and recover the first-order variation of the observations. Let $\mathcal{O}_i = X_i$ under coordinate masking and $\mathcal{O}_i = (X_i, V_i)$ under a prespecified response view. For one block, let

$$\overline{s}_b(a, h; \mathcal{O}_{1:b}) = \mathbb{E}_\xi\{\nabla_a \ell_b(a, h; \xi)\} + \frac{\rho}{m}(a - \mathbf{1}).$$

Supplementary Lemma 9 proves that the limit

$$\psi_0(o) = \lim_{b \to \infty} b\,[\mathbb{E}\{\overline{s}_b(a_0, h_0; \mathcal{O}_{1:b}) \mid \mathcal{O}_1 = o\} - \mathbb{E}\{\overline{s}_b(a_0, h_0; \mathcal{O}_{1:b})\}]$$

exists in $L^2(P)$. Put $\Omega_0 = \mathrm{var}\{\psi_0(\mathcal{O})\}$. Let $(\widehat{a}_{\mathrm{obs}}, \widehat{h}_{\mathrm{obs}})$ minimize the criterion that replaces each block loss by its mask expectation. A Monte Carlo average of J masks per block gives the same limit under the condition below.

Theorem 16. *Suppose Conditions 1 and 2 hold, m and the mask support are fixed, $b = b_R \to \infty$, and*

$$\frac{R(\log b_R)^2}{b_R} \to 0, \qquad \frac{b_R}{J_R} \to 0, \qquad \eta_R = o_{\mathbb{P}}\{(R b_R)^{-1}\}.$$

For an exact mask average, omit the condition involving J_R. If

$$v^\top H_0 v \geq \kappa \|v\|_2^2, \qquad v \in \mathrm{span}(\mathcal{C}_0),$$

then

$$\mathbb{P}(\widehat{h}_{\mathrm{obs}} = h_0) \to 1, \qquad \sqrt{R b_R}(\widehat{a}_{\mathrm{obs}} - a_0) \Longrightarrow \arg\min_{v \in \mathcal{C}_0} \left\{ Z_{\mathrm{obs}}^\top v + \frac{1}{2} v^\top H_0 v \right\},$$

where $Z_{\mathrm{obs}} \sim N(0, \Omega_0)$.

Theorem 16 gives population-metric inference at the total-observation rate $\sqrt{R b_R}$. Its covariance is generated by observations and can be positive under a deterministic view split. Random masks serve only as numerical integration; $J_R / b_R \to \infty$ removes their contribution from the first-order distribution. Interior contrasts use the sandwich covariance obtained by replacing Σ_0 with Ω_0, while Supplementary Proposition 42 gives a boundary distributional approximation.

The logarithmic factor in the block condition of Theorem 16 is not needed. The following corollary uses the Hilbert-space moment calculation already underlying the score expansion.

Corollary 16. *The conclusions of Theorem 16 hold if its block condition is replaced by $R/b_R \to 0$. All its other premises remain unchanged.*

Corollary 16 requires the finite-block bias to be negligible at the total-observation scale. It does not assert that this bias is zero. The covariance estimator in Proposition 34 and the spectral conclusion in Corollary 21 remain valid under this weaker block condition.

The observation covariance can be estimated without treating masks as independent observations. For an exact mask average, let $\widehat{s}_{r,b}$ be the block score at $(\widehat{a}_{\mathrm{obs}}, \widehat{h}_{\mathrm{obs}})$, let $\widehat{s}_{r,b}^{(-i)}$ be the corresponding score after observation $i \in I_r$ is removed, and put

$$\widehat{\psi}_{ri} = (b-1)\{\widehat{s}_{r,b} - \widehat{s}_{r,b}^{(-i)}\}, \qquad \overline{\psi} = \frac{1}{Rb}\sum_{r=1}^{R}\sum_{i\in I_r}\widehat{\psi}_{ri}.$$

Define

$$\widehat{\Omega}_0 = \frac{1}{Rb}\sum_{r=1}^{R}\sum_{i\in I_r}(\widehat{\psi}_{ri} - \overline{\psi})(\widehat{\psi}_{ri} - \overline{\psi})^\top.$$

The following proposition makes the observation-level covariance feasible. The implicit derivative in Supplementary Lemma 9 gives an equivalent calculation that avoids refitting every leave-one-out block.

Proposition 34. *Under Theorem 16, suppose the exact mask average is used. Then*

$$\widehat{\Omega}_0 \longrightarrow \Omega_0 \quad \textit{in probability}.$$

Proposition 34 supplies the sandwich covariance for interior contrasts and the Gaussian score covariance for boundary cone simulation. Its proof verifies the smooth-functional expansion used in classical jackknife theory (Shao & Wu, 1989) for the zero-diagonal bistochastic score. A deterministic scientific view is included as a one-point mask distribution.

Bounded responses are not needed for metric inference with a fixed prescribed observed view. The term involving only the squared responses is constant in the metric. Its derivative is zero. For the following extension, let $N = Rb$ and define ψ_0 directly from the population score. Let δ_o denote the point mass at the complete observation o. For $P_t = (1-t)P + t\delta_o$, put

$$\psi_0(o) = \frac{d}{dt}\nabla_a \mathcal{R}_{\infty,\rho}^{P_t}(a_0, h_0)\Big|_{t=0}, \qquad \Omega_0 = \mathrm{var}\{\psi_0(\mathcal{O})\}.$$

The superscript denotes the observation distribution used in the population risk. This definition agrees with the earlier one for bounded observations. It does not require finite-block conditional score projections to belong to $L^2(P)$.

Corollary 17. *For a prescribed response view with $\xi = \mathbf{1}$ and $\pi = 1$, suppose the observations are independent and identically distributed, Condition 2 holds, and m, q_V, and $\mathcal{H}$ are fixed. Assume the prediction-scale bounds in Condition 1 and suppose*

$$\|X\|_\infty \le B_X, \qquad \mathbb{E}\|V\|_2^2 < \infty, \qquad v^\top H_0 v \ge \kappa\|v\|_2^2 \quad \textit{for } v \in \mathrm{span}(\mathcal{C}_0).$$

If $b \to \infty$, $R/b \to 0$, and $\eta_R = o_{\mathbb{P}}(N^{-1})$, the conclusions of Theorem 16 and Proposition 34 hold with the preceding ψ_0 and Ω_0. The number of blocks $R \ge 1$ need not diverge.

Corollary 17 gives the total-observation limit and a consistent covariance estimator without a fourth response moment. The finite-block score remainder is $O_{\mathbb{P}}(b^{-1})$ and must be negligible at the $N^{-1/2}$ scale. The result permits a fixed number of increasing blocks.

The normalized prediction bandwidth also yields an increasing-dimensional statement. Fix $r_0 > 0$. For this result, risk separation is the infimum of $\mathcal{R}_{b,\rho}(a, h_{0,b}) - \mathcal{R}_{b,\rho}(a_{0,b}, h_{0,b})$ over feasible a with $\|a - a_{0,b}\|_2 \ge r_0$.

Theorem 17. *Suppose Condition 1 holds, b and $|\mathcal{H}|$ are bounded, $m = m_R$ increases, and the ridge sequence satisfies $\sup_R \rho_m < \infty$. Assume $(a_{0,b}, h_{0,b})$ is the unique finite-block minimizer, its bandwidth gap is bounded away from zero uniformly in R, its risk separation is at least $c\kappa_m$ for a fixed $c > 0$, and*

$$\mathcal{R}_{b,\rho}(a, h_{0,b}) - \mathcal{R}_{b,\rho}(a_{0,b}, h_{0,b}) \ge \kappa_m\|a - a_{0,b}\|_2^2$$

whenever $\|a - a_{0,b}\|_2 \le r_0$, *where* $0 < \kappa_m \le \overline{\kappa}$. *If*

$$\frac{m\log(mR)}{R} \to 0, \qquad \frac{1}{\kappa_m}\sqrt{\frac{m\log(mR)}{R}} \to 0, \qquad \frac{\eta_R}{\kappa_m} = o_{\mathbb{P}}(1),$$

then

$$\|\widehat{a} - a_{0,b}\|_2 = O_{\mathbb{P}}\left\{\frac{1}{\kappa_m}\sqrt{\frac{m\log(mR)}{R}} + \left(\frac{\eta_R}{\kappa_m}\right)^{1/2}\right\}.$$

Theorem 17 estimates a finite-block predictive target. The normalization of h makes the kernel and stochastic constants uniform in m; it does not supply population curvature. The factor κ_m^{-1} records the cost of weak metric identification. The risk separation may decrease at the same order as the curvature; the stochastic and optimization errors must be smaller. In a sparse regime with a fixed active-set size, raw active weights have order m and raw-coordinate curvature commonly has order m^{-2}. The normalized simplex parameter defined below gives the stable scale for this regime. The next result also lets b increase and targets the population smoother.

Fix a set S of size s and define

$$\Delta_S = \{\theta \in \mathbb{R}^m : \theta_j \ge 0,\ \mathbf{1}^\top\theta = 1,\ \theta_j = 0 \text{ for } j \notin S\}.$$

Let $\mathcal{T}$ be a fixed finite subset of $[\tau_-, \tau_+]$ and let $\mathcal{H}_m = \{\sqrt{m}\,\tau : \tau \in \mathcal{T}\}$. For $\rho_m = \bar{\rho}/m$, put

$$\widetilde{\mathcal{R}}_m(\theta, \tau) = \mathcal{R}_{\infty,\rho_m}\{\delta\mathbf{1} + m(1-\delta)\theta, \sqrt{m}\,\tau\}.$$

Let $(\theta_{0,m}, \tau_0)$ minimize this risk over $\Delta_S \times \mathcal{T}$. The constants in the following condition do not depend on m. We identify vectors supported on S with their s active coordinates and pad θ_∞ with zeros outside S.

Condition 4 (Sparse metric sequence)**.** *For fixed* s, $\bar{\rho}$, c_θ, κ_θ, c_d, χ, *and* $g_\tau > 0$,

$$\min_{j\in S}\theta_{0,m,j} \ge c_\theta, \qquad \theta_{0,m} \to \theta_\infty,$$

$$v^\top\nabla^2_\theta\widetilde{\mathcal{R}}_m(\theta,\tau)v \ge \kappa_\theta\|v\|_2^2, \quad \theta \in \Delta_S, \quad \tau \in \mathcal{T}, \quad \operatorname{supp}(v) \subset S, \quad \mathbf{1}^\top v = 0,$$

and

$$\inf_{\substack{\theta\in\Delta_S \\ \tau\neq\tau_0}}\{\widetilde{\mathcal{R}}_m(\theta,\tau) - \widetilde{\mathcal{R}}_m(\theta_{0,m},\tau_0)\} \ge g_\tau.$$

$$\min_{\tau\in\mathcal{T}}\min_{j\in S} d_j(\sqrt{m}\,\tau) \ge c_d, \qquad \max_{\tau\in\mathcal{T}}\max_{j\notin S} d_j(\sqrt{m}\,\tau) \le 0,$$

and

$$\max_{\tau\in\mathcal{T}}\max_{j\notin S}\max_{\ell\in S}\left[m\left\{\frac{\partial\mathcal{R}_{\infty,\rho_m}(\mathbf{1},\sqrt{m}\,\tau)}{\partial a_\ell} - \frac{\partial\mathcal{R}_{\infty,\rho_m}(\mathbf{1},\sqrt{m}\,\tau)}{\partial a_j}\right\} + (4C_H + 2\bar{\rho})(1-\delta)\right] \le -\chi.$$

Condition 4 states population separation on the scale of the simplex mass. Its curvature is imposed only on the fixed-dimensional active simplex. The score and exchange inequalities make support screening and the one-box certificate uniform over the finite bandwidth collection.

The condition can be checked through the one-dimensional risk curves in a prespecified-view product model. Let $\omega_j > 0$ be the fixed risk weight of signal coordinate j and put

$$\widetilde{r}_{j,m}(t,\tau) = \omega_j r_{j,m}\{\delta + m(1-\delta)t, \sqrt{m}\,\tau\}, \qquad 0 \le t \le 1.$$

For limits $\widetilde{r}_j$, define

$$\widetilde{F}_\tau(\theta) = \sum_{j\in S}\widetilde{r}_j(\theta_j,\tau) + \frac{\bar{\rho}(1-\delta)^2}{2}\sum_{j\in S}\theta_j^2, \qquad \theta \in \Delta_S.$$

Proposition 35. *Suppose the prespecified-view decomposition in Proposition 29 holds with fixed* $\omega_j > 0$, *while* s *and* $|\mathcal{T}|$ *are fixed, and*

$$\max_{j\in S}\max_{\tau\in\mathcal{T}}\|\widetilde{r}_{j,m}(\cdot,\tau) - \widetilde{r}_j(\cdot,\tau)\|_{C^2[0,1]} \to 0.$$

Assume (θ_∞, τ_0) *uniquely minimizes* $\widetilde{F}_\tau(\theta)$,

$$\min_{j\in S} \theta_{\infty,j} > 0, \qquad \inf_{\substack{\theta\in\Delta_S \\ \tau\neq\tau_0}} \{\widetilde{F}_\tau(\theta) - \widetilde{F}_{\tau_0}(\theta_\infty)\} > 0, \qquad \inf_{\substack{\theta\in\Delta_S,\ \tau\in\mathcal{T} \\ v\in\mathbf{1}^\perp,\ \mathrm{supp}(v)\subset S,\ v\neq 0}} \frac{v^\top \nabla^2 \widetilde{F}_\tau(\theta) v}{\|v\|_2^2} > 0,$$

and

$$\max_{j\in S} \max_{\tau\in\mathcal{T}} \frac{\partial_t \widetilde{r}_j(0,\tau)}{1-\delta} + (4C_H + 2\bar{\rho})(1-\delta) < 0,$$

then Condition 4 holds for all sufficiently large m.

Proposition 35 reduces the increasing-dimensional assumptions to convergence, curvature, and first derivatives of one-dimensional prediction-risk curves. The next bounded model verifies these properties and Condition 1 at the same time. The Gaussian model in Proposition 30 remains an exact fixed-dimensional comparison but is not used to justify the bounded-data theorem.

Let $Z_1, \ldots, Z_s, \varepsilon_1, \ldots, \varepsilon_s$, and $N_1, \ldots, N_{m-s}$ be independent Rademacher variables, and let $V_j = Z_j + \sigma\varepsilon_j$. The observed view is $(Z_1, \ldots, Z_s, N_1, \ldots, N_{m-s})$, while $V = (V_1, \ldots, V_s)$ is the response view. Use risk weights s^{-1}, $h = \sqrt{m}\,\tau$, and $\rho_m = \bar{\rho}/m$. For $0 \le t \le 1$, put

$$g_{\tau,m}(t) = \frac{1}{s}\Big[\Big\{1 - \tanh\Big(\frac{\delta + m(1-\delta)t}{m\tau^2}\Big)\Big\}^2 + \sigma^2\Big]$$

and

$$g_\tau(t) = \frac{1}{s}\Big[\Big\{1 - \tanh\Big(\frac{(1-\delta)t}{\tau^2}\Big)\Big\}^2 + \sigma^2\Big].$$

The following corollary gives a fully specified increasing-dimensional distribution for the sparse result.

Corollary 18. *Suppose* s *and* $\mathcal{T}$ *are fixed and*

$$\frac{2}{s(\max\mathcal{T})^2} > (4C_H + 2\bar{\rho})(1-\delta).$$

Then

$$\max_{\tau\in\mathcal{T}} \|g_{\tau,m} - g_\tau\|_{C^2[0,1]} \to 0,$$

and Conditions 1 and 4 hold with $\theta_{0,m} = \theta_\infty = s^{-1}\mathbf{1}_S$ *and* $\tau_0 = \min\mathcal{T}$ *for all sufficiently large* m.

Corollary 18 verifies the population premises of Theorem 18 by elementary formulas. The inequality holds, for example, with $\mathcal{T} = \{\tau\}$ and $\bar{\rho} = (2s\tau^2)^{-1}$ for sufficiently large fixed τ, because the proof of Lemma 4 gives $C_H = O(\tau^{-4})$. The remaining restrictions in that theorem concern the numbers of blocks, observations per block, and the certified optimization tolerance.

Uniform coordinate masks behave differently when s stays fixed. The next proposition quantifies this distinction and explains why a scientific response view is needed for the sparse increasing-dimensional result. Write $a_S(\theta) = \delta\mathbf{1}_s + m(1-\delta)\theta_S$.

Proposition 36. *Suppose* s *is fixed,* $h = \sqrt{m}\,\tau$*, and the first two derivatives of* $r_j\{\delta + m(1-\delta)t, h\}$ *with respect to* t *are uniformly bounded. For the paired mask or a uniform fixed-count mask with* k/m *bounded away from zero and one,*

$$\sup_{\theta\in\Delta_S} |Q_h\{a_S(\theta)\} - Q_h(\delta\mathbf{1}_s)| + \sup_{\theta\in\Delta_S} \|\nabla_\theta Q_h\{a_S(\theta)\}\|_2 + \sup_{\theta\in\Delta_S} \|\nabla^2_\theta Q_h\{a_S(\theta)\}\|_{\mathrm{op}} \le C/m,$$

and

$$\max_{j\in S} |d_j(h)| \le C/m.$$

If $\rho_m = \bar{\rho}/m$*, the ridge Hessian in* θ_S *converges to* $\bar{\rho}(1-\delta)^2 I_s$.

Proposition 36 shows that fixed-count masks dilute a fixed number of predictive coordinates because the loss averages over an increasing number of held-out features. A ridge of order m^{-1} then dominates their aggregate prediction risk. This is a property of the estimand, not a proof artifact.

The sparse result uses a prespecified response view with fixed risk weights. Under uniform coordinate masks, a nondegenerate sparse limit requires a smaller ridge and sample sizes that resolve an order-m^{-1} risk difference.

For a deterministic sequence $\bar{\eta}_R \to 0$, put

$$e_{R,b,m} = \sqrt{\frac{\log(m|\mathcal{T}|)}{R}} + \sqrt{\frac{m\log(mb)}{b}}, \qquad r_{R,b,m} = R^{-1/2} + \sqrt{\frac{m\log(mb)}{b}} + \bar{\eta}_R^{1/2}.$$

The following theorem establishes support recovery, a population-metric rate, and certified computation in one increasing-dimensional regime. Here $\widehat{a}$ is the certified restricted fit: for each bandwidth, coordinates outside the certified candidate set are fixed at δ, and the criterion is minimized on the resulting face to tolerance η_R. If the certificate fails, the full feasible set is used. Let $L_{\theta,R}$ be a certified Lipschitz constant for the restricted empirical gradient on Δ_S.

Theorem 18. *Suppose Conditions 1 and 4 hold, $\eta_R = O_{\mathbb{P}}(\bar{\eta}_R)$, and*

$$\frac{\log(m|\mathcal{T}|)}{R} \to 0, \qquad \frac{m\log(mb)}{b} \to 0, \qquad e_{R,b,m} = o(\lambda_R), \qquad \lambda_R \to 0.$$

Then, with probability tending to one, the score screen and the certificate in Proposition 18 select S for every $h \in \mathcal{H}_m$, and $\widehat{h} = \sqrt{m}\,\tau_0$. For

$$\widehat{\theta} = \frac{\widehat{a} - \delta\mathbf{1}}{m(1-\delta)},$$

it holds that

$$\|\widehat{\theta} - \theta_{0,m}\|_2 = O_{\mathbb{P}}(r_{R,b,m}).$$

For every prespecified $t_N \in (0, 1-\delta)$, $\mathbb{P}(\widehat{S} = S) \to 1$, and

$$\|\widehat{a}^{\mathrm{rep}} - s\theta_\infty\|_2 = O_{\mathbb{P}}\{r_{R,b,m} + m^{-1} + \|\theta_{0,m} - \theta_\infty\|_2\}.$$

With probability tending to one, the restricted empirical criterion has tangent curvature at least $\kappa_\theta/2$. For $0 < \eta_R < C$, projected gradient descent reaches its global fixed-bandwidth minimum within objective tolerance η_R after at most $C\frac{L_{\theta,R}}{\kappa_\theta}\log(C/\eta_R)$ iterations. With the evaluation accuracy ϵ defined before Proposition 17, chosen sufficiently small relative to η_R and with $\epsilon = o(\lambda_R)$, its total operation count is at most $C\log(b/\epsilon)$ times

$$O\Big[|\mathcal{T}|R(b^3 + b^2 m)\Big\{1 + \frac{L_{\theta,R}}{\kappa_\theta}\log(C/\eta_R)\Big\} + |\mathcal{T}|m^2\Big].$$

Theorem 18 avoids division by raw-coordinate curvature. The statistical screening term depends logarithmically on the full feature dimension, while the one-box calculation costs $O(|\mathcal{T}|m^2)$. The restricted tangent-Hessian certificate removes branch and bound in this regime and reduces global optimization to linearly convergent projected gradient steps. The representation metric has a fixed positive limit on S, which permits a spectral comparison on one compact metric set independent of m.

The support S_0, separation β_0, and population representation metric are defined in Section 4. The following corollary transfers simultaneous metric error to the selected distance.

Corollary 19. *Suppose $|S_0|$ is fixed, $\|\widehat{a} - a_0\|_\infty = O_{\mathbb{P}}(r_N)$, $r_N = o(t_N)$, and $t_N = o(\beta_0)$. Then*

$$\mathbb{P}(\widehat{S} = S_0) \to 1, \qquad \|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2 = O_{\mathbb{P}}(r_N).$$

Corollary 19 turns attenuation at the prediction boundary into exact removal from the representation distance. Proposition 29 gives this support under heterogeneous product distributions. In Proposition 30, every independent nuisance coordinate lies at the boundary and $\beta_0 = a_S - \delta > 0$.

Under Condition 3, the selected observation map defines the continuum operator in (11). The following proposition controls its eigenspace under metric perturbation.

Proposition 37. *Suppose Condition 3 holds and $\gamma_p > 0$. On a neighborhood of a_0^{rep},*

$$\max_{1\le j\le p+2} |\lambda_j(w) - \lambda_j(a_0^{\mathrm{rep}})| \le C_p\|w - a_0^{\mathrm{rep}}\|_2$$

and

$$\|\Pi_{\mathcal{E}_p(w)} - \Pi_{\mathcal{E}_p(a_0^{\mathrm{rep}})}\|_{L^2(P)\to L^2(P)} \le \frac{C_p}{\gamma_p}\|w - a_0^{\mathrm{rep}}\|_2.$$

Proposition 37 transfers uncertainty in the selected feature weights to the continuum operator. Its bound has no inverse graph radius.

Let α_d be the volume of the unit ball in $\mathbb{R}^d$. For a metric w supported on S_0, let $A_{n,k}(w)$ and $D_{n,k}(w)$ be the symmetric nearest-neighbor affinity and its degree matrix defined in Section 3. The scaled graph Laplacian is

$$L_{n,k}(w) = \frac{1}{n}\left(\frac{n\alpha_d}{k_n}\right)^{1+2/d}\{D_{n,k}(w) - A_{n,k}(w)\}.$$

The scalar does not change its eigenvectors. Let $\widehat{\lambda}_j$ be the eigenvalues of $L_{n,k}(\widehat{a}^{\mathrm{rep}})$. The following theorem compares the learned graph with the population eigenfunction evaluations Φ_0 defined in Section 4.

Theorem 19. *Suppose Condition 3 holds, $\mathbb{P}(\widehat{S} = S_0) \to 1$, $\widehat{a}^{\mathrm{rep}} \to a_0^{\mathrm{rep}}$ in probability, and $\gamma_p > 0$. The observations used to estimate the metric may also form the representation graph. If*

$$k_n/n \to 0, \qquad \frac{k_n^{1+4/d}}{n^{4/d}\log n} \to \infty,$$

then

$$\max_{1\le j\le p+2} |\widehat{\lambda}_j - \lambda_j(a_0^{\mathrm{rep}})| = O_{\mathbb{P}}\left\{\left(\frac{k_n}{n}\right)^{1/d} + \|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2\right\}$$

and

$$\inf_{O^\top O = I_p}\left\|\widehat{Y} - n^{-1/2}\Phi_0 O\right\|_F = O_{\mathbb{P}}\left\{\frac{(k_n/n)^{1/d} + \|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2}{\gamma_p}\right\}.$$

Theorem 19 proves convergence of the selected representation, including eigenvalue multiplicities within the selected space. Supplementary Lemma 13 establishes one simultaneous spectral event over the compact metric family. The event contains a metric estimated from the same observations and removes the need for sample splitting. The rate condition permits the finite covering probability and the graph perturbation between covering points. The fixed-metric spectral bound is due to Calder & García Trillos (2022). Related variational limits for balanced graph cuts are established by García Trillos (2019).

For the sparse increasing-dimensional sequence, put $w_\infty = s\theta_\infty$ and let Φ_∞ evaluate an orthonormal basis of $\mathcal{E}_p(w_\infty)$ at the representation observations. Define

$$\gamma_{p,\infty} = \lambda_2(w_\infty) \wedge \{\lambda_{p+2}(w_\infty) - \lambda_{p+1}(w_\infty)\}.$$

The following corollary connects the primitive increasing-dimensional metric result to the final spectral coordinates.

Corollary 20. *Suppose Theorem 18 and Condition 3 hold with $S_0 = S$, and $\gamma_{p,\infty} > 0$. If k_n satisfies Theorem 19, then*

$$\max_{1\le j\le p+2} |\widehat{\lambda}_j - \lambda_j(w_\infty)| = O_{\mathbb{P}}\left\{\left(\frac{k_n}{n}\right)^{1/d} + r_{R,b,m} + m^{-1} + \|\theta_{0,m} - \theta_\infty\|_2\right\},$$

and

$$\inf_{O^\top O = I_p}\|\widehat{Y} - n^{-1/2}\Phi_\infty O\|_F = O_{\mathbb{P}}\left[\frac{(k_n/n)^{1/d} + r_{R,b,m} + m^{-1} + \|\theta_{0,m} - \theta_\infty\|_2}{\gamma_{p,\infty}}\right].$$

The representation observations may equal the metric observations.

Corollary 20 formally yields a limiting population Laplace-type operator equipped with a fixed, strictly positive metric precisely on the selected coordinates. While the increasing number of nuisance coordinates actively affects both the empirical screening step and the initial finite-block approximation, it mathematically does not enlarge the true intrinsic dimension of this final limiting differential operator.

A denoised representation may replace the clean selected coordinates when its error decreases. Suppose

$$\widetilde{X}_{i,S_0} = \phi_{S_0}(Z_i) + e_i,$$

where, conditionally on Z_i,

$$\mathbb{P}(\|e_i\|_2 > t \mid Z_i) \le 2\exp\{-t^2/(2\sigma_n^2)\}.$$

Put $r_n = (k_n/n)^{1/d}$.

Proposition 38. *Suppose the assumptions of Theorem 19 hold, and construct the graph from $\widetilde{X}_{i,S_0}$. If*

$$\frac{\sigma_n\sqrt{\log n}}{r_n^3} \to 0,$$

then the eigenvalue bound in Theorem 19 has the additional term $O_{\mathbb{P}}\left(\frac{\sigma_n\sqrt{\log n}}{r_n^3}\right)$, and the eigenspace bound has this term divided by γ_p.

Proposition 38 permits measurement error that tends to zero or a preliminary cross-view denoiser whose uniform error has the stated order. A fixed nonzero error level does not satisfy this premise and is not treated as a point on the signal manifold.

The following proposition gives two directly implementable neighborhood orders.

Proposition 39. *For fixed $d \ge 2$, let $\ell_n \to \infty$ increase slowly enough that $k_n/n \to 0$. Both*

$$k_n = \left\lceil n^{4/(d+4)}(\log n)^{d/(d+4)}\ell_n \right\rceil$$

and the dimension-free choice $k_n = \lceil n^{2/3}(\log n)^{1/3}\ell_n \rceil$ satisfy the conditions of Theorem 19. The first gives graph error

$$O_{\mathbb{P}}\{n^{-1/(d+4)}(\log n)^{1/(d+4)}\ell_n^{1/d}\}.$$

Proposition 39 gives sparse nearest-neighbor orders for the final representation graph. The two choices agree when $d = 2$. The dimension-free choice uses more neighbors when $d > 2$, but it remains sparse because $k_n/n \to 0$.

Corollary 1 gives an explicit allocation across blocks, masks, optimization accuracy, support separation, and graph size. It applies to a representation graph formed from the metric observations. With noisy selected coordinates, Proposition 38 adds its explicit perturbation term. Independent nuisance coordinates outside S_0 do not enter either bound.

Mask averaging improves the metric term when the block criterion is smooth at the population target. Apply (8) to $\widehat{a}_{\mathrm{obs}}$ and denote the resulting support, metric, and embedding by $\widehat{S}_{\mathrm{obs}}$, $\widehat{a}^{\mathrm{rep}}_{\mathrm{obs}}$, and $\widehat{Y}_{\mathrm{obs}}$.

Corollary 21. *Suppose Theorem 16 and Condition 3 hold and $\gamma_p > 0$. If*

$$(Rb_R)^{-1/2} = o(t_{Rb_R}), \qquad t_{Rb_R} = o(\beta_0),$$

and k_n satisfies Theorem 19, then

$$\mathbb{P}(\widehat{S}_{\mathrm{obs}} = S_0) \to 1, \qquad \|\widehat{a}^{\mathrm{rep}}_{\mathrm{obs}} - a_0^{\mathrm{rep}}\|_2 = O_{\mathbb{P}}\{(Rb_R)^{-1/2}\},$$

and

$$\inf_{O^\top O = I_p} \left\| \widehat{Y}_{\mathrm{obs}} - n^{-1/2}\Phi_0 O \right\|_F = O_{\mathbb{P}}\left[\frac{(k_n/n)^{1/d} + (Rb_R)^{-1/2}}{\gamma_p}\right].$$

The representation observations may be the Rb_R metric observations, in which case $n = Rb_R$.

Corollary 21 transfers the observation-level rate to support recovery and the final spectral coordinates. It removes both a root-R metric term and an independent representation sample from this regime.

H AUXILIARY STATISTICAL AND SPECTRAL PROOFS

Rigorous uniform prediction-kernel bounds actively control the empirical scaling process and successfully govern the constrained estimator. Subsequently, the formal graph arguments establish a simultaneous spectral approximation valid over the compact metric family on the selected coordinates, comprehensively covering metrics strictly fitted directly to the original representation observations.

For one block, let ζ contain the complete observations and their mask, and write

$$q_b(a, h; \zeta) = \ell_b(a, h; \xi) + \frac{\rho}{2m}\|a - \mathbf{1}\|_2^2.$$

The following lemma supplies the regularity used in the finite-block and increasing-dimensional results.

Lemma 4. *Under Condition 1, every off-diagonal prediction-kernel entry satisfies*

$$k_- \le K_{ij}(a, h; \xi) \le 1, \qquad k_- = \exp\{-2B^2/(\pi_-\tau_-^2)\}.$$

There are constants $0 < c_u < C_u < \infty$ and $C_s < \infty$ that do not depend on b or m such that

$$c_u b^{-1/2} \le u_i \le C_u b^{-1/2}, \qquad \|(I_b + W)^{-1}\|_{\mathrm{op}} \le C_s.$$

The first three directional derivatives of q_b are bounded by constants that do not depend on b or m, and q_b is uniformly Lipschitz in the ℓ_∞ norm. For $\|v\|_2 = 1$,

$$\max_{i,j} |\dot\eta_{ij}(v)| \le \frac{2B^2}{\pi_-\tau_-^2\sqrt{m}}.$$

For $r = 1, 2, 3$ and coordinate indices $j_1, \dots, j_r$,

$$\left|\frac{\partial^r q_b}{\partial a_{j_1} \cdots \partial a_{j_r}}\right| \le \frac{C_r}{m^r} + \mathbf{1}\{r = 1\}\frac{\rho|a_{j_1} - 1|}{m} + \mathbf{1}\{r = 2,\ j_1 = j_2\}\frac{\rho}{m},$$

where C_r does not depend on b or m. We write $C_H = C_2$ for the Hessian constant.

Lemma 4 identifies the role of the normalized prediction bandwidth. The kernel lower bound no longer decreases exponentially with m, but its constants depend on B, π_-, and τ_-. At $\pi_- = 1/2$ and $\tau_- = 0.6$, the bounds $B = 3$ and $B = \sqrt{3}$ for the surface and torus designs give $k_- = e^{-100}$ and $e^{-100/3}$, respectively. These uniform constants describe asymptotic control and need not give useful numerical bounds at the experimental sample sizes.

Proof of Lemma 4. Condition 1 and $\mathbf{1}^\top a = m$ give

$$0 \le d_{a,\xi}^2(X_i, X_j) \le \frac{4B^2 m}{\pi_-}.$$

Because $h^2 \ge m\tau_-^2$, the kernel bound follows. Put $M = u_{\max}$ and $s = u_{\min}$, and let S be the sum of the other $b - 2$ scaling factors. The row equations at the corresponding indices give

$$\frac{1}{M} \ge k_-(s + S), \qquad \frac{1}{s} \le M + S.$$

For $r = M/s$ and $t = S/s \ge b - 2$,

$$r \le \frac{r + t}{k_-(1 + t)}.$$

For $k_-(b - 1) > 1$, the inequality bounds r in terms of k_-. For the finitely many remaining b, continuity of scaling on symmetric kernels with off-diagonal entries in $[k_-, 1]$ gives the same bound. The row equation at the smallest factor gives

$$\{C(b - 1)\}^{-1/2} \le s \le \{k_-(b - 1)\}^{-1/2}.$$

The ratio bound proves $c_u b^{-1/2} \le u_i \le C_u b^{-1/2}$. Hence

$$W_{ij} \ge c_W/(b - 1), \qquad i \ne j,$$

for $c_W > 0$ depending only on k_-. Write

$$W = c_W \frac{J_b - I_b}{b-1} + (1 - c_W)Q,$$

where Q is symmetric and stochastic. Because $\lambda_{\min}\{(J_b - I_b)/(b-1)\} = -(b-1)^{-1}$ and $\lambda_{\min}(Q) \geq -1$,

$$\lambda_{\min}(I_b + W) \geq c_W\{1 - (b-1)^{-1}\}.$$

For the maximum-norm bound, $z = (I_b + W)^{-1}f$, $z = f - Wz$, and $W_{ij} \leq C_u^2/b$ give

$$\|z\|_\infty \leq \|f\|_\infty + \frac{C_u^2}{\sqrt{b}}\|z\|_2 \leq (1 + C_u^2 C_s)\|f\|_\infty.$$

Here C_s is the preceding Euclidean inverse bound.

For a metric direction v,

$$|\dot{\eta}_{ij}(v)| \leq \frac{2B^2}{\pi_- m \tau_-^2}\|v\|_1 \leq \frac{2B^2}{\pi_- \tau_-^2 \sqrt{m}}.$$

Equations (20) and (21), the stochastic row sums, and the inverse bound give the first derivative bounds. Differentiating the same equations twice more gives the second and third derivative bounds. Every residual in (5) is bounded by $2B$, and its normalization by b completes the loss calculation. For coordinate derivatives, put $z = \log u$, $z_j = \partial_j z$, $z_{jk} = \partial_j \partial_k z$, and $c_j = \partial_j \log K$ off the diagonal. Then $\|c_j\|_\infty \leq C/m$ and the higher log-kernel derivatives are zero. Define the matrices G_j and H_{jk} by $(G_j)_{ir} = (c_j)_{ir} + (z_j)_i + (z_j)_r$ and $(H_{jk})_{ir} = (z_{jk})_i + (z_{jk})_r$. Entrywise products are denoted by $\circ$. Differentiating $W\mathbf{1} = \mathbf{1}$ gives

$$(I_b + W)z_j = -(W \circ c_j)\mathbf{1}, \quad (I_b + W)z_{jk} = -(W \circ G_j \circ G_k)\mathbf{1},$$
$$(I_b + W)z_{jkl} = -\{W \circ (G_j \circ G_k \circ G_l + G_j \circ H_{kl} + G_k \circ H_{jl} + G_l \circ H_{jk})\}\mathbf{1}.$$

The rigorous maximum-norm inverse bound, when combined with the strictly stochastic row sums, cleanly implies that $\|z_j\|_\infty \leq C/m$, $\|z_{jk}\|_\infty \leq C/m^2$, and $\|z_{jkl}\|_\infty \leq C/m^3$. Consequently, the corresponding first three coordinate derivatives of the scaled kernel W explicitly maintain absolute row sums strictly bounded by these identical orders. Because the responses are structurally bounded, using the normalization of the block prediction loss thus mathematically yields

$$\max_{j_1,\ldots,j_r} \left| \frac{\partial^r \ell_b}{\partial a_{j_1} \cdots \partial a_{j_r}} \right| \leq C_r m^{-r}.$$

For fixed B and π_-, the scaling and inverse bounds remain bounded as $\tau_- \to \infty$. Each derivative of order r contains r log-kernel derivative factors, each bounded by $2B^2/(\pi_- m \tau_-^2)$. Thus the unpenalized derivative constants can be chosen as $C_r = O(\tau_-^{-2r})$ for $r = 1, 2, 3$. The ridge contributes $\rho(a_j - 1)/m$ to the gradient, ρ/m to a diagonal Hessian entry, and zero to the third derivative. This proves the coordinate bounds. Summing the first-coordinate derivatives over coordinate increments proves the stated ℓ_∞ Lipschitz bound. This completes the proof. □

Proof of Proposition 21. Write $X_{\xi^c} - g = X_{\xi^c} - m_\xi + m_\xi - g$. The cross term has expectation zero because

$$\mathbb{E}(X_{\xi^c} - m_\xi \mid X_\xi) = 0.$$

This proves the first assertion.

Let (X'_ξ, X'_{ξ^c}) be an independent observation used by the population smoother. Conditional expectation gives

$$\mathbb{E}\{X'_{\xi^c} \mid X'_\xi = x'\} = m_\xi(x').$$

The kernel and both scaling factors depend only on X_ξ and X'_ξ. Integrating first over X'_{ξ^c} gives $T_{a,h,\xi} m_\xi(X_\xi)$. This completes the proof. □

Proof of Proposition 22. The derivative of the aggregate map satisfies

$$D\Psi(z)^\top D\Psi(z) = G_{\text{pred}}(z).$$

Because the aggregate map Ψ strictly forms a valid immersion, this resulting structural matrix remains strictly positive definite exactly on the local tangent space. Furthermore, both its smallest and largest associated tangent eigenvalues are entirely continuous in the base coordinate z. Applying the global compactness of $\mathcal{M}$ immediately guarantees strictly positive, finite uniform bounds.

For a minimizing geodesic γ from z to z', integration of $D\Psi\{\gamma(t)\}\dot{\gamma}(t)$ gives the upper Lipschitz bound. The lower bound holds locally by the uniform smallest singular value. Outside a fixed geodesic neighborhood of the diagonal, the ratio $\frac{\|\Psi(z)-\Psi(z')\|_2}{d_{\mathcal{M}}(z,z')}$ is continuous and positive because Ψ is one-to-one. Compactness gives a positive lower bound. This completes the proof. □

Proof of Proposition 23. Use $g_{\mathbf{1}}$ to define the unit tangent bundle

$$\mathbb{S}\mathcal{M} = \{(z, v) : z \in \mathcal{M},\ v \in T_z\mathcal{M},\ v^\top g_{\mathbf{1}}(z)v = 1\}.$$

This set and $\mathcal{A}_\delta \times \mathbb{S}\mathcal{M}$ are compact. The two functions

$$(a, z, v) \longmapsto v^\top g_a(z)v, \qquad (z, v) \longmapsto v^\top G_{\mathrm{pred}}(z)v$$

are continuous and strictly positive. The ratio of the first function to the second has positive finite minimum and maximum on $\mathcal{A}_\delta \times \mathbb{S}\mathcal{M}$. Homogeneity in v proves the quadratic-form inequalities.

For every piecewise continuously differentiable curve γ, the inequalities give

$$c_G^{1/2} \operatorname{len}_{G_{\mathrm{pred}}}(\gamma) \le \operatorname{len}_{g_a}(\gamma) \le C_G^{1/2} \operatorname{len}_{G_{\mathrm{pred}}}(\gamma).$$

Taking the infimum over curves joining two points proves the distance bounds. This completes the proof. □

The local risk expansion requires control of the nonlinear population scaling equation. The next lemma records the scaling correction, the inverse bound, and the operator remainder used in Theorem 9.

Lemma 5. *Under the assumptions of Theorem 9, fix a mask and put $\varepsilon = h^2$, $q_{a,\xi} = p_{a,\xi}^{1/2}$, and*

$$r_{a,\varepsilon,\xi}(z) = (2\pi\varepsilon)^{d/4} u_{a,h,\xi}\{\phi_\xi(z)\}.$$

There are functions $E_{a,\xi}$ and

$$s_{a,\xi} = -\frac{\Delta_{g_{a,\xi}} q_{a,\xi} + E_{a,\xi} q_{a,\xi}}{4 q_{a,\xi}}$$

such that, for $1 \le \nu < 4/3$,

$$\sup_{a \in \mathcal{A}_\delta} \left\| r_{a,\varepsilon,\xi} - q_{a,\xi}^{-1}(1 + \varepsilon s_{a,\xi}) \right\|_{L^\nu(P)} \le C_\nu \varepsilon^2.$$

For every scalar C^6 function f,

$$\sup_{a \in \mathcal{A}_\delta} \left\| T_{a,h,\xi} f - f - \frac{\varepsilon}{2} \left\{ \Delta_{g_{a,\xi}} f + \langle \nabla_{g_{a,\xi}} \log p_{a,\xi}, \nabla_{g_{a,\xi}} f \rangle_{g_{a,\xi}} \right\} \right\|_{L^2(P)} \le C_f \varepsilon^2.$$

At the exact log scaling, the derivative of the log row equation is $I + T_{a,h,\xi}$ and

$$\|(I + T_{a,h,\xi})^{-1}\|_{L^2(P) \to L^2(P)} \le 1.$$

All constants are uniform over the finite mask support.

Lemma 5 shows that the density and extrinsic-curvature terms in the scaling function are fixed by the row equation. The operator expansion includes the density drift and removes the scalar curvature term. The inverse bound does not deteriorate as h decreases.

Proof of Lemma 5. For fixed a and ξ, define

$$\mathcal{G}_{\varepsilon,a,\xi} g(z) = (2\pi\varepsilon)^{-d/2} \int_{\mathcal{M}} \exp\left\{ -\frac{d_{a,\xi}^2(z, z')}{2\varepsilon} \right\} g(z')\, dV_{g_{a,\xi}}(z').$$

Normal coordinates and a fourth-order Gaussian expansion give

$$\mathcal{G}_{\varepsilon,a,\xi}g = g + \frac{\varepsilon}{2}\{\Delta_{g_{a,\xi}}g + E_{a,\xi}g\} + \varepsilon^2\mathcal{R}_{\varepsilon,a,\xi}g,$$

where $\|\mathcal{R}_{\varepsilon,a,\xi}g\|_\infty$ is bounded by a constant times $\|g\|_{C^6}$. For fixed m, compactness of the C^6 embedding and C^5 metric families gives uniform derivative, ellipticity, and density bounds in a. Finite mask support gives uniformity in ξ.

The normalized scaling equation is

$$r_{a,\varepsilon,\xi}\mathcal{G}_{\varepsilon,a,\xi}(r_{a,\varepsilon,\xi}p_{a,\xi}) = 1.$$

Substitute $\widetilde{r} = q_{a,\xi}^{-1}(1+\varepsilon s)$ into its left side. The preceding expansion gives

$$\widetilde{r}\mathcal{G}_{\varepsilon,a,\xi}(\widetilde{r}p_{a,\xi}) = 1 + \varepsilon\left\{2s + \frac{\Delta_{g_{a,\xi}}q_{a,\xi} + E_{a,\xi}q_{a,\xi}}{2q_{a,\xi}}\right\} + O(\varepsilon^2).$$

The stated $s_{a,\xi}$ makes the residual $\delta_{a,\varepsilon,\xi} = \widetilde{r}\mathcal{G}_{\varepsilon,a,\xi}(\widetilde{r}p_{a,\xi}) - 1$ uniformly $O(\varepsilon^2)$ in sup norm. Let $r_{a,\varepsilon,\xi} = \widetilde{r}(1+e)$ denote the exact positive scaling and define

$$\mathcal{A}_{a,\varepsilon,\xi}g = \widetilde{r}\mathcal{G}_{\varepsilon,a,\xi}(\widetilde{r}p_{a,\xi}g).$$

The exact and approximate row equations give

$$-\delta_{a,\varepsilon,\xi} = \mathcal{A}_{a,\varepsilon,\xi}e + \frac{e}{1+e}.$$

The Gaussian kernel is positive semidefinite, and hence $\langle e, \mathcal{A}_{a,\varepsilon,\xi}e\rangle_{L^2(P)} \geq 0$. The candidate $\widetilde{r}$ is bounded above and away from zero. Multiplication by e and integration yield

$$\int \frac{e^2}{r_{a,\varepsilon,\xi}}\,dP \leq C\varepsilon^2\int |e|\,dP.$$

Lemma SM1.3 of Landa & Cheng (2023), applied with bandwidth 2ε, gives

$$P\{r_{a,\varepsilon,\xi} > t\} \leq Ct^{-2}, \qquad t \geq 1.$$

The geometric and density constants in this bound are uniform over the compact metric family and the finite mask support. For $3/2 < \ell \leq 2$, Hölder's inequality and the tail bound give

$$\left(\int |e|^{2/\ell}\,dP\right)^{\ell} \leq \left(\int \frac{e^2}{r_{a,\varepsilon,\xi}}\,dP\right)\left(\int r_{a,\varepsilon,\xi}^{1/(\ell-1)}\,dP\right)^{\ell-1} \leq C\varepsilon^2\int |e|\,dP.$$

For $\nu = 2/\ell \in [1, 4/3)$, this inequality and the uniform bounds for $\widetilde{r}$ prove the first assertion.

For the inverse calculation, define the log row map

$$\mathfrak{F}(v)(z) = v(z) + \log\int k_{a,h,\xi}(z,z')e^{v(z')}\,dP(z').$$

At $v = \log u_{a,h,\xi}$, differentiation gives

$$D\mathfrak{F}(v)[g] = g + T_{a,h,\xi}g.$$

Positive scaling preserves Gaussian positive semidefiniteness. Thus $T_{a,h,\xi}$ is a positive self-adjoint contraction on $L^2(P)$, and

$$\langle g, (I + T_{a,h,\xi})g\rangle_{L^2(P)} \geq \|g\|^2_{L^2(P)},$$

which proves the inverse bound.

Proposition 3 in the arXiv version arXiv:1711.06711v2 of Marshall & Coifman (2019), with Gaussian profile $\exp(-t/2)$ and reference weight equal to one, gives the following expansion after translating their positive-Laplacian convention to the sign used here:

$$T_{a,h,\xi}f = f + \frac{\varepsilon}{2q_{a,\xi}}\{\Delta_{g_{a,\xi}}(q_{a,\xi}f) - f\Delta_{g_{a,\xi}}q_{a,\xi}\} + O_{L^2(P)}(\varepsilon^2).$$

The same compactness argument makes this remainder uniform. The product rule and $2\nabla q_{a,\xi}/q_{a,\xi} = \nabla\log p_{a,\xi}$ give the asserted operator expansion. This completes the proof. □

Proof of Theorem 9. Lemma 5, applied coordinatewise, gives

$$T_{a,h,\xi} f = f + \frac{h^2}{2}\left\{\Delta_{g_{a,\xi}} f + \langle \nabla_{g_{a,\xi}} \log p_{a,\xi}, \nabla_{g_{a,\xi}} f\rangle_{g_{a,\xi}}\right\} + O(h^4) = f - h^2 \mathcal{B}_{a,\xi} f + O(h^4)$$

in $L^2(P)$, uniformly in a and in the mask.

Proposition 21 and $m_\xi\{\phi_\xi(z)\} = \psi_\xi(z)$ give

$$\mathcal{R}_{\infty,0}(a,h) - \mathcal{V} = \mathbb{E}_\xi\left[\frac{1}{|\xi^c|}\|\psi_\xi - T_{a,h,\xi}\psi_\xi\|^2_{L^2(P)}\right] = h^4 \mathcal{J}(a) + O(h^6)$$

uniformly over $a \in \mathcal{A}_\delta$. This proves the first assertion. After division by h^4, the penalized criteria converge uniformly on the compact set $\mathcal{A}_\delta$ to

$$\mathcal{J}(a) + \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2.$$

The argmin theorem (van der Vaart, 1998) gives the accumulation-point assertion and convergence under uniqueness. This completes the proof. □

Proof of Corollary 9. The proof of Theorem 9 gives

$$\sup_{a\in\mathcal{A}_\delta} \left|h^{-4}\{\mathcal{R}_{\infty,\rho_h}(a,h) - \mathcal{V}\} - \mathcal{J}_{\rho_\star}(a)\right| \le C\left\{h^2 + \left|\frac{\rho_h}{h^4} - \rho_\star\right|\right\}.$$

The minimizing property of a_h and the quadratic-growth inequality imply

$$\kappa_J \|a_h - a_\star\|_2^2 \le C\left\{h^2 + \left|\frac{\rho_h}{h^4} - \rho_\star\right|\right\}.$$

Taking square roots completes the proof. □

Proof of Proposition 24. The uniform expansion in the proof of Theorem 9 gives

$$\sup_{a\in\mathcal{A}_\delta} \left|h^{-4}\{\mathcal{R}_{\infty,\rho}(a,h) - \mathcal{V}\} - \mathcal{J}_{\rho_\star}(a)\right| \le C\left\{h^2 + \left|\frac{\rho}{h^4} - \rho_\star\right|\right\}.$$

The minimizing property of $a(h,\rho)$ and quadratic growth of $\mathcal{J}_{\rho_\star}$ prove the first inequality.

For every unit tangent vector v,

$$|v^\top\{g_a(z) - g_{a_\star}(z)\}v| \le \|D\phi(z)\|^2_{\mathrm{op}}\|a - a_\star\|_\infty.$$

Compactness gives the tensor bound uniformly in z. Uniform ellipticity of $g_{a_\star}$ then gives

$$(1 - C\|a - a_\star\|_2)g_{a_\star} \preceq g_a \preceq (1 + C\|a - a_\star\|_2)g_{a_\star}$$

when $\|a - a_\star\|_2$ is sufficiently small. Integrate along a curve, take square roots, and then take the infimum over curves joining z and z'. This gives the geodesic ratio bound in that neighborhood. Outside it, $\|a - a_\star\|_2$ is bounded below by a positive constant, while uniform ellipticity bounds every distance ratio above and away from zero. Increasing the constant gives the same bound on all of $\mathcal{A}_\delta$. This completes the proof. □

We next prove the finite-block approximation. The first lemma treats empirical scaling as a finite-dimensional convex equation and controls its parameter derivatives.

Let $\Theta = \mathcal{A}_\delta \times \mathcal{H}$, write $z_\theta = \log u_\theta$, and put

$$z_{b,\theta,i} = \log\{\sqrt{b}\, u_{b,\theta,i}\}, \qquad \|v\|^2_{2,b} = b^{-1}\sum_{i=1}^b v_i^2.$$

Lemma 6. *Suppose Condition 1 holds and $m\log(mb)/b \le c_0$. For $r = 0, 1, 2$,*

$$\mathbb{E}\Big[\sup_{\theta\in\Theta}\sup_{\|v_1\|_2,\dots,\|v_r\|_2\le 1}\Big\|D_{v_1}\cdots D_{v_r}\{z_{b,\theta} - z_\theta(X_{1:b})\}\Big\|^2_{2,b}\Big] \le C\frac{m\log(mb)}{b}.$$

The same bound holds for the corresponding scaled smoother applied to a bounded function of the observed coordinates and for its first two metric derivatives.

Lemma 6 supplies the derivative approximation needed for both risk transfer and the joint population limit.

Proof of Lemma 6. Integrating the population scaling equation gives

$$1 = \iint u_\theta(x) k_\theta(x, x') u_\theta(x') \, dP_\xi(x) \, dP_\xi(x').$$

Because $k_- \le k_\theta \le 1$, this identity and the pointwise scaling equation give

$$1 \le \int u_\theta \, dP_\xi \le k_-^{-1/2}, \qquad k_-^{1/2} \le u_\theta(x) \le k_-^{-1}$$

uniformly in x and θ. The population bistochastic kernel

$$w_\theta(x, x') = u_\theta(x) k_\theta(x, x') u_\theta(x')$$

is bounded below by a constant $c_0 > 0$. Hence $T_\theta = c_0 \Pi + (1 - c_0) Q_\theta$, where $\Pi f = \int f \, dP_\xi$ and Q_θ is a self-adjoint Markov operator. Thus

$$\inf_\theta \lambda_{\min}(I + T_\theta) \ge c_0.$$

Differentiating the population equation now gives uniform bounds for its first two metric derivatives.

For $\theta = (a, h)$, define the empirical row map

$$G_{b,\theta}(z)_i = e^{z_i} \frac{1}{b} \sum_{j \ne i} k_\theta(X_i, X_j) e^{z_j} - 1.$$

At $z_\theta^0 = \{z_\theta(X_i)\}_{i=1}^b$, conditional Hoeffding bounds, a union bound over i, and a Euclidean net for the $m - 1$ free metric coordinates give

$$\mathbb{E} \sup_{\theta \in \Theta} \|G_{b,\theta}(z_\theta^0)\|_{2,b}^2 \le C \frac{m \log(mb)}{b}.$$

The same calculation applies after one or two directional derivatives. The diagonal omission contributes $O(b^{-1})$. For a metric direction v, put

$$r_{\theta,v}(x) = \int u_\theta(x) k_\theta(x, x') u_\theta(x') D_v \log k_\theta(x, x') \, dP_\xi(x').$$

The kernel and population scaling derivatives are uniformly bounded by the population implicit equation

$$(I + T_\theta) D_v z_\theta = -r_{\theta,v}$$

and its second derivative. The preceding spectral bound controls the inverse of $I + T_\theta$.

The empirical row Jacobian is

$$J_{b,\theta}(z) = \operatorname{diag}\{G_{b,\theta}(z) + \mathbf{1}\} + \frac{1}{b} \operatorname{diag}(e^z) K_\theta \operatorname{diag}(e^z).$$

On a fixed sup-norm neighborhood of z_θ^0, every row sum is bounded above and away from zero and every off-diagonal entry of the row-normalized second term lies between c/b and C/b. If P is this row-normalized matrix, every entry of P^2 is at least c_2/b because $b \ge 3$. The Dobrushin contraction bound gives

$$\max\{|\lambda| : \lambda \in \operatorname{spec}(P), \ \lambda \ne 1\} \le (1 - c_2)^{1/2}.$$

The matrix P is similar to a symmetric matrix. Congruence by the square root of the row-sum matrix now gives

$$\inf_{\theta, z} \lambda_{\min}\{J_{b,\theta}(z)\} \ge c_J > 0$$

on this neighborhood. Its first two derivatives in z and θ are uniformly bounded in row-sum norm.

Lemma 4 bounds $z_{b,\theta}$ and z_θ^0, and hence their joining segment, in sup norm. The same eigenvalue bound holds there. Because $G_{b,\theta}(z_{b,\theta}) = 0$, integration gives

$$-G_{b,\theta}(z_\theta^0) = \left\{ \int_0^1 J_{b,\theta}\big(z_\theta^0 + t(z_{b,\theta} - z_\theta^0)\big) \, dt \right\} (z_{b,\theta} - z_\theta^0).$$

The matrix in braces has smallest eigenvalue at least c_J. Hence

$$\|z_{b,\theta} - z_\theta^0\|_{2,b} \le c_J^{-1} \|G_{b,\theta}(z_\theta^0)\|_{2,b}.$$

Integration of the exponential tail bound proves the assertion for $r = 0$.

For $r = 1$, implicit differentiation gives

$$J_{b,\theta}(z_{b,\theta}) D_v z_{b,\theta} = -D_v G_{b,\theta}(z_{b,\theta}).$$

Subtract the population equation at $X_{1:b}$. The $r = 0$ bound, resolvent identity, and differentiated empirical-process bound give the assertion. Differentiating again adds products of bounded terms and proves $r = 2$.

The empirical smoother equals

$$e^{z_{b,\theta,i}} \frac{1}{b} \sum_{j \ne i} k_\theta(X_i, X_j) e^{z_{b,\theta,j}} f(X_j).$$

Add and subtract the expression with z_θ^0 and then its population integral. The scaling bound and the same empirical-process calculation prove the smoother assertion for $r = 0, 1, 2$. This completes the proof. □

The next lemma also controls the random block score itself. This is needed because the limiting score in Theorem 15 varies across masks.

Put

$$s_{b,\xi}(a,h) = \nabla_a \ell_b(a,h;\xi), \qquad s_{\infty,\xi}(a,h) = \nabla_a r_\xi(a,h).$$

Lemma 7. *Under Condition 1, for fixed m and on a neighborhood of (a_0, h_0),*

$$\mathbb{E} \sup_{a,h} \|s_{b,\xi}(a,h) - s_{\infty,\xi}(a,h)\|_2^2 \le C \frac{m \log(mb)}{b},$$

and

$$\mathbb{E} \sup_{a,h} \|\nabla_a s_{b,\xi}(a,h) - \nabla_a s_{\infty,\xi}(a,h)\|_{\mathrm{op}} \le C \sqrt{\frac{m \log(mb)}{b}}.$$

Lemma 7 shows that mask variation, and not within-block smoothing noise, determines the root-R covariance when b grows at the rate in Theorem 15.

Proof of Lemma 7. Write

$$X_{j,\xi^c} = m_\xi(X_{j,\xi}) + e_j, \qquad \mathbb{E}(e_j \mid X_{j,\xi}) = 0.$$

The score and Hessian contain the smoother, its first two derivatives, and bounded residuals.

For the noise terms, zero diagonal and conditional independence give, for $r = 0, 1, 2$ and unit metric directions,

$$\mathbb{E}\Big[\Big\| \sum_j D_v^r W_{ij} e_j \Big\|_2^2 \;\Big|\; X_{1,\xi}, \dots, X_{b,\xi} \Big] \le C \sum_j |D_v^r W_{ij}|^2 \le \frac{C}{b}.$$

The last inequality follows from Lemma 4 and the differentiated scaling equations. Put $q = |\xi^c|$, let $V = X_{\xi^c}$ and let M have rows $m_\xi(X_{i,\xi})$. For $r = 1, 2$, write $A_r = D_v^r (I_b - W)^2$. The same derivative bounds give $\max_{i,j} |(A_r)_{ij}| \le C/b$ and $\max_i \sum_j |(A_r)_{ij}| \le C$. Conditional independence gives

$$\mathbb{E}(D_v^r \ell_b \mid X_{1,\xi}, \dots, X_{b,\xi}) = \frac{\mathrm{tr}(M^\top A_r M)}{bq} + \frac{1}{bq} \sum_i (A_r)_{ii} \mathbb{E}(\|e_i\|_2^2 \mid X_{i,\xi}).$$

The last term is bounded by C/b uniformly in the parameters and directions.

Conditional on the observed views, $D_v^r \ell_b = (bq)^{-1} \mathrm{tr}(V^\top A_r V)$. Changing one bounded response row alters this quadratic form by at most C/b, by the absolute row-sum bound. Conditional bounded differences gives tail $2\exp(-cbt^2)$. Joint b^{-2}-nets of metrics, unit directions, and bandwidths have

cardinality at most $(Cmb)^{Cm}$. Lemma 4 bounds interpolation error by Cb^{-2}. A union bound and tail integration give squared expected uniform error $Cm\log(mb)/b$.

Lemma 6 replaces WM and its first two derivatives by the corresponding population smoothers with the same squared expected order. The resulting empirical averages have bounded summands and metric derivatives through order three by the population scaling equation. Hoeffding's inequality on the same nets gives the same order relative to population averages. Taking suprema of first derivatives and absolute quadratic forms over unit directions gives the score bound and, by Cauchy–Schwarz, the Hessian bound. Boundedness covers $m\log(mb)/b$ bounded away from zero. This completes the proof. □

Proof of Proposition 31. For one task, let V_i denote its response, put $m_i = \mathbb{E}(V_i \mid X_{i,\xi})$, and write $V_i = m_i + \epsilon_i$. Conditional on the observed views, the errors are independent and centered. Because $W_{ii} = 0$,

$$\mathbb{E}\Big\|V_i - \sum_j W_{ij}V_j\Big\|_2^2 = \mathbb{E}\|\epsilon_i\|_2^2 + \mathbb{E}\Big\|m_i - \sum_j W_{ij}m_j\Big\|_2^2 + \sum_j \mathbb{E}\{W_{ij}^2\|\epsilon_j\|_2^2\}.$$

Lemma 4 bounds the final prediction variance term and its first two derivatives by C/b. Lemma 6 bounds the conditional-mean approximation and its derivatives. Lemma 7 gives the corresponding score bounds. When $m\log(mb)/b \le c_0$, Cauchy–Schwarz gives

$$\sup_{a,h} |\mathcal{R}_{b,\rho}(a,h) - \mathcal{R}_{\infty,\rho}(a,h)| \le C\sqrt{\frac{m\log(mb)}{b}}.$$

When $m\log(mb)/b > c_0$, the ridge terms cancel and both prediction risks lie in $[0, 4B^2]$, and hence the same inequality holds with $C \ge 4B^2 c_0^{-1/2}$. Taking expectations in the derivative bounds proves the same order for the first two derivatives. The kernel and scaling equations extend smoothly to an open positive neighborhood of $\mathcal{A}_\delta$; restriction to feasible paths proves the one-sided boundary assertion. This completes the proof. □

Proof of Corollary 12. Proposition 31 gives

$$\mathcal{R}_{\infty,\rho}(a_{0,b}, h_{0,b}) - \mathcal{R}_{\infty,\rho}(a_0, h_0) \le 2C_{\mathrm{bl}}\sqrt{\frac{m\log(mb)}{b}}.$$

For all sufficiently large b, the right side is smaller than the bandwidth gap in Condition 2, which gives $h_{0,b} = h_0$. Compactness, uniqueness, and uniform convergence give $a_{0,b} \to a_0$. The local quadratic inequality in Condition 2 then yields

$$\kappa\|a_{0,b} - a_0\|_2^2 \le 2C_{\mathrm{bl}}\sqrt{\frac{m\log(mb)}{b}}.$$

This completes the proof. □

The finite-mask error can also be bounded by a parameter covering argument. The subsequent oracle bound uses the sharper logarithmic estimate.

Proof of Theorem 11. Conditional on the blocks, the RJ centered mask losses are independent, bounded, and uniformly Lipschitz in a. Use a coordinatewise $(RJ)^{-1}$-net of $\mathcal{A}_\delta$, whose cardinality is at most $(CmRJ)^m$. Hoeffding's inequality on this net and the uniform Lipschitz extension give the stated bound. A union bound over $\mathcal{H}$ gives the stated probability and order. This completes the proof. □

Proof of Corollary 13. The minimizing inequality used in Theorem 2, both bounds in Corollary 3, and Proposition 31 give a population excess risk of order

$$\sqrt{\frac{\log(2m|\mathcal{H}|)}{R}} + \sqrt{\frac{\log(2m|\mathcal{H}|)}{RJ}} + \sqrt{\frac{m\log(mb)}{b}} + \eta_R.$$

Condition 2 converts excess risk into squared Euclidean error in a neighborhood of a_0. Uniqueness and compactness place the estimator in this neighborhood with probability tending to one. Taking square roots proves the result. This completes the proof. □

Proof of Proposition 32. Fix a mask and abbreviate its observed kernel by k. All constants are uniform over the finite mask support and $a \in \mathcal{A}_\delta$. Uniform ellipticity and the density bounds give

$$ch^d \le \int k_{a,h,\xi}(x,x')\,dP_\xi(x') \le Ch^d.$$

The observed support also has local volume at least cr^d for small r. The bounded support gives $|\log k(x,y) - \log k(x',y)| \le C\|x-x'\|_2/h^2$.

First consider the population scaling u. Its reciprocal is $\int k(x,y)u(y)\,dP_\xi(y)$, and hence its logarithm has the same Lipschitz bound. If $M = \max u$, the scaling is at least cM in a ball of radius ch^2 about a maximizer. The row equation there gives $1 \ge cM^2h^{2d}$. The upper kernel-mass bound then gives

$$\max_x u(x) \le Ch^{-d}, \qquad \min_x u(x) \ge \{Ch^d \max_x u(x)\}^{-1} \ge c.$$

Put $v_i = \sqrt{b}\,u_i$ for the empirical zero-diagonal scaling. Bernstein's inequality for balls and kernel sections gives, with probability at least $1 - b^{-A}$ for any fixed $A > 0$,

$$cbh^d \le \sum_{j\ne i} k_{a,h,\xi}(X_i,X_j) \le Cbh^d,$$

and at least cbh^{2d} other observations in every ball of radius ch^2 centered at a sample point. Uniformity in a follows from a net of polynomial size in h^{-1} and $bh^{2d} \ge C_A \log b$.

Let G_b be this row-sum and ball-count event. Conditional on one observed view and any other sample center, Bernstein's inequality applies to the remaining $b-2$ independent observations. Each fixed observation contributes at most one. The same nets and union bound give

$$\max_{j\le b} \operatorname*{ess\,sup}_x \mathbb{P}(G_b^c \mid X_{j,\xi} = x) \le C_A b^{-A}.$$

Let $M = v_i = \max_j v_j$. For a nearby index j, compare rows i and j after removing their two mutual terms. The log-kernel bound gives

$$\frac{1}{M} \ge c\left\{\frac{1}{v_j} - \frac{k(X_i,X_j)M}{b}\right\} + \frac{k(X_i,X_j)v_j}{b}, \qquad v_j \ge \frac{cM}{1+M^2/b}.$$

There are $D \ge cbh^{2d}$ such indices and their affinities to i are bounded below. Thus $1 \ge cDM^2/(b+M^2)$. For sufficiently large C_A, this implies $M^2 \le Cb/D \le Ch^{-2d}$. The row upper bound yields $\min_i v_i \ge c$. In particular, both the empirical and population scaling vectors lie between c and Ch^{-d}.

Write $K^0_{ij} = k(X_i,X_j)\mathbf{1}\{i \ne j\}$ and define $F_b(w) = K^0 w/b - w^{-1}$, with componentwise inversion. The full Gaussian matrix $K^0 + I$ is positive semidefinite. For vectors $w, \widetilde{w}$ in the preceding range,

$$\begin{aligned}&(w-\widetilde{w})^\top\{F_b(w) - F_b(\widetilde{w})\}\\ &\quad= \frac{1}{b}(w-\widetilde{w})^\top K^0(w-\widetilde{w}) + \sum_i \frac{(w_i-\widetilde{w}_i)^2}{w_i\widetilde{w}_i} \ge (ch^{2d} - b^{-1})\|w-\widetilde{w}\|_2^2 \ge c'h^{2d}\|w-\widetilde{w}\|_2^2.\end{aligned}$$

Set $t_i = u(X_i)$. Because $\int k(x,y)u(y)\,dP_\xi(y) = u(x)^{-1}$,

$$\int k(x,y)^2u(y)^2\,dP_\xi(y) \le \frac{\max u}{u(x)} \le Ch^{-d}.$$

Bernstein's inequality, the uniform bound Ch^{-d}, and a parameter net give

$$\sup_{a,x}\left|\frac{1}{b}\sum_{j=1}^b k_{a,h,\xi}(x,X_j)u_{a,h,\xi}(X_j) - \int k_{a,h,\xi}(x,x')u_{a,h,\xi}(x')\,dP_\xi(x')\right| \le C\sqrt{\frac{\log b}{bh^d}}.$$

Because $bh^d \ge C\log b$, the linear Bernstein term is absorbed. The derivative in u of $Ku - u^{-1} = 0$ is $K + \operatorname{diag}(u^{-2})$, bounded below by ch^{2d} in $L^2(P_\xi)$. Kernel differentiation and the row equation

bound parameter derivatives of u by powers of h^{-1} in L^2 and pointwise. Hence the log net size is $O(\log b)$.

Deleting the diagonal changes $F_b(t)$ by at most Ch^{-d}/b in each row. Because $F_b(v) = 0$, the monotonicity inequality gives

$$b^{-1/2}\|v - t\|_2 \le C\sqrt{\frac{\log b}{bh^{5d}}}.$$

For a bounded scalar response function g, the difference of the empirical operators with scalings v and t equals

$$\operatorname{diag}(v - t)\frac{K^0}{b}\operatorname{diag}(v)g + \operatorname{diag}(t)\frac{K^0}{b}\operatorname{diag}(v - t)g.$$

The row equation and $\min v_i \ge c$ bound the first term by $C|v_i - t_i|$ rowwise. For the second, $\max t_i \le Ch^{-d}$ and $\|K^0/b\|_{\mathrm{op}} \le Ch^d$ bound the multiplying operator. The normalized error is at most $Cb^{-1/2}\|v - t\|_2$. Bernstein's inequality for the operator with population scaling contributes at most $C\sqrt{\log b/(bh^{2d})} + C/(bh^{2d})$, which is bounded by $Cq_b(h)$.

For the fixed mask, let V_i be the response, $q = \dim V_i$, $m_i = \mathbb{E}(V_i \mid X_{i,\xi})$, and $\sigma_i^2 = q^{-1}\mathbb{E}(\|V_i - m_i\|_2^2 \mid X_{i,\xi})$. Conditional independence and $W_{ii} = 0$ give

$$\mathbb{E}(\ell_b \mid X_{1,\xi}, \dots, X_{b,\xi}) = \frac{1}{bq}\sum_i \Big\|m_i - \sum_j W_{ij}m_j\Big\|_2^2 + \frac{1}{b}\sum_j \sigma_j^2 + \frac{1}{b}\sum_{i,j} W_{ij}^2\sigma_j^2.$$

The middle term has the same expectation as the irreducible variance in the population risk. On G_b, the scaling bounds give $\max_{i,j} W_{ij} \le C/(bh^{2d})$. Because every column of W sums to one,

$$\frac{1}{b}\sum_{i,j} W_{ij}^2\sigma_j^2 \le \Big(\max_{i,j} W_{ij}\Big)\frac{1}{b}\sum_j \sigma_j^2.$$

On G_b^c, the same inequality holds with the maximum weight replaced by one. The conditional probability bound gives

$$\mathbb{E}\Big(\frac{1}{b}\sum_{i,j} W_{ij}^2\sigma_j^2\Big) \le C\Big\{\frac{1}{bh^{2d}} + b^{-A}\Big\}\mathbb{E}\sigma_1^2.$$

Smooth conditional means are bounded on the compact support. The operator comparison and Cauchy–Schwarz give squared-error difference $Cq_b(h)$; boundedness controls G_b^c. The empirical error for the population smoother has exactly its population expectation. Finite response second moments, $0 < h \le 1$, and finite mask averaging prove the bound. This completes the proof. □

Proof of Proposition 33. Fix a mask ξ of positive probability. Its observed support has dimension $d_\xi = d_{\xi,\mathcal{S}} + |\xi \cap \mathcal{N}|$. The assumed density and local-volume bounds give the kernel-mass estimate and the radius-h^2 ball count used in Proposition 32, with d replaced by d_ξ. One-sided volumes at a boundary change only the constants. The Gaussian log-kernel comparison and positive semidefiniteness of the full Gaussian matrix are unchanged. The same proof thus gives population and normalized empirical scaling upper bounds Ch^{-d_ξ}, a monotonicity constant ch^{2d_ξ}, and block error $Cq_{b,\xi}^{\mathrm{prod}}(h)$.

Because $0 < h \le 1$ and $d_\xi \le d_{\max}$, $bh^{2d_{\max}} \ge C\log b$ implies every mask-specific sample-size condition. Taking the maximum over the finite mask support proves the result. □

Proof of Theorem 12. The block losses are bounded. Proposition 20 gives a loss Lipschitz constant of order h_R^{-2} for every block. A finite net of $\mathcal{A}_\delta$, Hoeffding's inequality over the R blocks, and the same argument over the RJ masks give

$$\sup_{a \in \mathcal{A}_\delta} |\widehat{\mathcal{R}}_{\rho_R}(a, h_R) - \mathcal{R}_{\infty,\rho_R}(a, h_R)| = O_{\mathbb{P}}\{\Delta_{R,J,b}(h_R)\}.$$

Proposition 32 supplies the block term in this expression. The approximate minimizing inequality yields

$$h_R^{-4}\left[\mathcal{R}_{\infty,\rho_R}(\widehat{a}_{\mathrm{geo}}, h_R) - \mathcal{R}_{\infty,\rho_R}\{a(h_R, \rho_R), h_R\}\right] = O_{\mathbb{P}}\left\{\frac{\Delta_{R,J,b}(h_R)}{h_R^4}\right\}.$$

The uniform expansion in Theorem 9 and quadratic growth of $\mathcal{J}_{\rho_\star}$ now give

$$\|\widehat{a}_{\mathrm{geo}} - a_\star\|_2^2 = O_{\mathbb{P}}\left\{h_R^2 + \left|\frac{\rho_R}{h_R^4} - \rho_\star\right| + \frac{\Delta_{R,J,b}(h_R)}{h_R^4}\right\}.$$

Taking square roots proves the parameter rate. The tensor and distance calculation in Proposition 24 proves the final assertion. This completes the proof. □

Proof of Corollary 14. The continuous signal map has compact domain, and each nuisance coordinate has compact support. Thus the response coordinates are bounded. The concentration argument in Theorem 12 is unchanged. Proposition 33 replaces its block term. Hence

$$\sup_{a\in\mathcal{A}_\delta} |\widehat{\mathcal{R}}_{\rho_R}(a, h_R) - \mathcal{R}_{\infty,\rho_R}(a, h_R)| = O_{\mathbb{P}}\{\Delta^{\mathrm{prod}}_{R,J,b}(h_R)\}.$$

Apply the uniform expansion in Theorem 10 and the quadratic-growth condition in Corollary 11. The approximate minimizing inequality gives

$$\|\widehat{a}_{\mathrm{geo}} - a_\star^{\mathrm{mask}}\|_2 = O_{\mathbb{P}}(e^{\mathrm{prod}}_{R,J,b}).$$

The tensor and geodesic calculation in Corollary 11 gives the two remaining terms. This completes the proof. □

We now prove the constrained local limit. The first-order population gradient distinguishes the tangent cone from the critical cone.

Proof of Theorem 13. Uniform convergence to the finite-block risk, local uniqueness, and $\eta_R = o_{\mathbb{P}}(1)$ give $(\widehat{a}, \widehat{h}) \to (a_{0,b}, h_{0,b})$ in probability. The bandwidth gap gives $\mathbb{P}(\widehat{h} = h_{0,b}) \to 1$.

For local quadratic growth, take feasible $d_k \to 0$ with convergent $d_k/\|d_k\|_2$. If the limiting direction is outside the critical cone, the first-order term divided by $\|d_k\|_2^2$ diverges. On the cone, the normalized quadratic term is at least $\kappa/2 + o(1)$. The uniform empirical gradient error on a fixed neighborhood is $O_{\mathbb{P}}(R^{-1/2})$ by the bounded derivatives and finite dimension. The approximate minimizing inequality gives

$$c\|\widehat{a} - a_{0,b}\|_2^2 \le O_{\mathbb{P}}(R^{-1/2})\|\widehat{a} - a_{0,b}\|_2 + \eta_R.$$

Thus $\sqrt{R}(\widehat{a} - a_{0,b}) = O_{\mathbb{P}}(1)$.

For v in a fixed compact set, put $a_R(v) = a_{0,b} + v/\sqrt{R}$. Feasibility of $a_R(v)$ converges to $v \in \mathcal{T}_{0,b}$. A Taylor expansion of the empirical criterion gives

$$R\{\widehat{\mathcal{R}}_\rho(a_R(v), h_{0,b}) - \widehat{\mathcal{R}}_\rho(a_{0,b}, h_{0,b})\} = \sqrt{R}\, g_{0,b}^\top v + v^\top Z_R + \frac{1}{2} v^\top H_{0,b} v + o_{\mathbb{P}}(1)$$

uniformly on compact subsets of $\mathcal{T}_{0,b}$, where

$$Z_R = R^{-1/2} \sum_{r=1}^{R} \{\nabla_a q_b(a_{0,b}, h_{0,b}; \zeta_r) - \mathbb{E}\nabla_a q_b(a_{0,b}, h_{0,b}; \zeta_r)\}.$$

Lemma 4 and the multivariate central limit theorem give $Z_R \Longrightarrow Z_{0,b}$. The population minimizing inequality gives $g_{0,b}^\top v \ge 0$ for $v \in \mathcal{T}_{0,b}$. The first term diverges for every tangent direction outside $\mathcal{C}_{0,b}$ and equals zero on $\mathcal{C}_{0,b}$.

Write $q_z(v) = z^\top v + v^\top H_{0,b} v/2$ on $\mathcal{C}_{0,b}$. For $u, v \in \mathcal{C}_{0,b}$, their difference belongs to $\mathrm{span}(\mathcal{C}_{0,b})$, and

$$q_z(tu + (1-t)v) \le t q_z(u) + (1-t) q_z(v) - \frac{\kappa}{2} t(1-t)\|u - v\|_2^2, \qquad 0 \le t \le 1.$$

Also, $q_z(v) \ge -\|z\|_2\|v\|_2 + \kappa\|v\|_2^2/2$. Hence there is a unique minimizer v_z. The variational inequality at v_z gives

$$q_z(v) - q_z(v_z) \ge \frac{\kappa}{2}\|v - v_z\|_2^2, \qquad v \in \mathcal{C}_{0,b}.$$

These inequalities give continuity in every score z and every Hessian with the same curvature bound on the cone span, including singular score covariances. The optimization error contributes $R\eta_R = o_{\mathbb{P}}(1)$.

Put $v_R = \sqrt{R}(\widehat{a} - a_{0,b})$. From any subsequence, joint tightness gives a further subsequence on which (v_R, Z_R) converges in distribution. Use a joint almost-sure representation that includes $R\eta_R$, the indicator of $\widehat{h} \neq h_{0,b}$, and the absolute uniform expansion remainders on the intersection of $\mathcal{T}_{0,b}$ with each closed ball of positive integer radius. Write (v, z) for the limit of (v_R, Z_R). The linear constraints give $v_R \in \mathcal{T}_{0,b}$ and make $a_{0,b} + u/\sqrt{R}$ feasible eventually for every fixed $u \in \mathcal{C}_{0,b}$. Comparison with $u = 0$ bounds the nonnegative term $\sqrt{R}\, g_{0,b}^\top v_R$, which implies $v \in \mathcal{C}_{0,b}$. Comparison with any fixed $u \in \mathcal{C}_{0,b}$, after dropping this nonnegative term and taking limits, gives $q_z(v) \leq q_z(u)$. The uniform remainders make these comparisons simultaneous on every compact set. Uniqueness gives $v = v_z$. Every subsequential limit has the asserted distribution. This completes the proof. □

Proof of Theorem 14. Let g_r^0 be $\widehat{g}_r$ at $(a_{0,b}, h_{0,b})$. Lemma 4, parameter consistency, and bandwidth selection give

$$\max_{r \leq R} \|\widehat{g}_r - g_r^0\|_2 = o_{\mathbb{P}}(1), \qquad \widehat{H}_{0,b} - H_{0,b} = o_{\mathbb{P}}(1).$$

The second assertion follows from convergence of the empirical averages at the population argument and the bounded third derivatives. The independent bounded scores satisfy

$$\frac{1}{R}\sum_{r=1}^{R} g_r^0 (g_r^0)^\top - \mathbb{E}\{g_r^0 (g_r^0)^\top\} = o_{\mathbb{P}}(1), \qquad \frac{1}{R}\sum_{r=1}^{R} g_r^0 - g_{0,b} = o_{\mathbb{P}}(1).$$

Replacing g_r^0 by $\widehat{g}_r$ and centering gives $\widehat{\Sigma}_{0,b} \to \Sigma_{0,b}$, also at boundary targets.

Suppose $A_{0,b}^-$ is empty. The Karush–Kuhn–Tucker equations give $g_{0,b} \in \operatorname{span}(\mathbf{1})$. Hence $\mathcal{C}_{0,b} = \mathcal{T}_{0,b} = \mathbf{1}^\perp$. Write $v = Qw$. The minimizer in Theorem 13 is $-Q(Q^\top H_{0,b} Q)^{-1} Q^\top Z_{0,b}$. Its covariance on the Q coordinates is $V_{Q,b}$. The eigenvalues of $Q^\top H_{0,b} Q$ are at least κ. Hessian consistency makes the eigenvalue truncation inactive with probability tending to one, and continuity of matrix inversion proves $\widehat{V}_{Q,b} \to V_{Q,b}$. Slutsky's theorem gives the standardized contrast limit. This completes the proof. □

Proof of Theorem 15. Theorem 2, Proposition 31, Condition 2, and $\eta_R = o_{\mathbb{P}}(1)$ give

$$(\widehat{a}, \widehat{h}) \longrightarrow (a_0, h_0) \quad \text{in probability.}$$

The bandwidth gap gives $\mathbb{P}(\widehat{h} = h_0) \to 1$.

Let

$$S_{Rr} = \nabla_a q_{b_R}(a_0, h_0; \zeta_r), \qquad Z_R = R^{-1/2} \sum_{r=1}^{R} \{S_{Rr} - \mathbb{E} S_{Rr}\}.$$

Lemma 7 and Proposition 31 give

$$\sqrt{R}\|\mathbb{E} S_{Rr} - g_0\|_2 \leq C \left\{ \frac{Rm \log(mb_R)}{b_R} \right\}^{1/2} = o(1)$$

and $\operatorname{var}(S_{Rr}) \longrightarrow \Sigma_0$. The scores are uniformly bounded by Lemma 4. The Lindeberg condition for this triangular array follows, and hence

$$Z_R \Longrightarrow Z_0.$$

The empirical Hessian has the decomposition

$$\nabla_a^2 \widehat{\mathcal{R}}_\rho(a, h_0) - H_0 = \{\nabla_a^2 \widehat{\mathcal{R}}_\rho(a, h_0) - \nabla_a^2 \mathcal{R}_{b_R,\rho}(a, h_0)\} + \{\nabla_a^2 \mathcal{R}_{b_R,\rho}(a, h_0) - H_0\}.$$

Bounded third derivatives in Lemma 4 give uniform convergence of the first term. Proposition 31 approximates the second by $\nabla_a^2 \mathcal{R}_{\infty,\rho}(a, h_0) - H_0$, bounded by $C\|a - a_0\|_2$. Thus convergence is uniform on shrinking neighborhoods of a_0.

For v in a compact set, put $a_R(v) = a_0 + v/\sqrt{R}$. A Taylor expansion now gives, uniformly on compact subsets of $\mathcal{T}_0$,

$$R\{\widehat{\mathcal{R}}_\rho(a_R(v), h_0) - \widehat{\mathcal{R}}_\rho(a_0, h_0)\} = \sqrt{R}\, g_0^\top v + Z_R^\top v + \frac{1}{2} v^\top H_0 v + o_{\mathbb{P}}(1).$$

The population Karush–Kuhn–Tucker inequality gives $g_0^\top v \geq 0$ on $\mathcal{T}_0$. This term equals zero on $\mathcal{C}_0$ and diverges outside $\mathcal{C}_0$. The local quadratic-growth argument in the proof of Theorem 13, the $O_{\mathbb{P}}(R^{-1/2})$ score error, and the uniform Hessian convergence give $\sqrt{R}(\widehat{a} - a_0) = O_{\mathbb{P}}(1)$. Curvature on $\mathrm{span}(\mathcal{C}_0)$ gives coercivity, a unique minimizer, and the same quadratic separation inequality. The computed-solution discrepancy is $R\eta_R = o_{\mathbb{P}}(1)$. Apply the subsequence comparison argument in the proof of Theorem 13 with a_0, $\mathcal{T}_0$, $\mathcal{C}_0$, H_0, and Z_0 in place of their finite-block counterparts. Comparison with zero places every limit in $\mathcal{C}_0$, and comparison with each fixed critical direction identifies the unique quadratic minimizer. This completes the proof. □

Proof of Corollary 15. For an interior a_0, $\mathcal{C}_0 = \mathbf{1}^\perp$. Writing $v = Qw$, the minimizer in Theorem 15 is $-Q(Q^\top H_0 Q)^{-1} Q^\top Z_0$. Its covariance is $QV_Q Q^\top$. The proof of Theorem 14 applies to the triangular array because the first three block derivatives remain uniformly bounded. At (a_0, h_0), the variances of the empirical first and second score moments are $O(R^{-1})$. Lemma 7 gives convergence of their population covariance to Σ_0, and Proposition 31 gives convergence of the population Hessian to H_0. Parameter consistency then gives

$$\widehat{\Sigma}_{0,b_R} \longrightarrow \Sigma_0, \qquad \widehat{H}_{0,b_R} \longrightarrow H_0 \quad \text{in probability.}$$

Positive curvature makes the eigenvalue truncation inactive with probability tending to one. Matrix inversion and multiplication prove $\widehat{V}_Q \to V_Q$, and Slutsky's theorem proves the standardized limit. This completes the proof. □

The observation-level limit uses the first-order projection of a mask-averaged block score. The positive stable-bandwidth kernel makes the scaling and score smooth functionals of the empirical distribution.

For a bounded symmetric vector-valued kernel h_2 of two complete observations, write

$$U_b(h_2) = \binom{b}{2}^{-1} \sum_{1 \leq i < j \leq b} h_2(\mathcal{O}_i, \mathcal{O}_j).$$

The following lemma gives the second-order expansion needed for both the observation limit and its block-jackknife covariance. It treats diagonal deletion as part of the finite-block statistic and hence does not replace the estimator by a full-diagonal approximation. For one task, write $O = X_\xi$, let V be its bounded response vector, put $q = \dim(V)$, and let $m_\xi(o) = \mathbb{E}(V \mid O = o)$. Under coordinate masking, $V = X_{\xi^c}$ and $q = |\xi^c|$. A prespecified response instead uses its own dimension, even when $O = X$. Put

$$s_{\infty,\xi}(a, h) = \nabla_a \frac{\mathbb{E}\|V - T_{a,h,\xi} m_\xi(O)\|_2^2}{q}.$$

Lemma 8. *Suppose Condition 1 holds, m and the mask support are fixed, and (a, h) ranges over a neighborhood of (a_0, h_0). For every mask ξ, there are bounded $\mathbb{R}^m$-valued functions ψ_ξ and $h_{2,\xi}$ of one and two complete observations, a vector $\beta_\xi \in \mathbb{R}^m$, and a remainder $r_{\xi,b}$ such that*

$$\nabla_a \ell_b(a_0, h_0; \xi) - s_{\infty,\xi}(a_0, h_0) = \frac{1}{b} \sum_{i=1}^{b} \psi_\xi(\mathcal{O}_i) + \frac{1}{2} U_b(h_{2,\xi}) + \frac{\beta_\xi}{b} + r_{\xi,b},$$

where

$$\mathbb{E}\{\psi_\xi(\mathcal{O})\} = 0, \qquad \mathbb{E}\{h_{2,\xi}(o, \mathcal{O})\} = 0, \qquad \mathbb{E}\|r_{\xi,b}\|_2^2 \leq C \frac{(\log b)^3}{b^3}.$$

If $r_{\xi,b-1}^{(-i)}$ denotes the corresponding remainder after deleting $\mathcal{O}_i$, then

$$\frac{1}{b} \sum_{i=1}^{b} \left\| (b-1)\{r_{\xi,b} - r_{\xi,b-1}^{(-i)}\} \right\|_2^2 = O_{\mathbb{P}}\left\{ \frac{(\log b)^3}{b} \right\}.$$

The same assertions hold after one derivative with respect to a.

Lemma 8 identifies all terms of order at least b^{-1}. The first projection determines the root-Rb limit, while the canonical quadratic term and the diagonal-deletion bias are negligible at that scale under Theorem 16.

Proof of Lemma 8. Fix a task with observed view $O = X_\xi$ and response $V = X_{\xi^c}$ or a prespecified response view; use $O = X$ for a full observed view. Let P be the distribution of $\mathcal{O}$ preceding Theorem 16. For a probability measure Q near P, denote its marginals by Q_O and $Q_{O,V}$. Define $z_Q = \log u_Q$ by

$$F(Q_O, z_Q)(o) = e^{z_Q(o)} \int k_{a,h,\xi}(o, o') e^{z_Q(o')}\, dQ_O(o') - 1 = 0.$$

Embed the bounded complete observation in a fixed cube $D \subset \mathbb{R}^{d_0}$ and choose an integer $s > d_0/2$. Write $H^s(D)$ for the Sobolev space with square-integrable weak derivatives through order s, and define

$$\|\nu\|_* = \sup_{\|f\|_{H^s(D)} \le 1} |\nu f|.$$

Sobolev embedding bounds point evaluations on D. Continuous multiplication in H^s and bounded Gaussian mixed derivatives make integration continuous from the dual space to the observed-view H^s space. Because response coordinates are smooth functions on D, the same statement holds with them in the integrand.

At $Q = P$, let $T = T_{a,h,\xi}$ and

$$B_\nu(o) = u(o) \int k_{a,h,\xi}(o, o') u(o')\, d\nu_O(o').$$

The population scaling bounds in Lemma 6 imply that the transition density of T is bounded below by a fixed $\epsilon_0 > 0$. Thus $T = \epsilon_0 \Pi + (1 - \epsilon_0) T_0$, where $\Pi f = Pf$ and T_0 is a self-adjoint Markov operator. Decomposition into constants and mean-zero functions gives

$$\|(I + T)^{-1}\|_{L^2(P_O) \to L^2(P_O)} + \|(I + T)^{-1}\|_{L^\infty(P_O) \to L^\infty(P_O)} \le C.$$

Extend scaling to the observed-view cube by its reciprocal integral equation. Solve $g + Tg = f$ on the support of P_O and set $g = f - Tg$ elsewhere. Kernel smoothness extends the maximum-norm bound and gives $\|g\|_{H^s} \le \|f\|_{H^s} + C\|g\|_\infty \le C\|f\|_{H^s}$. The implicit function theorem makes $Q \mapsto z_Q$ smooth near P, with uniformly bounded derivatives and next metric derivatives on a smaller neighborhood.

For $\nu(1) = 0$, implicit differentiation gives $z_\nu = -(I + T)^{-1} B_\nu$. For bounded functions g_1, g_2, define

$$\mathcal{Q}(g_1, g_2) = g_1 g_2 + g_1 T g_2 + g_2 T g_1 + T(g_1 g_2),$$

$$\mathcal{M}_\nu(g)(o) = g(o) B_\nu(o) + u(o) \int k_{a,h,\xi}(o, o') u(o') g(o')\, d\nu_O(o').$$

The second derivative is

$$z_{\nu_1,\nu_2} = -(I + T)^{-1} \{ \mathcal{Q}(z_{\nu_1}, z_{\nu_2}) + \mathcal{M}_{\nu_1}(z_{\nu_2}) + \mathcal{M}_{\nu_2}(z_{\nu_1}) \}.$$

One more differentiation gives products of B_{ν_j}, z_{ν_j}, and z_{ν_j,ν_k} acted on by $(I + T)^{-1}$. The inverse and Sobolev multiplication bounds give, for $r = 1, 2, 3$,

$$\|D_Q^r z_P[\nu_1, \ldots, \nu_r]\|_\infty \le C \prod_{j=1}^{r} \|\nu_j\|_*.$$

The same bound holds for one additional metric derivative because the kernel derivatives and the differentiated resolvent are uniformly bounded.

Define

$$G_Q(o) = e^{z_Q(o)} \int k_{a,h,\xi}(o, o') e^{z_Q(o')} v'\, dQ_{O,V}(o', v')$$

and

$$L_{\xi,a,h}(Q) = \frac{1}{q} \int \|v - G_Q(o)\|_2^2\, dQ_{O,V}(o, v).$$

The preceding derivative bounds show that $S_\xi(Q) = \nabla_a L_{\xi,a_0,h_0}(Q)$ is three times Fréchet differentiable in $\|\cdot\|_*$. Its derivatives satisfy

$$\|D_Q^r S_\xi(P)[\nu_1,\dots,\nu_r]\|_2 \le C\prod_{j=1}^{r}\|\nu_j\|_*, \qquad r=1,2,3.$$

The elements $\delta_{\mathcal{O}_i} - P$ are independent, centered, and bounded in $H^{-s}(D)$. Only terms with at most three distinct indices have nonzero expectation in the sixth-moment expansion. Hence

$$\mathbb{E}\|P_b - P\|_*^6 \le C\frac{(\log b)^3}{b^3}.$$

For diagonal deletion, replace the scaling equation by $F(Q_O, z) - te^{2z} = 0$ for t near zero, and the predictor at (o, v) by

$$G_{Q,t}(o,v) = e^{z_{Q,t}(o)}\int k_{a,h,\xi}(o,o')e^{z_{Q,t}(o')}v'\,dQ_{O,V}(o',v') - te^{2z_{Q,t}(o)}v.$$

At $Q = P_b$, $t = 1/b$, this is the zero-diagonal smoother. The same inverse gives joint smoothness in (Q, t, a). Let $A_\xi(Q)$ be the t-derivative at zero of its risk gradient. Taylor expansion gives bounded smooth A_ξ and remainder $\Delta_{\xi,b}$ satisfying

$$\nabla_a \ell_b(a_0,h_0;\xi) = S_\xi(P_b) + \frac{1}{b}A_\xi(P_b) + \Delta_{\xi,b}, \qquad \mathbb{E}\|\Delta_{\xi,b}\|_2^2 \le Cb^{-4}.$$

The bound is uniform for leave-one-out measures in the dual-norm neighborhood. The empirical norm has differences $O(b^{-1})$ and mean $O(b^{-1/2})$. Bounded differences makes its complement exponentially small; Lemma 4 controls its contribution to the moment bounds. Put

$$\psi_\xi(\mathfrak{o}) = D_Q S_\xi(P)[\delta_{\mathfrak{o}} - P], \qquad g_{2,\xi}(\mathfrak{o},\mathfrak{o}') = D_Q^2 S_\xi(P)[\delta_{\mathfrak{o}} - P, \delta_{\mathfrak{o}'} - P].$$

Both functions are bounded, and bilinearity gives $\mathbb{E}\{g_{2,\xi}(\mathfrak{o},\mathcal{O})\} = 0$. Expanding $S_\xi(P_b)$ to second order and $A_\xi(P_b)$ to first order gives

$$S_\xi(P_b) - S_\xi(P) = D_Q S_\xi(P)[P_b - P] + \frac{1}{2}D_Q^2 S_\xi(P)[P_b - P, P_b - P] + R_{3,b},$$
$$b^{-1}A_\xi(P_b) = b^{-1}A_\xi(P) + b^{-1}D_Q A_\xi(P)[P_b - P] + R_{A,b},$$

where the squared means of $R_{3,b}$ and $R_{A,b}$ are bounded by $C(\log b)^3/b^3$. Symmetry of the second derivative gives

$$\frac{1}{2}D_Q^2 S_\xi(P)[P_b - P, P_b - P] = \frac{b-1}{2b}U_b(g_{2,\xi}) + \frac{1}{2b^2}\sum_{i=1}^{b} g_{2,\xi}(\mathcal{O}_i,\mathcal{O}_i).$$

We may take $h_{2,\xi} = g_{2,\xi}$ and

$$\beta_\xi = A_\xi(P) + \frac{1}{2}\mathbb{E}\{g_{2,\xi}(\mathcal{O},\mathcal{O})\}.$$

The difference between $(b-1)U_b(g_{2,\xi})/b$ and $U_b(g_{2,\xi})$, the centered diagonal average, and the centered first derivative of A_ξ have squared means of orders b^{-4}, b^{-3}, and b^{-3}. Combining these terms with $R_{3,b}$, $R_{A,b}$, and $\Delta_{\xi,b}$ proves the asserted bound for $r_{\xi,b}$.

Apply the same Taylor calculation jointly to P_b and the b leave-one-out empirical measures. If $P_{b-1}^{(-i)}$ omits $\mathcal{O}_i$, then

$$\|P_b - P_{b-1}^{(-i)}\|_* \le C/b.$$

The integral Taylor remainder and the third derivative bound give

$$(b-1)\|R_{3,b} - R_{3,b-1}^{(-i)}\|_2 \le C\{\|P_b - P\|_* + \|P_{b-1}^{(-i)} - P\|_*\}^2.$$

The quadratic remainders from A_ξ and diagonal deletion satisfy a smaller bound. The sixth-moment inequality above, exchangeability, and Markov's inequality now give

$$\frac{1}{b}\sum_{i=1}^{b}\left\|(b-1)\{r_{\xi,b} - r_{\xi,b-1}^{(-i)}\}\right\|_2^2 = O_{\mathbb{P}}\left\{\frac{(\log b)^3}{b}\right\}.$$

One metric differentiation leaves each bound unchanged. This completes the proof. □

Lemma 9. *Suppose Condition 1 holds, m and the mask support are fixed, and (a, h) ranges over a neighborhood of (a_0, h_0). There is a bounded mean-zero $\mathbb{R}^m$-valued function ψ_0 of one complete observation such that*

$$\overline{s}_b(a_0, h_0; \mathcal{O}_{1:b}) - g_0 = \frac{1}{b}\sum_{i=1}^{b} \psi_0(\mathcal{O}_i) + R_b, \qquad \mathbb{E}\|R_b\|_2^2 \le C\frac{(\log b)^2}{b^2}.$$

The block Hessian differs uniformly from $\nabla_a^2 \mathcal{R}_{\infty,\rho}(a, h)$ on this neighborhood by $O_{\mathbb{P}}\{(\log b/b)^{1/2}\}$.

Lemma 9 separates observation variation from numerical integration over masks. The influence function gives the first-order change of the population predictive score under a perturbation of the observation distribution.

Proof of Lemma 9. For each mask, apply Lemma 8. Average its linear projection over the finite mask support and put

$$\psi_0(o) = \mathbb{E}_\xi\{\psi_\xi(o)\}.$$

The function is bounded and mean zero. A bounded canonical second-order U-statistic has second moment of order b^{-2}. The deterministic diagonal-deletion term has squared norm of the same order, and the third-order remainder is smaller. Mask averaging and the deterministic ridge derivative give

$$\overline{s}_b(a_0, h_0; \mathcal{O}_{1:b}) - g_0 = \frac{1}{b}\sum_{i=1}^{b} \psi_0(\mathcal{O}_i) + R_b, \qquad \mathbb{E}\|R_b\|_2^2 \le C\frac{(\log b)^2}{b^2}.$$

The metric derivative version of Lemma 8 gives uniform approximation to the population Hessian at order $O_{\mathbb{P}}\{(\log b/b)^{1/2}\}$. The finite-dimensional parameter covering uses the bounded third derivatives in Lemma 4. These derivatives also give, for every deterministic $r_b \downarrow 0$,

$$\sup_{a \in \mathcal{A}_\delta:\, \|a-a_0\|_2 \le r_b} \|\nabla_a \overline{s}_b(a, h_0) - H_0\|_{\mathrm{op}} \le C r_b + O_{\mathbb{P}}\{(\log b/b)^{1/2}\}.$$

Thus the Hessian converges to H_0 on shrinking neighborhoods. Conditioning on $\mathcal{O}_1 = o$ and centering removes the b^{-1} bias and canonical quadratic term, identifying ψ_0 in Theorem 16. This completes the proof. □

Proof of Theorem 16. Let $N = Rb_R$. Average the expansion in Lemma 9 over the R disjoint blocks. It gives

$$\sqrt{N}\left\{\frac{1}{R}\sum_{r=1}^{R} \overline{s}_{b_R}(a_0, h_0; \mathcal{O}_{I_r}) - g_0\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi_0(\mathcal{O}_i) + o_{\mathbb{P}}(1),$$

because

$$\sqrt{Rb_R}\frac{\log b_R}{b_R} = \left\{\frac{R(\log b_R)^2}{b_R}\right\}^{1/2} \longrightarrow 0.$$

The multivariate central limit theorem gives convergence of the first term to $N(0, \Omega_0)$.

If J_R masks approximate the mask expectation, the conditional score variance of their average is $O\{(RJ_R)^{-1}\}$. Multiplication by N gives order $b_R/J_R = o(1)$, and hence mask integration does not change the limit. Uniform risk convergence and population identification first give consistency of $\widehat{a}_{\mathrm{obs}}$, and the bandwidth gap gives $\mathbb{P}(\widehat{h}_{\mathrm{obs}} = h_0) \to 1$. Lemma 9 and uniform convergence of the empirical averages then give uniform Hessian convergence to H_0 on shrinking neighborhoods of a_0.

For v in a compact subset of $\mathcal{T}_0$, expand the averaged criterion at $a_0 + v/\sqrt{N}$. The expansion is

$$N\{\widehat{\mathcal{R}}(a_0 + v/\sqrt{N}, h_0) - \widehat{\mathcal{R}}(a_0, h_0)\} = \sqrt{N} g_0^\top v + Z_N^\top v + \frac{1}{2} v^\top H_0 v + o_{\mathbb{P}}(1),$$

where $Z_N \Longrightarrow Z_{\mathrm{obs}}$. Local quadratic growth, the $O_{\mathbb{P}}(N^{-1/2})$ score error, and uniform Hessian convergence give $\sqrt{N}(\widehat{a}_{\mathrm{obs}} - a_0) = O_{\mathbb{P}}(1)$ by the argument in Theorem 13. The first term restricts the limit to $\mathcal{C}_0$. Curvature on its span gives a unique quadratic minimizer by the separation inequality proved there, and $N\eta_R = o_{\mathbb{P}}(1)$. The constrained argmin theorem proves the result. This completes the proof. □

Proof of Corollary 16. Let P_b be one block's empirical distribution and P the observation distribution. In the Hilbert space of Lemma 8, only bounded terms with at most three distinct indices remain in the centered sixth-moment expansion. Hence

$$\mathbb{E}\|P_b - P\|_*^6 \le Cb^{-3}.$$

The uniform third-derivative bound and the exponentially small complement of the fixed dual-norm neighborhood give $\mathbb{E}\|r_{\xi,b}\|_2^2 \le Cb^{-3}$ in that lemma. The canonical quadratic term has second moment $O(b^{-2})$, and the diagonal-deletion bias has norm $O(b^{-1})$. Thus the remainder R_b in Lemma 9 satisfies

$$\mathbb{E}\|R_b\|_2^2 \le Cb^{-2}, \qquad \mathbb{E}\Big\|\frac{\sqrt{Rb}}{R}\sum_{r=1}^{R} R_{b,r}\Big\|_2 \le C\sqrt{R/b} \to 0.$$

One metric differentiation of the same smooth map gives uniform approximation to the population Hessian because $\mathbb{E}\|P_b - P\|_*^2 \le C/b$. Its continuity gives convergence to H_0 on shrinking neighborhoods of a_0. The conditional variance of mask integration is still $O\{(RJ)^{-1}\}$, which is negligible after multiplication by Rb when $b/J \to 0$. Theorem 6 now gives the root-Rb parameter rate. The score central limit theorem and the constrained quadratic expansion in the proof of Theorem 16 apply without further changes.

The leave-one-out remainder in Lemma 8 converges to zero in mean square as $b \to \infty$, without a logarithmic block condition. Parameter consistency gives Proposition 34. The root-Rb rate and threshold conditions give Corollary 21 by its uniform graph argument. This completes the proof. □

Proof of Proposition 34. Apply Lemma 8 jointly to a block and its b leave-one-out versions, and then average over masks. The expansion has the form

$$\overline{s}_b(a_0, h_0; \mathcal{O}_{I_r}) - g_0 = \frac{1}{b}\sum_{j\in I_r}\psi_0(\mathcal{O}_j) + \frac{1}{2}U_b(h_2) + \frac{\beta}{b} + \mathcal{R}_{3,r}.$$

Here h_2 is bounded and canonical, and Lemma 8 gives $\mathbb{E}\|\mathcal{R}_{3,r}\|_2^2 \le C(\log b)^3/b^3$. The linear terms in the leave-one-out difference give

$$(b-1)\{\overline{s}_b(a_0, h_0; \mathcal{O}_{I_r}) - \overline{s}_{b-1}(a_0, h_0; \mathcal{O}_{I_r\setminus\{i\}})\} = \psi_0(\mathcal{O}_i) - \overline{\psi}_{0,r} + e_{ri},$$

where $\overline{\psi}_{0,r} = b^{-1}\sum_{j\in I_r}\psi_0(\mathcal{O}_j)$. Conditional degeneracy gives an average squared contribution of order b^{-1} from the quadratic pseudo-values. The pseudo-value of β/b has norm $O(b^{-1})$. The leave-one-out bound in Lemma 8 controls the third-order term. Hence

$$\frac{1}{Rb}\sum_{r=1}^{R}\sum_{i\in I_r}\|e_{ri}\|_2^2 = o_{\mathbb{P}}(1).$$

The score functional and its leave-one-out derivative are locally Lipschitz in (a, h). Theorem 16 gives $\widehat{a}_{\mathrm{obs}} \to a_0$ and $\mathbb{P}(\widehat{h}_{\mathrm{obs}} = h_0) \to 1$, which leaves the last bound unchanged when the population arguments are replaced by the fitted arguments.

The observations are independent and ψ_0 is bounded. Hence

$$\frac{1}{Rb}\sum_{r=1}^{R}\sum_{i\in I_r}\psi_0(\mathcal{O}_i)\psi_0(\mathcal{O}_i)^\top \longrightarrow \Omega_0 \quad \text{in probability,}$$

while $R^{-1}\sum_r \overline{\psi}_{0,r}\overline{\psi}_{0,r}^\top = O_{\mathbb{P}}(b^{-1})$. Centering the estimated pseudo-values and applying Cauchy–Schwarz to the e_{ri} terms proves $\widehat{\Omega}_0 \to \Omega_0$. This completes the proof. □

Proof of Corollary 17. Let P_X be the distribution of X and let M be the vector measure defined by $M(B) = \mathbb{E}\{V\mathbf{1}_{\{X\in B\}}\}$. For fixed (a, h), let $u = u_{a,h}(P_X)$ be the population scaling and put

$$g(x) = u(x)\int k_{a,h}(x, x')u(x')\,M(dx').$$

The part of the population loss that depends on the metric, apart from its ridge, is

$$F_{a,h}(P_X, M) = \frac{1}{q_V}\left\{-2\int g^\top \, dM + \int \|g\|_2^2 \, dP_X\right\}.$$

The omitted term is $\mathbb{E}\|V\|_2^2/q_V$. For an integer $\ell > m/2$, use the $H^{-\ell}$ norm on a cube containing the support of X, and its product norm $\|\cdot\|_*$ for vector measures. Point masses are uniformly bounded. Gaussian smoothness and the uniform kernel lower bound permit the implicit-function argument of Lemma 8 for P_X. Because $F_{a,h}$ is quadratic in M, its first three measure derivatives and required metric derivatives are locally bounded, uniformly over $a \in \mathcal{A}_\delta$ and $h \in \mathcal{H}$.

For block r, write

$$P_{X,r} = \frac{1}{b}\sum_{i\in I_r}\delta_{X_i}, \qquad M_r = \frac{1}{b}\sum_{i\in I_r} V_i\delta_{X_i}, \qquad \nu_r = P_{X,r} - P_X, \quad \zeta_r = M_r - M.$$

The second response moment and independence imply

$$\mathbb{E}\{\|\nu_r\|_*^2 + \|\zeta_r\|_*^2\} \le C/b, \qquad \max_{r\le R}(\|\nu_r\|_* + \|\zeta_r\|_*) = o_{\mathbb{P}}(1).$$

A union bound and $R/b \to 0$ give the second assertion. The Taylor remainder of $S_{a,h} = \nabla_a F_{a,h}$ is at most $C(\|\nu_r\|_* + \|\zeta_r\|_*)^2$, averaging to $O_{\mathbb{P}}(b^{-1})$ by Markov's inequality. The averaged linear term applies the population derivative to the centered full empirical measure. Thus ψ_0 is centered and $\|\psi_0(X,V)\|_2 \le C(1+\|V\|_2)$.

For diagonal deletion, let $t = 1/b$ and let u_t solve the perturbed equation in Lemma 8. Define g_t by the preceding integral with u_t, $P_{X,r}$, and M_r. At observation i, the fitted response is $g_t(X_i) - tu_t(X_i)^2V_i$. The corresponding squared residual is

$$(1+tu_t(X_i)^2)^2\|V_i\|_2^2 - 2(1+tu_t(X_i)^2)V_i^\top g_t(X_i) + \|g_t(X_i)\|_2^2.$$

At $t = 0$, the coefficient of $\|V_i\|_2^2$ is independent of a. Bounded scaling derivatives give exact block-score error relative to $S_{a,h}(P_{X,r}, M_r)$, with the ridge canceled, at most $Cb^{-1}\{1 + b^{-1}\sum_{i\in I_r}\|V_i\|_2^2\}$. The same bound holds after one metric differentiation. Averaging over blocks gives

$$\nabla_a\widehat{\mathcal{R}}_\rho(a_0,h_0) - \nabla_a\mathcal{R}_{\infty,\rho}(a_0,h_0) = \frac{1}{N}\sum_{i=1}^N \psi_0(\mathcal{O}_i) + O_{\mathbb{P}}(b^{-1}).$$

The calculation gives uniform Hessian convergence on a fixed neighborhood, and convergence to H_0 on shrinking neighborhoods of a_0. Expanding $F_{a,h}$ and using convergence of $N^{-1}\sum_i\|V_i\|_2^2$ to its expectation gives uniform risk consistency. Condition 2 implies metric and bandwidth consistency. Local quadratic growth then gives metric error $O_{\mathbb{P}}(N^{-1/2} + b^{-1} + \eta_R^{1/2})$. Because $\sqrt{N}/b = \sqrt{R/b} \to 0$, the score central limit theorem and the constrained quadratic expansion in the proof of Theorem 16 prove its asserted limit. Only the finite second moment of ψ_0 is used.

For covariance estimation, first evaluate all scores at (a_0, h_0). Write $\psi_i = \psi_0(\mathcal{O}_i)$ and $\overline{\psi}_r = b^{-1}\sum_{i\in I_r}\psi_i$. Let $\nu_{r,-i}$ and $\zeta_{r,-i}$ be the centered measures of the block with observation i removed. Taylor expansion between that block and the complete block gives

$$(b-1)\{\overline{s}_b - \overline{s}_{b-1}^{(-i)}\} = \psi_i - \overline{\psi}_r + e_{ri}.$$

The preceding derivative bounds and the explicit diagonal-deletion expression yield

$$\|e_{ri}\|_2 \le C\Big[(\|\nu_{r,-i}\|_* + \|\zeta_{r,-i}\|_*)(1+\|V_i\|_2) + \frac{1}{b}\Big\{1 + \|V_i\|_2^2 + \frac{1}{b}\sum_{j\in I_r}\|V_j\|_2^2\Big\}\Big].$$

The bounds hold simultaneously because $\max_i\|V_i\|_2^2/N = o_{\mathbb{P}}(1)$ under the second-moment assumption and $N/b^2 = R/b \to 0$. Thus every leave-one-out measure remains in the fixed neighborhood.

Leave-one-out independence from V_i bounds the first term's average square by $O_{\mathbb{P}}(b^{-1})$. Jensen's inequality bounds the second by $C(Nb^2)^{-1}\sum_i(1+\|V_i\|_2^4)$. Finite second moments suffice because

$$\frac{1}{Nb^2}\sum_{i=1}^N\|V_i\|_2^4 \le \frac{\max_i\|V_i\|_2^2}{b^2}\frac{1}{N}\sum_{i=1}^N\|V_i\|_2^2 = o_{\mathbb{P}}(1).$$

It follows that $N^{-1}\sum_{r,i}\|e_{ri}\|_2^2 = o_{\mathbb{P}}(1)$. Uniform metric derivatives give the same conclusion at the fitted arguments: the additional leading term is bounded by $C\|\widehat{a}_{\mathrm{obs}} - a_0\|_2(1 + \|V_i\|_2)$, and the bandwidth is correct with probability tending to one. Finally,

$$\frac{1}{N}\sum_i \psi_i\psi_i^\top \longrightarrow \Omega_0, \qquad \frac{1}{R}\sum_r \|\overline{\psi}_r\|_2^2 = O_{\mathbb{P}}(b^{-1}).$$

Centering the pseudo-values and applying Cauchy–Schwarz proves $\widehat{\Omega}_0 \to \Omega_0$ in probability. The argument is unchanged when R is fixed. This completes the proof. □

Boundary inference can be implemented by simulating the cone limit. Let $\lambda_{0,b}$ and $\mu^-_{0,b} \in \mathbb{R}^m_+$ satisfy the population Karush–Kuhn–Tucker equations

$$g_{0,b} + \lambda_{0,b}\mathbf{1} - \mu^-_{0,b} = 0,$$

with the usual complementary slackness conditions. A boundary coordinate is strongly active when its corresponding multiplier is positive. On such a coordinate, every direction in $\mathcal{C}_{0,b}$ equals zero.

Choose $t_R, s_R \to 0$ with $\sqrt{R}t_R \to \infty$ and $\sqrt{R}s_R \to \infty$, and define

$$\widehat{A}^-_{0,b} = \{j : \widehat{a}_j \le \delta + t_R\}.$$

Estimate the multipliers from $\overline{g}$ and the estimated free coordinates by the formulas in the proof below. Classify an estimated lower constraint as strongly active when its multiplier exceeds s_R. In cone minimization, replace eigenvalues of the estimated Hessian below R^{-1} by R^{-1} on the span of the estimated cone. The following proposition proves consistency of the estimated constraint geometry and of the resulting distribution.

Proposition 40. *Under Theorem 13, the active set and the set of strongly active coordinates are consistently estimated. The conditional distribution obtained by drawing $Z^* \sim N(0, \widehat{\Sigma}_{0,b})$ and minimizing*

$$(Z^*)^\top v + \frac{1}{2}v^\top \widehat{H}_{0,b}v$$

over the estimated critical cone

$$\left\{\ v : \mathbf{1}^\top v = 0, v_j = 0 \text{ for every strongly active } j,\ v_j \ge 0 \text{ for the remaining } j \in \widehat{A}^-_{0,b}\ \right\}$$

converges to the limit in Theorem 13.

Proposition 40 estimates the limiting distribution at a boundary. Confidence sets based on a limiting quantile require the distribution function to be continuous and strictly increasing at that quantile. Distributional convergence alone does not justify a zero-radius confidence set when the limit is a point mass. A weakly active coordinate permits inward variation, while a strongly active coordinate is fixed to first order.

Proof of Proposition 40. Theorem 13 gives $\widehat{a}_j - a_{0,b,j} = O_{\mathbb{P}}(R^{-1/2})$. If $j \in A^-_{0,b}$, this order and $\sqrt{R}t_R \to \infty$ imply $j \in \widehat{A}^-_{0,b}$ with probability tending to one. A coordinate outside $A^-_{0,b}$ stays above $\delta + t_R$ with probability tending to one. At least one coordinate of $a_{0,b}$ exceeds δ because its coordinates sum to $m > m\delta$. Thus the trace gradient and the active coordinate gradients are linearly independent. Let F denote the nonempty free-coordinate set. The multiplier formulas are $\lambda_{0,b} = -|F|^{-1}\sum_{j\in F} g_{0,b,j}$ and $\mu^-_{0,b,j} = g_{0,b,j} + \lambda_{0,b}$ on the active set. The finite-dimensional score bound and the parameter rate give $O_{\mathbb{P}}(R^{-1/2})$ errors in these formulas with the empirical score. The finite set of positive multipliers has positive minimum when nonempty. The conditions on s_R separate it from zero multipliers consistently. The Hessian and covariance argument in the proof of Theorem 14 applies without an interior restriction. With probability tending to one, the cones agree, the Hessian curvature is at least $\kappa/2$ on their span, and truncation is inactive. The separation inequality in Theorem 13 gives continuity even for singular Gaussian scores. Covariance convergence and continuous mapping prove the conditional limit. This completes the proof. □

The same construction applies to the population target when b_R satisfies the joint-limit condition.

Proposition 41. *Under Theorem 15, replace $a_{0,b}$, $H_{0,b}$, and $\Sigma_{0,b}$ in Proposition 40 by a_0, H_0, and Σ_0. The resulting conditional cone distribution converges to the limit in Theorem 15.*

Proposition 41 gives a feasible boundary approximation centered at the estimand in (10).

Proof of Proposition 41. Theorem 15 gives $\widehat{a} - a_0 = O_{\mathbb{P}}(R^{-1/2})$. Lemma 7, the condition $Rm\log(mb_R)/b_R \to 0$, and convergence of the empirical averages give consistent population score, Hessian, and covariance estimators. The active-set, multiplier, and quadratic-program arguments in the proof of Proposition 40 now apply with the population quantities. This completes the proof. □

The observation-level limit uses the same boundary geometry with a different normalization and covariance. Use the Hessian of the mask-averaged empirical criterion and the exact mask-averaged jackknife estimator $\widehat{\Omega}_0$ from Proposition 34. The jackknife uses exact mask expectations even when Monte Carlo averaging is used to fit the point estimator.

Proposition 42. *Under Theorem 16, in Proposition 40, replace R by Rb_R, $a_{0,b}$ by a_0, $H_{0,b}$ by H_0, and $\Sigma_{0,b}$ by Ω_0. The resulting conditional cone distribution converges to the limit in Theorem 16.*

Proposition 42 gives an observation-level distributional approximation for boundary signal and nuisance weights. Its use for confidence sets requires the quantile continuity stated after Proposition 40.

Proof of Proposition 42. Theorem 16 gives $\widehat{a}_{\mathrm{obs}} - a_0 = O_{\mathbb{P}}\{(Rb_R)^{-1/2}\}$. Lemma 9 gives consistent score and Hessian estimators, and Proposition 34 gives the covariance estimator. The active-set and multiplier arguments in Proposition 40 apply at the $\sqrt{Rb_R}$ normalization. Continuity of the quadratic-program solution proves the conditional convergence. This completes the proof. □

Proof of Theorem 17. Lemma 4 gives dimension-free directional derivative bounds for every fixed block size. Vector Bernstein bounds and a net of the unit sphere give, on a fixed neighborhood of $a_{0,b}$,

$$\sup_a \|\nabla_a \widehat{\mathcal{R}}_\rho(a, h_{0,b}) - \nabla_a \mathcal{R}_{b,\rho}(a, h_{0,b})\|_2 = O_{\mathbb{P}}\Big\{\sqrt{\frac{m\log(mR)}{R}}\Big\}.$$

The same argument for the Hessian gives uniform convergence on this neighborhood when $m\log(mR)/R \to 0$. The criterion error and η_R are $o_{\mathbb{P}}(\kappa_m)$. Excess risk is at most twice that error plus η_R. The bandwidth gap gives $\widehat{h} = h_{0,b}$ with probability tending to one; separation by $c\kappa_m$ places $\widehat{a}$ in the radius-r_0 neighborhood. Put $d = \widehat{a} - a_{0,b}$. The population quadratic growth and the approximate minimizing inequality yield

$$\kappa_m \|d\|_2^2 \le \sup_a \|\nabla_a \widehat{\mathcal{R}}_\rho(a, h_{0,b}) - \nabla_a \mathcal{R}_{b,\rho}(a, h_{0,b})\|_2 \|d\|_2 + \eta_R.$$

Solving this quadratic inequality proves the stated rate. This completes the proof. □

Proof of Proposition 35. For $a = \delta\mathbf{1} + m(1-\delta)\theta$ with $\theta \in \Delta_S$, the product decomposition gives

$$\widetilde{\mathcal{R}}_m(\theta, \tau) = C_m + \sum_{j\in S} \widetilde{r}_{j,m}(\theta_j, \tau) + \frac{\bar{\rho}}{2m^2} \sum_{j\in S} \{\delta + m(1-\delta)\theta_j - 1\}^2,$$

where C_m does not depend on θ or τ. Uniform C^2 convergence gives

$$\max_{\tau\in\mathcal{T}} \|\widetilde{\mathcal{R}}_m(\cdot, \tau) - C_m - \widetilde{F}_\tau(\cdot)\|_{C^2(\Delta_S)} \to 0.$$

The limiting risk separation and compactness imply τ_0 selection and $\theta_{0,m} \to \theta_\infty$. The positive active coordinates and the tangent curvature over the finite collection $\{\widetilde{F}_\tau : \tau \in \mathcal{T}\}$ then give the first three assertions of Condition 4 for all sufficiently large m.

At $a = \mathbf{1}$, the corresponding simplex coordinate equals m^{-1}. For $j \in S$,

$$d_j(\sqrt{m}\,\tau) = -m\omega_j r'_{j,m}(1, \sqrt{m}\,\tau) = -\frac{\partial_t \widetilde{r}_{j,m}(m^{-1}, \tau)}{1-\delta},$$

while $d_j(\sqrt{m}\,\tau) = 0$ for $j \notin S$. Uniform C^1 convergence and the strict last inequality in Proposition 35 give the score separation.

The ridge gradient equals zero at $a = \mathbf{1}$. Hence, for $j \notin S$ and $\ell \in S$,

$$m\left\{\frac{\partial \mathcal{R}_{\infty,\rho_m}(\mathbf{1}, \sqrt{m}\,\tau)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho_m}(\mathbf{1}, \sqrt{m}\,\tau)}{\partial a_j}\right\} = \frac{\partial_t \widetilde{r}_{\ell,m}(m^{-1}, \tau)}{1-\delta}.$$

Uniform convergence and the same strict inequality give the exchange margin for some fixed $\chi > 0$. This proves Condition 4. This completes the proof. □

Proof of Corollary 18. The tensor-product calculation in Proposition 28 gives $\widetilde{r}_{j,m}(t, \tau) = g_{\tau,m}(t)$ for every $j \in S$. The arguments of the hyperbolic tangent and their first two derivatives converge uniformly on $[0, 1] \times \mathcal{T}$. Hence

$$\max_{\tau \in \mathcal{T}} \|g_{\tau,m} - g_\tau\|_{C^2[0,1]} \to 0.$$

For $x \geq 0$, put $u = \tanh x$. Direct differentiation gives

$$\frac{d}{dx}(1-u)^2 = -2(1-u)(1-u^2) < 0, \quad \frac{d^2}{dx^2}(1-u)^2 = 2(1-u^2)(1-u)(1+3u) > 0.$$

The arguments of the limiting curves range over a fixed compact interval. Thus their second derivatives with respect to t have a positive uniform lower bound. Strict convexity and permutation symmetry on Δ_S give the unique minimizer $\theta_\infty = s^{-1}\mathbf{1}_S$. Each limiting curve is strictly increasing in τ at every $t > 0$. Because every point of Δ_S has a positive coordinate, the finite collection has the unique minimizing bandwidth $\min \mathcal{T}$ and a positive gap. Uniform C^2 convergence gives the same minimizer and the required tangent curvature for all sufficiently large m.

At the origin,

$$g'_\tau(0) = -\frac{2(1-\delta)}{s\tau^2}.$$

The strict inequality in the corollary and uniform C^1 convergence give the score separation and the one-box exchange margin in Condition 4. The variables are uniformly bounded, and the deterministic observed view has $\pi = 1$. Condition 1 follows. This completes the proof. □

Proof of Proposition 36. For the paired mask, $\omega_{j,m} = m^{-1}$. For the uniform fixed-count mask,

$$\omega_{j,m} = \frac{k}{m(m-1)} = O(m^{-1}).$$

The separable risk is

$$Q_h\{a_S(\theta)\} = \sum_{j \in S} \omega_{j,m} r_j\{\delta + m(1-\delta)\theta_j, h\}.$$

Because s is fixed, the asserted derivative bounds and the mean value theorem give the risk, gradient, and Hessian orders after multiplication by $\omega_{j,m}$. For a signal coordinate,

$$|d_j(h)| = m\omega_{j,m}|r'_j(1, h)|.$$

The derivative with respect to a_j equals the derivative with respect to θ_j divided by $m(1-\delta)$, and hence $|r'_j(1, h)| \leq C/m$. This proves the score order. The active ridge Hessian under the change of variables is

$$\{m(1-\delta)\}^2 \frac{\rho_m}{m} I_s = \rho_m m(1-\delta)^2 I_s.$$

Substitution of $\rho_m = \bar{\rho}/m$ proves the final assertion. This completes the proof. □

Proof of Theorem 18. Write

$$\widehat{\widetilde{\mathcal{R}}}_m(\theta, \tau) = \widehat{\mathcal{R}}_{\rho_m}\{\delta\mathbf{1} + m(1-\delta)\theta, \sqrt{m}\,\tau\}, \qquad \widetilde{\mathcal{R}}_{b,m}(\theta, \tau) = \mathcal{R}_{b,\rho_m}\{\delta\mathbf{1} + m(1-\delta)\theta, \sqrt{m}\,\tau\}.$$

The stated orders and Proposition 15 give

$$\max_{h \in \mathcal{H}_m} \max_{1 \leq j \leq m} |\widehat{d}_j(h) - d_j(h)| = o_{\mathbb{P}}(1).$$

The left side is $O_{\mathbb{P}}(e_{R,b,m}) = o_{\mathbb{P}}(\lambda_R)$. Condition 4 then implies

$$\mathbb{P}\{\widehat{S}_{\mathrm{cand}}(h) = S \text{ for every } h \in \mathcal{H}_m\} \to 1.$$

On the same event, Proposition 18 verifies the exchange certificate for every bandwidth. Proposition 16 then shows that each unrestricted empirical global minimum lies on Δ_S.

The coordinate bounds in Lemma 4 imply that every active θ-derivative of one block loss is uniformly bounded. Because s and $|\mathcal{T}|$ are fixed, the entropy-integral maximal inequality gives

$$\sup_{\theta\in\Delta_S,\,\tau\in\mathcal{T}} \|\nabla_\theta \widehat{\widetilde{\mathcal{R}}}_m(\theta,\tau) - \nabla_\theta \widetilde{\mathcal{R}}_{b,m}(\theta,\tau)\|_2 = O_{\mathbb{P}}(R^{-1/2}).$$

The coordinatewise derivative argument in Proposition 31, multiplied by $m(1-\delta)$ on the fixed active set, gives

$$\sup_{\theta\in\Delta_S,\,\tau\in\mathcal{T}} \|\nabla_\theta \widetilde{\mathcal{R}}_{b,m}(\theta,\tau) - \nabla_\theta \widetilde{\mathcal{R}}_m(\theta,\tau)\|_2 \le C\sqrt{\frac{m\log(mb)}{b}}.$$

The same argument after one more derivative gives

$$\sup_{\theta\in\Delta_S,\,\tau\in\mathcal{T}} \|\nabla^2_\theta \widehat{\widetilde{\mathcal{R}}}_m(\theta,\tau) - \nabla^2_\theta \widetilde{\mathcal{R}}_m(\theta,\tau)\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Condition 4 then gives tangent curvature at least $\kappa_\theta/2$ for every restricted empirical criterion with probability tending to one. The corresponding Hessian upper bound gives $L_{\theta,R} = O_{\mathbb{P}}(1)$ uniformly over $\mathcal{T}$. The third-derivative bound gives a deterministic Hessian Lipschitz constant $L_{3,\theta}$. A regular covering of Δ_S with radius $\kappa_\theta/(8L_{3,\theta})$ has at most $C_s(L_{3,\theta}/\kappa_\theta)^{s-1}$ cells. Interval Hessian bounds on these cells certify curvature $\kappa_\theta/4$ with probability tending to one. This cell count is constant in m, R, and b and is absorbed into the operation constant. The corresponding criterion bounds and the bandwidth gap give $\mathbb{P}(\widehat{h} = \sqrt{m}\,\tau_0) \to 1$.

On this event, put $d = \widehat{\theta} - \theta_{0,m}$. Population optimality gives $\nabla_\theta \widetilde{\mathcal{R}}_m(\theta_{0,m},\tau_0)^\top d \ge 0$. Integrating the gradient difference and using tangent curvature and the objective gap gives

$$\kappa_\theta \|d\|_2^2 \le C\Big\{R^{-1/2} + \sqrt{\frac{m\log(mb)}{b}}\Big\}\|d\|_2 + \eta_R.$$

Because $\eta_R = O_{\mathbb{P}}(\bar{\eta}_R)$, solving the quadratic inequality proves the θ rate.

The restricted computation fixes all coordinates outside S at δ, also for approximate minimizers. Certification alone guarantees this only for exact unrestricted minimizers. Condition 4 and the θ rate give

$$\min_{j\in S} \widehat{\theta}_j \ge c_\theta/2$$

with probability tending to one. Hence $\widehat{a}_j - \delta \ge m(1-\delta)c_\theta/2 > t_N$ for $j \in S$ and all sufficiently large m, while $\widehat{a}_j - \delta = 0$ for $j \notin S$. This proves $\mathbb{P}(\widehat{S} = S) \to 1$.

On the certified support, the active representation weights satisfy

$$\widehat{a}_j^{\mathrm{rep}} = \frac{s\{\delta + m(1-\delta)\widehat{\theta}_j\}}{m(1-\delta) + s\delta}, \qquad j \in S.$$

The derivative of this map with respect to θ_S is bounded, and

$$\max_{j\in S}\left|\frac{s\{\delta + m(1-\delta)\theta_{0,m,j}\}}{m(1-\delta)+s\delta} - s\theta_{0,m,j}\right| \le C/m.$$

The triangle inequality gives the representation-metric rate.

On the restricted simplex, the projected-gradient calculation in Proposition 19 applies with curvature $\kappa_\theta/2$ and gradient Lipschitz constant $L_{\theta,R}$. Its linear contraction gives the stated iteration count. The full-dimensional score and one-box calculations cost $O\{|\mathcal{T}|R(b^3+b^2m)+|\mathcal{T}|m^2\}$. Each restricted gradient step costs $O\{R(b^3+b^2m)\}$ for each bandwidth. Adding these terms proves the operation bound.

For finite-precision projected gradients on the certified face, let $\kappa_* > 0$ and L_* be its certified curvature and smoothness bounds. Because s is fixed, a scaled-score error at most ϵ gives a θ-gradient error at most $C\epsilon$. The usual quadratic upper bound with the feasible comparison point $\theta + (\kappa_*/L_*)(\theta_* - \theta)$ gives

$$F(\theta^{(t+1)}) - F(\theta_*) \le (1 - \kappa_*/L_*)\{F(\theta^{(t)}) - F(\theta_*)\} + C\epsilon.$$

The initial gap is at most $4B^2 + \bar{\rho}/2$. Iterate the inequality with $\epsilon \le c\eta_R\kappa_*/L_*$. Stop when the bound plus evaluation-interval widths is at most η_R. The residual error is at most $\eta_R/2$, and $C(L_*/\kappa_*)\log(C/\eta_R)$ iterations suffice. The scaling calculation in Proposition 17 supplies the additional $\log(b/\epsilon)$ factor. Twice-differentiated scaling equations give the active Hessian entries with the same inverse bound and cost, up to a constant depending on s. Choosing $\epsilon = o(\lambda_R)$ preserves screening for the same criterion. This completes the proof. □

Proof of Corollary 19. For any fixed $C > 0$, the event $\|\widehat{a} - a_0\|_\infty \le Cr_N$ has probability tending to one as $C \to \infty$. On this event, $Cr_N < t_N$ for all sufficiently large N. If $j \notin S_0$, then $a_{0j} = \delta$ and $\widehat{a}_j < \delta + t_N$. If $j \in S_0$, then $\widehat{a}_j > \delta + \beta_0 - Cr_N > \delta + t_N$. Hence $\widehat{S} = S_0$.

On this event, the map

$$a_{S_0} \longmapsto \frac{|S_0| a_{S_0}}{\mathbf{1}^\top a_{S_0}}$$

has a uniformly bounded derivative because its denominator is at least $|S_0|\delta$ and $|S_0|$ is fixed. On the selected coordinates, $\|\widehat{a}_{S_0} - a_{0,S_0}\|_2 \le \sqrt{|S_0|}\|\widehat{a} - a_0\|_\infty$. Integration of the derivative along the line segment gives the representation-metric rate. This completes the proof. □

We now prove the spectral results. All continuum forms act on the common Hilbert space $L^2(P)$ because the sampling measure does not depend on the selected metric.

Proof of Proposition 37. The Dirichlet form of (11) is

$$\mathfrak{e}_w(f, g) = \frac{\sigma_\varphi}{2} \int_{\mathcal{M}} p_w^{-2/d} \langle \nabla_{g_w} f, \nabla_{g_w} g \rangle_{g_w} \, dP.$$

In a fixed atlas, its coefficient matrix is a C^4 function of (w, z). Condition 3 and the positive lower and upper bounds for the selected weights give uniform ellipticity and

$$|\mathfrak{e}_w(f, g) - \mathfrak{e}_{w'}(f, g)| \le C\|w - w'\|_2 \|f\|_{H^1} \|g\|_{H^1}.$$

The min–max characterization and uniform equivalence of the H^1 norms imply

$$\max_{1 \le j \le p+2} |\lambda_j(w) - \lambda_j(w')| \le C_p \|w - w'\|_2.$$

Let f be a unit eigenfunction of $\mathcal{L}_w$ with eigenvalue among the first p nonconstant eigenvalues. Uniform elliptic regularity gives $\|f\|_{H^2} \le C_p$. The coordinate expression of the two operators gives

$$\|(\mathcal{L}_{a_0^{\mathrm{rep}}} - \mathcal{L}_w) f\|_{L^2(P)} \le C_p \|w - a_0^{\mathrm{rep}}\|_2.$$

Expand f in an orthonormal eigenbasis of $\mathcal{L}_{a_0^{\mathrm{rep}}}$. The eigenvalue bound and $\gamma_p > 0$ bound the squared coefficients outside $\mathcal{E}_p(a_0^{\mathrm{rep}})$ by $\frac{C_p^2}{\gamma_p^2}\|w - a_0^{\mathrm{rep}}\|_2^2$. The same calculation with the two metrics exchanged gives the projection inequality. This completes the proof. □

Let

$$\mathcal{W}_{S_0} = \Big\{ w : w_j = 0 \text{ for } j \notin S_0, \quad \delta/A_m \le w_j \le A_m/\delta \text{ for } j \in S_0, \quad \sum_{j \in S_0} w_j = |S_0| \Big\}.$$

Let $\lambda_{j,n}(w)$ be the eigenvalues of $L_{n,k}(w)$. Let $Y_{n,p}(w)$ contain Euclidean-normalized graph eigenvectors numbered 2 through $p+1$. For any orthonormal basis of $\mathcal{E}_p(w)$, let Φ_w be its $n \times p$ evaluation matrix. Let $r_i(w)$ be the w-distance from X_i to its k_nth nearest neighbor.

For differentiable self-tuned affinities, Cheng et al. (2027) obtain faster pointwise graph-Laplacian convergence rates. Their smooth bandwidth dependence differs from the maximum of two neighbor radii used here. The following lemma makes the fixed-metric spectral probability uniform over a finite collection of metrics. Its constants are independent of the collection.

Put

$$\zeta_{n,k} = (k_n/n)^{1/d}.$$

Lemma 10. *Suppose Condition 3 holds. There are constants $c, C, C_0 > 0$ such that, for every finite $\mathcal{N} \subset \mathcal{W}_{S_0}$,*

$$\mathbb{P}\Big\{ \max_{w\in\mathcal{N}} \max_{1\le j\le p+2} |\lambda_{j,n}(w) - \lambda_j(w)| > C_p \zeta_{n,k} \Big\} \le C|\mathcal{N}| n^{C_0} \exp\{-ck_n\zeta_{n,k}^4\}.$$

If the continuum gap is at least $\gamma > 0$ on $\mathcal{N}$, the same bound holds for

$$\max_{w\in\mathcal{N}} \inf_{O^\top O = I_p} \|Y_{n,p}(w) - n^{-1/2}\Phi_w O\|_F > C_p\zeta_{n,k}/\gamma.$$

If $k_n/\log n \to \infty$ and $k_n/n \to 0$, then, with probability tending to one,

$$c\zeta_{n,k} \le r_i(w) \le C\zeta_{n,k}, \qquad 1 \le i \le n, \quad w \in \mathcal{W}_{S_0}.$$

Lemma 10 separates the published fixed-metric spectral estimate from the finite covering argument needed for a learned metric. The last assertion is simultaneous over the complete metric family and supplies the scale used in the perturbation calculation.

Proof of Lemma 10. Write $s = |S_0|$ and $r = \zeta_{n,k}$. The maps F_w have singular values in a fixed positive interval. In a fixed atlas, continuity of g_w, its derivatives through order six, and its inverse on compact $\mathcal{W}_{S_0} \times \mathcal{M}$ gives uniform ellipticity, curvature, chart, and injectivity-radius bounds. The Jacobian of F_w is bounded above and away from zero, giving uniform bounds for p_w and its derivatives. The graph normalization in equation (2.9) of Calder & García Trillos (2022) equals $L_{n,k}(w)$, and its limiting operator equals $\mathcal{L}_w$ because the factor σ_φ is included in (11).

Orientability is not needed for the cited scalar estimates. In local coordinates, $dV_g = \sqrt{\det g}\, dx$ transforms with the absolute Jacobian. A partition of unity gives scalar integration by parts because the partition derivatives sum to zero. The cell construction in Proposition 2.12 and the normal-coordinate calculation in Theorem 3.7 of Calder & García Trillos (2022) use this positive measure. Their Lemma 4.7 uses invariance under geodesic flow. Metric duality identifies the product of base and tangent volume measures with Liouville measure on the cotangent bundle. Hamilton's equations preserve this measure because their coordinate divergence equals zero by cancellation of mixed derivatives. Thus the energy comparisons and min–max argument apply without orientation, with constants uniform over the same metric family.

Fix $w_* \in \mathcal{W}_{S_0}$ and choose $q \ge p+2$ at the end of the eigenvalue multiplicity containing $\lambda_{p+2}(w_*)$. Then $\lambda_q(w_*) < \lambda_{q+1}(w_*)$. The min–max bound in the proof of Proposition 37 gives a neighborhood on which this gap is at least $2g_* > 0$. In Theorem 2.5(1) of Calder & García Trillos (2022), choose the transport parameter $\widetilde{\delta} = c_* r$ and the density parameter $\theta = c_*$, with $c_* > 0$ sufficiently small. For small r, large k_n, and sufficiently small c_*, its conditions hold uniformly, its first $q+1$ eigenvalue errors are less than g_*, and failure probability is at most $Cn\exp(-ck_n)$. In particular,

$$\lambda_{q+1,n}(w) - \lambda_q(w) \ge g_*.$$

Let $f_1, \dots, f_q$ be an $L^2(P)$-orthonormal continuum eigenbasis for these first q eigenvalues, and put

$$U_{ij} = n^{-1/2} f_j(Z_i), \qquad \Lambda = \operatorname{diag}\{\lambda_1(w), \dots, \lambda_q(w)\}, \qquad A = L_{n,k}(w), \qquad E = AU - U\Lambda.$$

Uniform elliptic regularity bounds the C^3 norms of these eigenfunctions. Theorem 3.7 of Calder & García Trillos (2022), with its consistency parameter equal to a sufficiently large constant times r, gives $\|E\|_F \le C_q r$ except on a set of probability at most $Cqn\exp(-ck_n r^4)$. Hoeffding's inequality for the q^2 products $f_j f_l$ gives

$$\|U^\top U - I_q\|_{\mathrm{op}} \le C_q r$$

except on a set of probability at most $Cq^2\exp(-cnr^2)$.

Let $\Pi_{n,q}$ project onto the first q graph eigenvectors. Expanding each column of U in a graph eigenbasis and using the preceding gap gives

$$\|(I - \Pi_{n,q})U\|_F \le g_*^{-1}\|E\|_F \le C_q r.$$

Put $V = \Pi_{n,q}U$, $G = V^\top V$, and $Q = VG^{-1/2}$. Then

$$\|G - I_q\|_{\mathrm{op}} \le C_q r, \qquad AV = V\Lambda + \Pi_{n,q}E,$$

and Q is an orthonormal basis of the first q graph eigenspace. Hence

$$Q^\top AQ = G^{1/2}\Lambda G^{-1/2} + G^{-1/2}V^\top EG^{-1/2}, \quad \|Q^\top AQ - \Lambda\|_{\mathrm{op}} \le C_q r.$$

Weyl's inequality gives the bound through $p + 2$. A finite cover of $\mathcal{W}_{S_0}$ makes the indices and constants uniform. Each g_* separates a complete multiplicity group; internal gaps are unnecessary.

For the eigenspace assertion, use $U = n^{-1/2}\Phi_w$ with a continuum eigenbasis numbered 2 through $p+1$, and let Λ contain their eigenvalues. The preceding pointwise and empirical Gram bounds still hold. Let $\Pi_{n,p}$ project onto graph eigenvectors numbered 2 through $p + 1$. The eigenvalue bound and the two boundary gaps give, for sufficiently small r,

$$\|(I - \Pi_{n,p})U\|_F \le \frac{2}{\gamma}\|AU - U\Lambda\|_F \le \frac{C_p r}{\gamma}.$$

For $V = \Pi_{n,p}U$ and $G = V^\top V$,

$$\|G - I_p\|_{\mathrm{op}} \le C_p\{r + r^2/\gamma^2\}, \qquad \|VG^{-1/2} - U\|_F \le C_p r/\gamma.$$

Orthogonal alignment of $VG^{-1/2}$ gives the Frobenius bound for any orthonormal continuum basis, using only the two external gaps.

Because $nr^2 \ge k_n r^4$ for $r \le 1$, all three failure probabilities are bounded by $Cn^{C_0}\exp\{-ck_n\zeta_{n,k}^4\}$, and a union bound over $\mathcal{N}$ proves the first two assertions. If r exceeds the fixed geometric smallness threshold, the degree bound in the proof of Lemma 11 gives $\|A\|_{\mathrm{op}} \le Cr^{-2}$. This bound and uniform boundedness of the continuum eigenfunction evaluations give both error inequalities after increasing C_p. For the remaining bounded values of k_n, increasing C and C_0 makes the failure-probability bound at least one.

For $z \in \mathcal{M}$, $w \in \mathcal{W}_{S_0}$, and small t, uniform normal coordinates give

$$c_1 t^d \le P\{d_w(Z, z) \le t\} \le C_1 t^d.$$

The class of sets

$$\{z' : d_w(z, z') \le t\} : z \in \mathcal{M},\ w \in \mathcal{W}_{S_0},\ t > 0$$

has finite Vapnik–Chervonenkis dimension as a fixed-dimensional ellipsoid family restricted to $\phi_{S_0}(\mathcal{M})$. The relative Vapnik–Chervonenkis inequality and $k_n/\log n \to \infty$ give, with probability tending to one, uniform counts below and above k_n for balls of probabilities ck_n/n and Ck_n/n, respectively. The volume bounds imply

$$c\zeta_{n,k} \le r_i(w) \le C\zeta_{n,k}$$

for all sample centers and all w. This completes the proof. □

The next lemma gives a deterministic modulus for changing the learned metric. It includes changes in nearest-neighbor order and in the support of the compactly supported affinity.

Lemma 11. *On the radius event in Lemma 10, for all $w, w' \in \mathcal{W}_{S_0}$,*

$$|d_w(x_i, x_j) - d_{w'}(x_i, x_j)| \le C\|w - w'\|_2 d_w(x_i, x_j),$$
$$|r_i(w) - r_i(w')| \le C\|w - w'\|_2 r_i(w),$$

and

$$\|L_{n,k}(w) - L_{n,k}(w')\|_{\mathrm{op}} \le C\frac{\|w - w'\|_2}{\zeta_{n,k}^2}.$$

Lemma 11 quantifies the price of extending a fixed-metric theorem between covering points. A covering radius of order $\zeta_{n,k}^3$ produces a graph-operator error of order $\zeta_{n,k}$.

Proof of Lemma 11. Uniform lower and upper bounds for the active weights give

$$|d_w^2(x_i, x_j) - d_{w'}^2(x_i, x_j)| \le C\|w - w'\|_2 d_w^2(x_i, x_j).$$

The first inequality follows from the square-root identity, with a larger constant when $\|w - w'\|_2$ is not small. Apply the resulting two-sided bound to every pairwise distance and take the k_nth order statistic. This proves the radius inequality.

Put $r_{ij}(w) = r_i(w) \vee r_j(w)$. If either affinity at (i, j) is nonzero, the two distance-to-radius ratios are bounded by a fixed constant. The preceding two inequalities and the Lipschitz property of φ give

$$|A_{n,k,ij}(w) - A_{n,k,ij}(w')| \le C\|w - w'\|_2.$$

The bound also holds when one ratio crosses one because Lipschitz continuity and compact support imply $\varphi(1) = 0$.

Partition the unit sphere into finitely many cones of angle less than $\pi/3$. A cone with more than k_n incoming neighbors would give its farthest point at least k_n closer neighbors. Uniform equivalence of the diagonal norms on $\mathcal{W}_{S_0}$ permits one such partition. Each row of the union of the two graphs thus has at most $C_s k_n$ nonzero entries.

The absolute row sum of the difference between the two unscaled Laplacians is at most $Ck_n\|w - w'\|_2$. Both matrices are symmetric, and the scaling in the definition of $L_{n,k}$ gives

$$\|L_{n,k}(w) - L_{n,k}(w')\|_{\mathrm{op}} \le \frac{C}{n}\left(\frac{n}{k_n}\right)^{1+2/d} k_n\|w - w'\|_2 = C\frac{\|w - w'\|_2}{\zeta_{n,k}^2}.$$

This completes the proof. □

The final auxiliary lemma transfers continuum eigenspace perturbations to the empirical norm used by graph eigenvectors.

Lemma 12. *Suppose Condition 3 holds and the continuum gap is at least $\gamma > 0$ on a compact set $\mathcal{U} \subset \mathcal{W}_{S_0}$. The eigenspaces admit finitely many local $L^2(P)$-orthonormal frames for which*

$$\sup_{w \in \mathcal{U}} \|n^{-1}\Phi_w^\top \Phi_w - I_p\|_{\mathrm{op}} = O_{\mathbb{P}}\left\{\sqrt{\frac{\log n}{n}}\right\},$$

and, for $0 < \epsilon \le 1$,

$$\sup_{\substack{w, w' \in \mathcal{U} \\ \|w - w'\|_2 \le \epsilon}} \inf_{O^\top O = I_p} n^{-1/2}\|\Phi_w - \Phi_{w'} O\|_F = O_{\mathbb{P}}\left\{\frac{\epsilon}{\gamma} + \sqrt{\frac{\log n}{n}}\right\}.$$

Lemma 12 closes the normalization step between $L^2(P)$ eigenfunctions and their evaluations on the observations. It does not require an independent representation sample.

Proof of Lemma 12. On each neighborhood, choose a contour enclosing eigenvalues 2 through $p+1$, at distance at least $c\gamma$ from the spectrum. The resolvent identity and elliptic regularity give bounded C^2 projections of fixed smooth bases, with differences $C\|w - w'\|_2/\gamma$. Their polar factors give local orthonormal frames, uniformly bounded in C^2 and C/γ-Lipschitz in supremum norm, on a finite cover of $\mathcal{U}$.

Cover each parameter neighborhood and the unit sphere in $\mathbb{R}^p$ by n^{-2}-nets. Their combined cardinality is polynomial in n because p and $|S_0|$ are fixed. Hoeffding's inequality and a union bound give

$$\sup_f \left|\frac{1}{n}\sum_{i=1}^n f(Z_i)^2 - \|f\|_{L^2(P)}^2\right| = O_{\mathbb{P}}\left\{\sqrt{\frac{\log n}{n}}\right\}$$

over all unit linear combinations in the local frames. Applying this argument to frame products controls empirical inner products. The uniform supremum bound makes the n^{-2} error negligible.

Write $\phi_{w,j}$ for the jth function in the local orthonormal frame at w. The frame Lipschitz bound and orthogonal alignment give, within one neighborhood,

$$n^{-1/2}\|\Phi_w - \Phi_{w'}O\|_F \le \sqrt{p}\max_{1\le j\le p}\|\phi_{w,j} - \phi_{w',j}\|_\infty \le C\|w - w'\|_2/\gamma.$$

A Lebesgue number places sufficiently close pairs in one neighborhood. Bounded frames give the same inequality for the remaining pairs, after increasing C, without an additional sampling term. This completes the proof. □

Lemma 13. *Suppose Condition 3 holds and*

$$k_n/n \to 0, \qquad \frac{k_n^{1+4/d}}{n^{4/d}\log n} \to \infty.$$

For $\zeta_{n,k} = (k_n/n)^{1/d}$ and a constant C_p,

$$\mathbb{P}\Big\{\sup_{w\in\mathcal{W}_{S_0}}\max_{1\le j\le p+2}|\lambda_{j,n}(w) - \lambda_j(w)| > C_p\zeta_{n,k}\Big\} \to 0.$$

If $\mathcal{U} \subset \mathcal{W}_{S_0}$ and

$$\inf_{w\in\mathcal{U}}[\lambda_2(w) \wedge \{\lambda_{p+2}(w) - \lambda_{p+1}(w)\}] \ge \gamma > 0,$$

then

$$\mathbb{P}\Big\{\sup_{w\in\mathcal{U}}\inf_{O^\top O = I_p}\|Y_{n,p}(w) - n^{-1/2}\Phi_w O\|_F > C_p\zeta_{n,k}/\gamma\Big\} \to 0.$$

Lemma 13 is simultaneous in the selected metric. It applies to a metric computed from the same observations as the graph.

Proof of Lemma 13. Put $\epsilon_n = \zeta_{n,k}^3$. The compact polytope $\mathcal{W}_{S_0}$ has an ϵ_n-net $\mathcal{N}_n$ satisfying

$$|\mathcal{N}_n| \le (C/\epsilon_n)^{|S_0|-1}.$$

Lemma 10 and a union bound give a failure probability bounded by $C\epsilon_n^{-(|S_0|-1)}n^{C_0}\exp\{-ck_n\zeta_{n,k}^4\} = o(1)$, because $k_n\zeta_{n,k}^4/\log n \to \infty$. On the complementary event, the fixed-metric eigenvalue error is at most $C_p\zeta_{n,k}$ at every net point and the neighborhood radii satisfy the bounds in Lemma 10 over the full metric family.

For $w \in \mathcal{W}_{S_0}$, choose $w' \in \mathcal{N}_n$ with $\|w - w'\|_2 \le \epsilon_n$. Lemma 11 gives

$$\|L_{n,k}(w) - L_{n,k}(w')\|_{\mathrm{op}} \le C\epsilon_n/\zeta_{n,k}^2 = C\zeta_{n,k}.$$

The min–max bound in the proof of Proposition 37 gives $|\lambda_j(w) - \lambda_j(w')| \le C_p\epsilon_n$. Weyl's inequality and the triangle inequality prove the first assertion.

For the eigenspace assertion, construct the net inside $\mathcal{U}$. The fixed-metric graph gaps are at least $\gamma/2$ for all sufficiently large n. The Davis–Kahan theorem (Davis & Kahan, 1970), applied to the entire selected spectral space, and the preceding graph perturbation give

$$\inf_{O^\top O = I_p}\|Y_{n,p}(w) - Y_{n,p}(w')O\|_F \le C_p\zeta_{n,k}/\gamma.$$

Lemma 12 gives

$$\inf_{O^\top O = I_p} n^{-1/2}\|\Phi_w - \Phi_{w'}O\|_F = O_{\mathbb{P}}\left\{\epsilon_n/\gamma + \sqrt{\frac{\log n}{n}}\right\}$$

simultaneously over neighboring pairs. Because $d \ge 2$ and $k_n/\log n \to \infty$, the second term is $O(\zeta_{n,k})$. The fixed-metric estimate at w' and the triangle inequality prove the second assertion. This completes the proof. □

Proof of Theorem 19. On an event with probability tending to one, $\widehat{S} = S_0$ and $\widehat{a}^{\mathrm{rep}}$ lies in the neighborhood of a_0^{rep} in Proposition 37. Lemma 13 holds simultaneously for every metric in this neighborhood. This simultaneous event applies to $\widehat{a}^{\mathrm{rep}}$ without conditioning or sample independence and gives

$$\max_{1\le j\le p+2} |\widehat{\lambda}_j - \lambda_j(\widehat{a}^{\mathrm{rep}})| = O_{\mathbb{P}}(\zeta_{n,k})$$

and the corresponding empirical eigenspace error relative to $\mathcal{E}_p(\widehat{a}^{\mathrm{rep}})$.

Proposition 37 gives eigenvalue error $O_{\mathbb{P}}(\|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2)$ and projection error $O_{\mathbb{P}}(\|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2/\gamma_p)$. Choose bases by the polar factor of the two projection ranges. Lemma 12 bounds the empirical normalization error by $O_{\mathbb{P}}\{\sqrt{\log n/n}\} = O_{\mathbb{P}}(\zeta_{n,k})$ and the empirical distance between the two continuum eigenspaces by $O_{\mathbb{P}}\left\{\frac{\|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2}{\gamma_p} + \zeta_{n,k}\right\}$. The triangle inequality and orthogonal Procrustes minimization prove both rates. This completes the proof. □

Proof of Corollary 20. Condition 4 gives $\min_{j\in S} w_{\infty,j} \ge sc_\theta > 0$. Choose a fixed compact neighborhood $\mathcal{U}$ of w_∞ on which every active weight is at least $sc_\theta/2$ and the weights sum to s. Theorem 18 gives

$$\mathbb{P}(\widehat{a}^{\mathrm{rep}} \in \mathcal{U}) \to 1, \quad \|\widehat{a}^{\mathrm{rep}} - w_\infty\|_2 = O_{\mathbb{P}}\{r_{R,b,m} + m^{-1} + \|\theta_{0,m} - \theta_\infty\|_2\}.$$

All ellipticity, density, covering, and graph-radius constants in Lemmas 10–12 are uniform on $\mathcal{U}$ and do not depend on m. Their proof and Proposition 37, centered at w_∞, give the eigenvalue and eigenspace bounds with the metric error above. The simultaneous graph event includes a metric estimated from the representation observations. This completes the proof. □

Proof of Proposition 38. The sub-Gaussian tail and a union bound give

$$\epsilon_n = \max_{1\le i\le n} \|e_i\|_2 = O_{\mathbb{P}}(\sigma_n\sqrt{\log n}).$$

Every selected pairwise distance changes by at most $C\epsilon_n$, and the same bound holds for each k_nth order statistic. On the uniform radius event in Lemma 13, the clean radii are of order r_n. The condition on σ_n gives $\epsilon_n/r_n = o_{\mathbb{P}}(1)$, hence the noisy radii have the same order with probability tending to one. The Lipschitz property of φ gives an edge-weight perturbation bounded by $C\epsilon_n/r_n$.

In $\mathbb{R}^{|S_0|}$, cones of angle below $\pi/3$ each contain at most k_n incoming positive-weight neighbors. Thus a point receives at most Ck_n such weights, with C depending on $|S_0|$. Affinity and degree differences have absolute row sums at most $Ck_n\epsilon_n/r_n$. Multiplication by the scaling factor in $L_{n,k}$ gives

$$\|L_{n,k}^{\mathrm{noisy}}(w) - L_{n,k}^{\mathrm{clean}}(w)\|_{\mathrm{op}} \le C\frac{\epsilon_n}{r_n^3}$$

uniformly over $w \in \mathcal{W}_{S_0}$. Weyl's inequality and the gap-only sin-theta bound prove the two assertions. This completes the proof. □

Proof of Proposition 39. For the dimension-adapted order,

$$\frac{k_n^{1+4/d}}{n^{4/d}\log n} = \ell_n^{1+4/d}\{1 + o(1)\} \longrightarrow \infty.$$

Its ratio to n tends to zero when ℓ_n increases at the stated slow rate. Its graph radius satisfies

$$(k_n/n)^{1/d} = n^{-1/(d+4)}(\log n)^{1/(d+4)}\ell_n^{1/d}\{1 + o(1)\}.$$

For the dimension-free order,

$$\frac{k_n^{1+4/d}}{n^{4/d}\log n} = n^{2(d-2)/(3d)}(\log n)^{-2(d-2)/(3d)}\ell_n^{1+4/d}\{1 + o(1)\}.$$

This tends to infinity for $d = 2$ because $\ell_n \to \infty$, and the power of n is positive for $d > 2$. The condition $k_n/n \to 0$ follows from the prescribed growth of ℓ_n. This completes the proof. □

Proof of Corollary 21. Theorem 16 gives

$$\|\widehat{a}_{\mathrm{obs}} - a_0\|_\infty = O_{\mathbb{P}}\{(Rb_R)^{-1/2}\}.$$

The two threshold conditions and Corollary 19 imply

$$\mathbb{P}(\widehat{S}_{\mathrm{obs}} = S_0) \to 1, \qquad \|\widehat{a}^{\mathrm{rep}}_{\mathrm{obs}} - a^{\mathrm{rep}}_0\|_2 = O_{\mathbb{P}}\{(Rb_R)^{-1/2}\}.$$

The simultaneous metric event in Theorem 19 applies when the representation graph uses the same observations. Substitution of the last metric rate into its eigenspace bound proves the assertion. This completes the proof. □

I NUMERICAL STUDIES

The numerical studies distinguish representation quality from verification of a limit theorem. The bounded binary model gives an identified population metric, an explicit covariance, and a global optimization certificate. The higher-dimensional studies assess neighborhood preservation under independent nuisance variation. The CITE-seq analysis uses one measured view to define the information that the representation should preserve.

I.1 SYNTHETIC OBSERVATIONS AND COMMON EVALUATION

Each synthetic design uses 100 independently generated data sets with $n = 2048$ observations and 48 independent nuisance coordinates. In the torus design, U_i and V_i are independent uniform angles on $[0, 2\pi]$. Put

$$s_i = \sqrt{2}(\cos U_i, \sin U_i, \cos V_i, \sin V_i), \qquad X_i = (s_i, s_i, N_i),$$

where the coordinates of $N_i \in \mathbb{R}^{48}$ are independent uniform variables on $[-\sqrt{3}, \sqrt{3}]$. The repeated measurements supply information for predicting an excluded coordinate. The representation has dimension four, and its neighborhood quality is assessed against s_i.

In the surface design, U_i, V_i are independent uniform variables on $[-1, 1]$, and the six signal measurements are

$$\sqrt{3}U_i, \quad \sqrt{3}V_i, \quad \frac{\sqrt{45}}{2}(U_i^2 - 1/3), \quad \frac{\sqrt{45}}{2}(V_i^2 - 1/3), \quad 3U_iV_i, \quad 2\sin(\pi U_i)\cos(\pi V_i).$$

The same 48-coordinate nuisance distribution is used, and the representation has dimension two. Every feature has population mean zero and variance one in both designs. No estimated scaling constants or evaluation labels enter fitting. The surface has a boundary and is a prediction experiment; it is not an example of the closed-manifold limit theorem.

The 2048 metric observations are partitioned into 16 disjoint blocks of size 128. Each block uses two independent uniform masks containing exactly half of the coordinates. The ridge parameter is 0.0005, the lower bound is $\delta = 0.02$, and the prediction bandwidth is selected by the fitted criterion from $\sqrt{m}\{0.6, 0.9\}$. The representation threshold is $t_N = 0.5$. The graph uses $k_n = 24$ and the triangular profile $\varphi(t) \propto (1 - t)_+$. All graph representations use the same centered eigenproblem for $D - A$. The metric observations are reused in the final graph.

The comparisons are isotropic Laplacian Eigenmaps, principal component analysis, spectral self-supervised feature selection, and two self-supervised SLAPS adjacency generators. Two further fits remove feature selection or remove feature exclusion from the proposed procedure. The first uses every fitted weight in the graph. The second uses all coordinates to construct the prediction affinity and assesses prediction of the complete feature vector, still deleting the diagonal. These comparisons separate the roles of feature prediction and support selection.

Every representation column is centered and normalized to unit Euclidean norm before evaluation. Trustworthiness and continuity use ten neighbors and the latent evaluation coordinates (Venna & Kaski, 2001; Lee & Verleysen, 2009). Continuity reverses the arguments of trustworthiness. Classification uses four angular sectors on the torus and the four quadrants on the surface. A seven-nearest-neighbor classifier is evaluated by five-fold stratified splitting after every representation has

been fitted. No class label enters metric estimation, feature selection, or graph learning. The reference eigenspace uses the same graph construction with unit weights on the signal measurements and zero nuisance weights. Its Procrustes discrepancy is the Frobenius distance between the optimally aligned, column-normalized coordinate matrices divided by $\sqrt{p}$, where p is the representation dimension. It measures departure from that signal-only graph and does not identify the unknown population predictive metric.

Table 2: Synthetic neighborhood, classification, and eigenspace errors.

Method	Trustworthiness	Continuity	Accuracy	Eigenspace error
Surface				
PALE	0.9954 (0.0010)	0.9958 (0.0009)	0.972 (0.005)	0.160 (0.020)
PALE, all weights	0.9951 (0.0010)	0.9955 (0.0009)	0.971 (0.005)	0.163 (0.019)
Prediction, all features	0.9888 (0.0013)	0.9892 (0.0013)	0.925 (0.008)	0.244 (0.025)
Isotropic LE	0.8061 (0.0418)	0.8177 (0.0306)	0.676 (0.053)	0.828 (0.166)
PCA	0.5265 (0.0133)	0.5456 (0.0157)	0.301 (0.026)	1.275 (0.050)
SSFS	0.7415 (0.1696)	0.7759 (0.1740)	0.579 (0.241)	0.886 (0.379)
SLAPS-DAE FP	0.8158 (0.0280)	0.8453 (0.0137)	0.686 (0.037)	0.781 (0.086)
SLAPS-DAE MLP	0.7663 (0.0520)	0.7721 (0.0477)	0.625 (0.069)	0.740 (0.145)
Torus				
PALE	0.9992 (0.0022)	0.9992 (0.0005)	0.953 (0.006)	0.069 (0.056)
PALE, all weights	0.9993 (0.0004)	0.9992 (0.0003)	0.953 (0.006)	0.078 (0.047)
Prediction, all features	0.9975 (0.0001)	0.9972 (0.0002)	0.936 (0.006)	0.137 (0.007)
Isotropic LE	0.9650 (0.0189)	0.9543 (0.0080)	0.828 (0.045)	0.496 (0.127)
PCA	0.9963 (0.0004)	0.9963 (0.0004)	0.930 (0.007)	0.179 (0.008)
SSFS	0.5021 (0.0097)	0.5019 (0.0082)	0.252 (0.011)	1.389 (0.015)
SLAPS-DAE FP	0.9792 (0.0007)	0.9692 (0.0012)	0.861 (0.009)	0.372 (0.005)
SLAPS-DAE MLP	0.8246 (0.0423)	0.7869 (0.0456)	0.636 (0.084)	0.796 (0.097)

Table 2 gives mean trustworthiness 0.9954 on the surface and 0.9992 on the torus for the proposed method. Numbers in parentheses represent standard deviations. Both neighborhood criteria improve relative to isotropic Laplacian Eigenmaps, principal component analysis, SSFS, and the SLAPS denoising comparisons. The classification gains after excluding target features from their own distances are 0.0467 on the surface and 0.0177 on the torus. The corresponding errors relative to the reference eigenspace decrease from 0.244 to 0.160 and from 0.137 to 0.069.

The thresholded metric selects averages of 8.58 and 3.13 nuisance coordinates out of 48 on the surface and torus. It selects 4.75 of the six surface measurements and 7.99 of the eight torus measurements. These counts refer to the generating features; the population predictive support is not known in these two designs. They do not constitute a measurement of exact support recovery for an identified population metric. The fit without thresholding has similar neighborhood and classification accuracy, which separates the effect of estimating the distance from the additional thresholding step.

I.2 Close comparisons and numerical optimization

The SLAPS comparison uses the full-parameter and multilayer-perceptron adjacency generators of Fatemi et al. (2021). We use their two-layer graph denoising network and call the comparison SLAPS-DAE. The network has 32 hidden units and is trained for 200 epochs with Adam, learning rate 0.01, and weight decay 0.0005. Independent corruption masks omit each feature with probability $1/2$. Feature and adjacency dropout probabilities are 0.2 and 0.25. Every ten epochs, three separate corruption masks assess prediction error and select a checkpoint. Class labels do not select that checkpoint. The learned adjacency is symmetrized and passed to the common eigenmap solver. This procedure isolates the self-supervised adjacency objective; it is not the supervised SLAPS classifier.

Spectral self-supervised feature selection uses the adaptive eigenvector selection and XGBoost gain scores of Segal et al. (2024). It considers twice the representation dimension in the initial spectrum.

For each candidate eigenvector, 500 subsamples containing 95 percent of observations estimate the variability of normalized logistic-regression coefficients. The least variable eigenvectors supply the feature scores. The two one-dimensional medoids are computed exactly. Only the requested leading eigenvectors are computed; the normalized operator is unchanged. The final subset contains as many coordinates as the proposed method selected on the same observations. This count depends on the observations and not on the latent variables or evaluation labels. Its graph uses the common neighborhood rule and eigenproblem.

Sequential quadratic programming proposes feasible metric candidates. Initial values include the uniform metric and metrics concentrated on 2, $\lceil\sqrt{m}\rceil$, or $\lceil m/3\rceil$ coordinates ranked by the criterion gradient at the uniform metric. Every fit searches the full trace-constrained polytope. These initial values do not fix a support and are not a screening certificate. The largest iteration count is 120 per initial value in the synthetic and CITE-seq studies. Finite bandwidth candidates are compared through the same empirical objective.

Proposition 20 supplies global lower bounds independently of the candidate solver. The higher-dimensional fits use a fixed computational budget and report their achieved objective bounds. These bounds are too large to establish the asymptotic optimization orders; their performance comparisons do not claim such orders. The binary experiment below uses a verified global curvature bound and satisfies its prescribed statistical tolerance in every replication. All numerical summaries include the complete set of replications.

I.3 Interior and boundary inference

The model in Proposition 43 connects the data distribution, population target, limiting covariance, and implemented objective explicitly. We use $h = 1.4$, $\delta = 0.02$, and $\sigma = 0.4$. Let ρ_c be the ridge value for which the derivative of the population risk along the trace constraint equals zero at $a_1 = 2 - \delta$. The three ridge values are $\rho_c/2$, ρ_c, and $2\rho_c$, giving a strictly active boundary, a weak boundary, and an interior target. The block allocations are $(R, b) = (4, 1024)$, $(8, 4096)$, and $(8, 16384)$. Each of the nine settings uses 100 independently generated data sets. A deterministic observed view contains (Z, N), while the response is $Z + 0.4\varepsilon$.

Every fit uses objective tolerance $\eta_R = \{Rb\log(Rb)\}^{-1}$. The eight-state counts give the exact individual-observation criterion and the exact grouped leave-one-out covariance calculation. Proposition 44 controls the tangent Hessian between 8193 equally spaced points, including the errors in symmetric scaling and differentiated row equations. The scalar score equation or a boundary solution then gives a global objective gap by Proposition 19.

The active-coordinate threshold is $(Rb)^{-1/4}$, and the multiplier threshold is $0.1(Rb)^{-1/4}$. Interior intervals use the normal critical value 1.96. At the weak boundary, the absolute cone limit has critical value 1.645. Intervals are intersected with the trace-feasible interval. The strictly active case is evaluated by boundary-selection frequency, because its point-mass limit alone does not justify a nominal zero-width confidence interval.

Table 3: Boundary and interior inference.

Regime	N	Bias $\times 10^3$	Coverage	Length $\times 10^3$	Covariance error $\times 10^3$	Gap ratio
Strict boundary	4,096	0.00	–	–	0.283	0.00013
Strict boundary	32,768	0.00	–	–	0.075	0.00514
Strict boundary	131,072	0.00	–	–	0.020	0.09311
Weak boundary	4,096	-1.47	0.98	9.26	0.287	0.00013
Weak boundary	32,768	-0.47	0.98	3.17	0.079	0.00514
Weak boundary	131,072	-0.32	0.96	1.65	0.020	0.09311
Interior	4,096	1.47	0.94	16.73	0.213	0.00013
Interior	32,768	0.44	0.93	5.87	0.054	0.00516
Interior	131,072	0.11	0.96	2.93	0.014	0.09334

Table 3 reports coverage from 0.93 to 0.98 in the interior and weak-boundary settings. Each exact binomial 95 percent interval for a coverage estimate contains the nominal value 0.95. Covariance

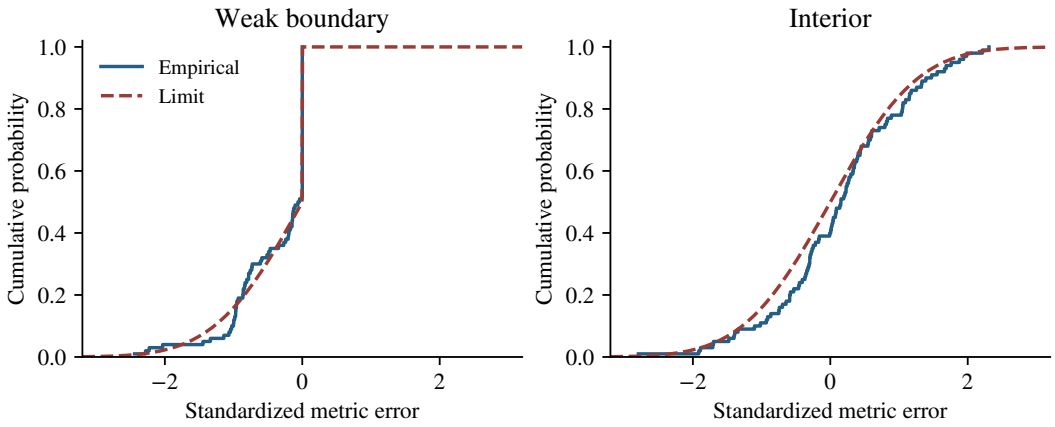

Figure 2: Metric-error distributions for $N = 131072$.

error decreases at the larger block size. Every strictly active fit selects the exact population boundary. The largest ratio of a certified objective gap to its required tolerance is below 0.094. Thus computation contributes less than the prescribed error allowance in all 900 data sets. Covariance error is the Frobenius distance from the analytic covariance, and interval length is measured on the first metric coordinate.

Figure 2 compares empirical cumulative distributions with the limits in Proposition 43. The weak boundary places positive probability at zero; the interior target has a continuous normal limit. The distinction changes the appropriate critical value and cannot be represented by an interior normal approximation at every target.

I.4 Inference under a second-moment response condition

We examine the inference conclusion of Corollary 17 when the response has a finite second moment and an infinite fourth moment. In the interior model of Proposition 43, replace the Rademacher response error by $T/\sqrt{3}$, where T has Student's t distribution with three degrees of freedom and is independent of the observed view. We use $V = Z + 0.4\varepsilon$, $h = 1.4$, $\delta = 0.02$, and $\rho = 2\rho_c$. The conditional response mean and variance agree with the Rademacher model. Thus both models have the same population metric, curvature, and score covariance. We use one block with $b = N$ and consider $N = 500$ and $N = 1000$. Each of the four settings uses 100 independently generated data sets.

The four-state calculation in Section K.2 evaluates the exact loss and all observation-level leave-one-out scores. Each fitted metric satisfies the global objective tolerance $1/(N \log N)$. We use the empirical Hessian and the covariance in Proposition 34 to construct two-sided 95 percent normal intervals for the interior metric. Table 4 reports the root mean squared error, denoted by RMSE, and the Monte Carlo standard deviation, denoted by MC SD, for the estimated first weight. SE denotes the estimated standard error, and MC SE is the Monte Carlo standard error of coverage.

Table 4 shows coverage of 0.97 and 0.94 under the Student response, with Monte Carlo standard errors 0.017 and 0.024. The median interval length decreases from 0.0457 to 0.0329 when N increases from 500 to 1000. The estimated standard errors follow the Monte Carlo variation under both response distributions. These results support the second-moment inference conclusion in this model.

I.5 CITE-seq scientific views

The CITE-seq cord-blood mononuclear-cell data contain paired RNA and antibody-derived protein measurements (Stoeckius et al., 2017). We use GEO accession GSE100866 and its CBMC 8K RNA and antibody count matrices. Their 8617 cell barcodes agree. We select the 8005 cells with at least 500 human RNA counts and at least 90 percent human counts among human and mouse RNA

Table 4: Metric estimation and interval coverage under bounded and unbounded responses.

Noise	N	RMSE	MC SD	Median SE	Coverage	MC SE	Median length
$t_3/\sqrt{3}$	500	0.0122	0.0122	0.0117	0.97	0.017	0.0457
$t_3/\sqrt{3}$	1000	0.0089	0.0089	0.0084	0.94	0.024	0.0329
Rademacher	500	0.0127	0.0126	0.0123	0.96	0.020	0.0484
Rademacher	1000	0.0088	0.0086	0.0086	0.93	0.026	0.0338

counts. A fixed random partition assigns 1024 cells to independent preprocessing, 2048 to metric fitting, 1024 to selection of tuning values, and 1024 to confirmation. These four sets are disjoint.

The scientific task is to construct a representation from 13 protein measurements that predicts a 128-gene RNA view. RNA counts are divided by each cell's total human count, multiplied by 10000, and transformed by $\log(1 + x)$. The RNA panel contains the 128 largest pilot variances among human genes observed in at least 30 pilot cells. Protein counts are transformed by $\log(1 + x)$ and centered by the within-cell mean of the transformed protein counts. Pilot means and standard deviations standardize both views, and all standardized values are clipped to $[-3, 3]$. No cell-type label or biological annotation enters these operations.

Metric fitting uses the protein view for distances and the RNA view for assessment. The 2048 fitting cells form 16 disjoint blocks of size 128 with a deterministic observed view. Candidate prediction scales are $h/\sqrt{13} \in \{0.6, 0.9, 1.2\}$. The ridge candidates are 0.005, 0.05, and 0.5, and the candidate representation dimensions are 5, 10, and 20. A ridge regression with penalty one predicts the RNA view from each representation. Separate validation-cell prediction errors select the ridge value and dimension for each method. The proposed method selects ridge 0.05 and dimension 20; row-normalized prediction selects ridge 0.5 and dimension 20. The latter comparison has the same metric constraints, response view, and initialization rules but replaces symmetric scaling by row normalization.

Isotropic Laplacian Eigenmaps, principal component analysis, partial least squares, and full-feature ridge regression provide additional comparisons. Partial least squares models the relation between the two measured views (Wold et al., 2001). Its selected dimension is ten, as is that of principal component analysis; isotropic Laplacian Eigenmaps uses dimension 20. Implementations of the linear representations, regressions, and classification evaluation use scikit-learn (Pedregosa et al., 2011). Every prediction head is fitted on the same 2048 training cells. The protein observations of the confirmation cells may enter graph construction; their RNA values enter only the final error calculation.

The complete metric fit is repeated under 100 independently drawn partitions of the same training observations into disjoint blocks. These repetitions quantify sensitivity to block assignment, not variation across independent biological samples. We report mean squared RNA prediction error on the fixed 1024 confirmation cells. Paired bootstrap resampling of these cells uses 1000 replications. The resulting uncertainty is conditional on this cohort and is not a claim of generalization across donors.

Table 5: Protein-to-RNA prediction in CITE-seq.

Method	MSE	Partition SD	Difference	95 percent interval
PALE	0.5532	0.00190	0.0000	–
PALE, all weights	0.5518	0.00013	-0.0014	[-0.0026, -0.0002]
Row-normalized prediction	0.5547	0.00002	0.0016	[-0.0005, 0.0038]
Isotropic LE	0.5560	–	0.0028	[0.0003, 0.0054]
PCA	0.5968	–	0.0436	[0.0393, 0.0482]
PLS	0.5919	–	0.0388	[0.0344, 0.0433]
Full-feature ridge	0.5869	–	0.0338	[0.0299, 0.0380]

Table 5 gives mean squared error 0.5532 for the proposed representation, compared with 0.5560 for isotropic Laplacian Eigenmaps, 0.5968 for principal component analysis, and 0.5919 for partial

least squares. The protein neighborhoods predict RNA variation that these linear representations do not represent as accurately. The mean number of selected protein measurements is 9.78 out of 13. Using all fitted weights gives error 0.5518, indicating that thresholding is not needed for the best prediction in this small observed view.

The difference column subtracts the proposed method's mean squared error from the comparison error. The paired 95 percent bootstrap interval for isotropic Laplacian Eigenmaps is $[0.0003, 0.0054]$. The corresponding interval for row-normalized prediction is $[-0.0005, 0.0038]$ and does not establish a difference between the two normalization rules. The error reductions relative to the three linear comparisons range from 0.0338 to 0.0436. Partition standard deviations summarize the 100 repeated fits; a dash denotes a comparison that does not use random blocks.

I.6 The continuous identified model

The sphere experiment connects the population minimizer in Theorem 8 to support recovery, spectral approximation, prediction, and certified computation. We use $s = 3$, $\sigma = 0.4$, and independent sphere blocks, with observed dimensions $m \in \{12, 48, 192\}$. Each of the six combinations of dimension and fitting sample size $N \in \{1024, 4096\}$ uses 100 independently generated data sets. An additional N observations are generated for evaluation. Their responses do not enter metric fitting or the prediction head.

The coordinates have population variance one; no estimated scaling is applied. The response is the first sphere block plus independent errors taking values -0.4 and 0.4 with equal probability. Every observed coordinate is available to the distance, while the response view is excluded. The ridge is $\rho_m = (3\tau^2 m)^{-1}$ with $\tau = 300$, and $\delta = 0.02$. The fixed prediction scale satisfies the explicit exchange and curvature inequalities. Because $|X_j| \le \sqrt{3}$ and $h^2 = m\tau^2$, its prediction kernel satisfies $\exp(-6/300^2) \le K_{ij} \le 1$ off the diagonal. This nearly constant kernel tests the identification and optimization guarantees for a linear response; the signal is also detectable by coordinate correlations. The final graph remains local: it uses $k_n = \lceil n^{3/4} \rceil$ for $n = 2N$, which satisfies the spectral rate condition for intrinsic dimension two as a sequence.

For $N = 1024$, the observations form 16 blocks of size 64; for $N = 4096$, they form 32 blocks of size 128. The score cutoff is $(3\tau^2)^{-1}$. The candidate support is determined from the observed gradient and is accepted only after the global exchange and curvature calculations in Proposition 8. Neither the generating coordinate indices nor the population weights enter this calculation. The returned point lies on the certified face and has a verified objective gap at most $1/(N \log N)$. The representation threshold is 0.5, and its dimension is three.

The representation graph contains the fitting and evaluation covariates. An intercept and least-squares coefficients are estimated using only the fitting responses. We measure squared error from the known conditional mean on the independent evaluation observations. The response mean is available only for evaluation. The oracle-support graph uses unit weights on the three generating coordinates. Isotropic Laplacian Eigenmaps use all observed coordinates, and PCA and PLS each use three components. PLS receives the same fitting responses as PALE.

Table 6: Support, metric, and spectral estimation in the continuous model.

N	m	Exact support	$\|\widehat{\theta} - \theta_0\|_2$	Spectral error	Gap/tolerance
1024	12	97/100	0.1939 (0.1021)	0.1732 (0.2111)	3.91e-06
1024	48	99/100	0.1874 (0.0958)	0.1657 (0.1938)	6.50e-06
1024	192	99/100	0.1792 (0.0976)	0.1448 (0.1743)	1.69e-05
4096	12	100/100	0.0927 (0.0461)	0.0452 (0.0115)	3.25e-05
4096	48	100/100	0.0981 (0.0471)	0.0463 (0.0123)	4.51e-05
4096	192	100/100	0.0974 (0.0513)	0.0466 (0.0134)	9.53e-05

Table 6 shows exact support recovery in 97–99 of the 100 samples at $N = 1024$ and in all 100 samples at $N = 4096$, at each dimension. Numbers in parentheses represent standard deviations. The spectral discrepancy is the Frobenius distance from $n^{-1/2}X^{(1)}$, minimized over orthogonal rotations and divided by $\sqrt{3}$. At $m = 192$, its mean decreases from 0.1448 to 0.0466. All 600

fits satisfy the requested global objective tolerance. The largest gap divided by that tolerance is 9.53×10^{-5}. A calculation using common population coordinate bounds and enlarged arithmetic errors also satisfies the tolerance in every sample.

Table 7: Conditional-mean prediction error in the continuous model.

N	m	PALE	Oracle support	Isotropic LE	PCA	PLS
1024	12	0.0463	0.0038	0.7327	0.7279	0.0107
		(0.1011)	(0.0013)	(0.0516)	(0.0548)	(0.0025)
1024	48	0.0419	0.0038	0.9292	0.9263	0.0502
		(0.0927)	(0.0011)	(0.0331)	(0.0341)	(0.0065)
1024	192	0.0338	0.0038	0.9856	0.9807	0.1838
		(0.0851)	(0.0011)	(0.0107)	(0.0125)	(0.0093)
4096	12	0.0021	0.0010	0.7383	0.7350	0.0026
		(0.0012)	(0.0002)	(0.0601)	(0.0554)	(0.0006)
4096	48	0.0023	0.0010	0.9344	0.9334	0.0127
		(0.0013)	(0.0002)	(0.0236)	(0.0248)	(0.0016)
4096	192	0.0023	0.0010	0.9829	0.9825	0.0527
		(0.0014)	(0.0002)	(0.0084)	(0.0081)	(0.0034)

Table 7 shows why the selected geometry matters for prediction. At $N = 4096$ and $m = 192$, conditional-mean error is 0.0023 for PALE, 0.9829 for isotropic eigenmaps, and 0.0527 for PLS. PALE error remains close across the three observed dimensions. PLS is more accurate at $N = 1024$ and $m = 12$, where the response depends linearly on a small observed vector; its error increases with the number of nuisance coordinates. The irreducible response variance is 0.16, which is not included in this conditional-mean error table.

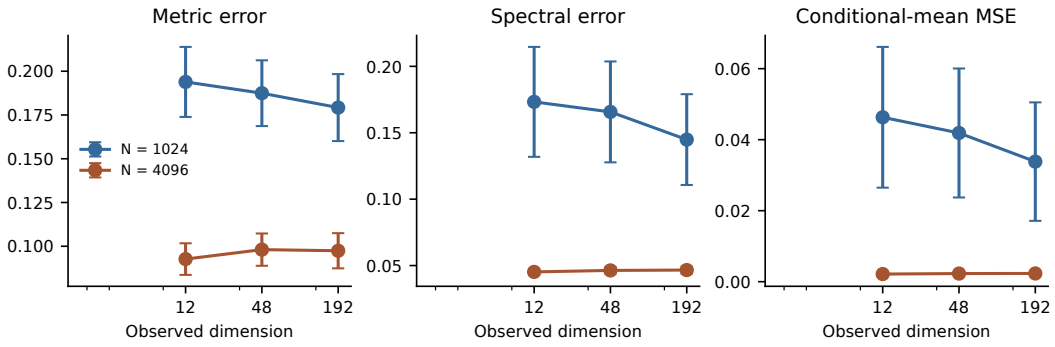

Figure 3: Metric, spectral, and prediction errors in the continuous model.

Figure 3 shows the decrease in estimation and prediction errors with sample size. Error bars are 1.96 Monte Carlo standard errors over the 100 independent samples. These experiments use a prescribed response view and do not assert the same support for uniform random coordinate masks.

I.7 Dependence with unchanged measurement marginals

We examine whether signal selection and prediction persist when a nuisance block becomes dependent on the signal. We use the dependent sphere model in Section A.1, with $s = 3$ and $m = 48$. Its density relative to the independent model is $1 + \epsilon X_1^{(1)} X_1^{(2)}/3$, where $\epsilon \in \{0, 0.25, 0.5, 0.75\}$. Exact rejection sampling preserves both sphere marginals. The response remains $V = X^{(1)} + \zeta$, where the coordinates of ζ independently take values -0.4 and 0.4 with equal probability. The covariance between the first coordinates of the signal and nuisance blocks is $\epsilon/3$, but the conditional response mean remains $X^{(1)}$.

Each combination of $N \in \{1024, 4096\}$ and ϵ uses 100 independently generated data sets. We use the estimator, graph, and least-squares decoder of Section I.6, with all tuning values unchanged. The

representation graph contains the N fitting and N evaluation covariates; evaluation responses enter neither the metric nor the decoder. The oracle-support and isotropic graphs use the same graph rule and decoder. Three-component PLS uses the same fitting responses. The generating support is used only by the explicitly identified oracle comparison and for evaluation.

The reference weights θ_0 and reference eigenfunctions are those of the independent model. Their values are known, whereas the population weights under nonzero dependence are not computed analytically. The two errors in Table 8 measure discrepancy from this reference. They are not estimation errors from the unknown perturbed optimum. The theorem gives a local population conclusion. This experiment measures sensitivity to dependence without determining the theorem's admissible neighborhood.

Table 8: Support and reference errors under dependence.

N	ϵ	Exact support	$\|\widehat{\theta} - \theta_0\|_2$	Spectral error	Gap/tolerance
1024	0	100/100	0.1757 (0.0893)	0.1505 (0.1730)	6.49e-06
1024	0.25	100/100	0.1817 (0.0990)	0.1621 (0.1932)	6.48e-06
1024	0.5	100/100	0.1765 (0.0927)	0.1236 (0.1280)	6.54e-06
1024	0.75	99/100	0.1853 (0.0913)	0.1475 (0.1617)	6.48e-06
4096	0	100/100	0.0976 (0.0425)	0.0459 (0.0121)	4.52e-05
4096	0.25	100/100	0.0886 (0.0463)	0.0440 (0.0119)	4.51e-05
4096	0.5	100/100	0.0814 (0.0475)	0.0433 (0.0114)	4.52e-05
4096	0.75	100/100	0.0883 (0.0439)	0.0442 (0.0124)	4.52e-05

Table 8 reports exact signal selection in 99–100 of 100 samples at $N = 1024$ and all 100 samples at $N = 4096$, at every perturbation level. Numbers in parentheses represent standard deviations. All 800 fits satisfy the global exchange, curvature, and objective certificates, with maximum gap divided by tolerance 4.52×10^{-5}. The finite-sample selection results show that the criterion distinguishes the signal from a dependent nuisance block in this design.

Table 9: Conditional-mean prediction error under dependence.

N	ϵ	PALE	Oracle support	Isotropic LE	PLS
1024	0	0.0364 (0.0818)	0.0038 (0.0011)	0.9312 (0.0269)	0.0508 (0.0063)
1024	0.25	0.0412 (0.0936)	0.0039 (0.0011)	0.9183 (0.0293)	0.0518 (0.0061)
1024	0.5	0.0218 (0.0586)	0.0037 (0.0011)	0.8709 (0.0356)	0.0574 (0.0066)
1024	0.75	0.0330 (0.0751)	0.0036 (0.0012)	0.8136 (0.0276)	0.0645 (0.0073)
4096	0	0.0023 (0.0012)	0.0010 (0.0003)	0.9337 (0.0255)	0.0128 (0.0013)
4096	0.25	0.0021 (0.0012)	0.0010 (0.0002)	0.8729 (0.0321)	0.0151 (0.0018)
4096	0.5	0.0020 (0.0011)	0.0010 (0.0003)	0.8073 (0.0181)	0.0208 (0.0021)
4096	0.75	0.0021 (0.0014)	0.0009 (0.0002)	0.7728 (0.0214)	0.0290 (0.0027)

Table 9 relates this selection to prediction. At $N = 4096$, mean conditional-mean error ranges from 0.0020 to 0.0023 for PALE, compared with 0.0128 to 0.0290 for PLS and 0.7728 to 0.9337 for isotropic eigenmaps. The advantage over these comparisons persists across all four perturbation levels at this sample size. The oracle-support comparison gives the error when the generating support and equal signal weights are known. These results concern the prescribed-response model and do not establish the same selected set for arbitrary dependent measurements or random coordinate masks.

I.8 Independent molecular data and nonlinear comparisons

The independent molecular analysis tests whether a fitted protein representation predicts RNA measurements in cells that did not enter its graph. The public 10x Genomics PBMC data contain 7865 cells from one healthy donor, with 17 antibody measurements including three isotype controls (10x Genomics, 2018). These measurements were collected separately from the CITE-seq cord-blood data. All cells satisfy the prespecified minimum of 500 RNA counts. The analysis uses 1024 pilot cells, 2048 fitting cells, 1024 validation cells, and 1024 further cells for evaluation; the four sets are disjoint.

RNA and protein transformations are the same as in the cord-blood analysis. The 128-gene response panel is selected by pilot variance among genes detected in at least 30 pilot cells. Pilot means and standard deviations scale both views, and transformed values are clipped to $[-3, 3]$. All 17 antibody measurements are available to every method. No cell annotation enters fitting or tuning.

The metric criterion uses disjoint blocks of size 128. Its prediction scales are $h/\sqrt{17} \in \{0.6, 0.9, 1.2\}$, and its ridge candidates are 0.005, 0.05, and 0.5. Spectral dimensions are 5, 10, and 20. Every spectral comparison uses the same 48-neighbor construction and a linear ridge prediction head with penalty one. The training graph and its eigenvectors are fixed before evaluation. New cells use the extension in Section P, with training radii and eigenvalues unchanged. Prediction values from every method are clipped to the response range.

The nonlinear prediction comparisons are Gaussian kernel ridge regression, neural regression, and metric learning for kernel regression, denoted by MLKR (Weinberger & Tesauro, 2007). Kernel ridge regression searches scales $\sqrt{17}\{0.3, 0.6, 1.2, 2.4\}$ and penalties 0.01, 0.1, 1, and 10. Neural regression uses one hidden layer of 64 or 128 units, or two hidden layers of 128 and 64 units, with penalties 10^{-4} or 10^{-2}. Early stopping uses a training-only validation subset. MLKR minimizes the average of its leave-one-out squared losses over all 128 responses; its four candidates use ranks 5 or 17 and initial scales $0.6\sqrt{17}$ or $1.2\sqrt{17}$. The fitted MLKR distances are also passed through the same spectral graph and linear prediction head as PALE. This comparison is denoted by MLKR spectral. PCA, PLS, and full-feature ridge regression are included.

Each method is selected by its own validation prediction error. PALE and its all-weight ablation select ridge 0.05 and dimension 20. MLKR spectral also selects dimension 20. The kernel-ridge comparison selects scale $1.2\sqrt{17}$ and penalty one. The selected neural network has hidden widths 128 and 64. Its selected penalty is 10^{-4}. Native MLKR selects full rank 17. MLKR uses a 120-iteration limit, which is not a convergence certificate. The neural fits use batches of 128, learning rate 0.001, and at most 500 epochs; early stopping uses ten percent of the training cells and 20 epochs without validation improvement.

The PALE fit is repeated over 100 independent random block assignments of the same training cells. Baseline predictors without this random block construction are fixed. Paired bootstrap intervals use 1000 resamples of the 1024 evaluation cells, after averaging the cellwise losses over block assignments. These repetitions quantify algorithmic and within-cohort evaluation variation. They do not represent independent donors or evaluate transfer of a model between donors.

Table 10: Prediction of RNA measurements in independently collected PBMC data.

Method	MSE	Partition SD	Difference	95 percent interval
PALE	0.4765	0.00031	0.0000	–
PALE, all weights	0.4760	0.00039	-0.0004	[-0.0015, 0.0009]
Isotropic LE	0.4912	–	0.0147	[0.0061, 0.0260]
MLKR spectral	0.4715	–	-0.0050	[-0.0086, -0.0015]
PCA	0.5059	–	0.0294	[0.0246, 0.0342]
PLS	0.5059	–	0.0294	[0.0246, 0.0342]
Full-feature ridge	0.5059	–	0.0294	[0.0244, 0.0341]
Kernel ridge	0.4761	–	-0.0004	[-0.0043, 0.0029]
Neural regression	0.4696	–	-0.0069	[-0.0112, -0.0028]
MLKR	0.4628	–	-0.0137	[-0.0182, -0.0088]

Table 10 gives RNA prediction error 0.4765 for PALE, compared with 0.4912 for isotropic eigenmaps and 0.5059 for PLS. The paired interval for the isotropic-minus-PALE difference is $[0.0061, 0.0260]$. PALE selects a mean of 10.71 antibody coordinates and excludes all three isotype controls in every fit. Thus the predictive metric excludes assay-control measurements and gives lower RNA prediction error than the isotropic graph.

Kernel ridge regression gives error 0.4761, with a paired difference interval containing zero. MLKR spectral, neural regression, and native MLKR give smaller errors of 0.4715, 0.4696, and 0.4628, respectively. The all-weight PALE comparison gives 0.4760, and its paired difference interval also

contains zero. These results establish the improvement over isotropic and linear representations in a second collection. They do not establish a general prediction advantage over nonlinear regression.

I.9 Prediction at a common representation dimension

We compare prediction under the same dimension constraint for every representation. Each method uses five or ten coordinates for the 13 protein measurements in CITE-seq and the 17 antibody measurements in PBMC. Twenty spectral coordinates provide a separate reference for prediction with a larger representation.

The pilot transformations, 128-gene response panels, 2048 fitting observations, and 1024 validation observations are as specified in Sections I and I.8. Separate evaluation samples contain 2885 CITE-seq cells and 2745 PBMC cells. These cells do not enter preprocessing, fitting, or validation and do not overlap the earlier evaluation samples. For this comparison, each graph contains only fitting observations. New cells use (23). The CITE-seq and PBMC graphs have 24 and 48 neighbors, respectively.

We evaluate 100 metrics fitted under independent random block assignments of each fitting sample. The metric criterion and its ridge parameter do not depend on the representation dimension. PALE and its all-weight comparison use ridge 0.05 and the same fitted distances at all three dimensions. PALE selects an average of 9.78 protein features in CITE-seq and 10.71 in PBMC. The same selected features are used at each dimension. Every representation uses a linear ridge decoder with penalty one. Coordinates are scaled on the fitting sample, and predictions are clipped to $[-3, 3]$. Isotropic eigenmaps, PCA, PLS, and MLKR spectral receive the same fitting covariates and, for the response-dependent methods, the same RNA responses. MLKR spectral chooses its metric separately at each dimension using validation error. Its four candidates have ranks five or the full observed dimension and initial scales $0.6\sqrt{m}$ or $1.2\sqrt{m}$. The optimization limit is 120 iterations. PCA and PLS are evaluated at five and ten coordinates.

Table 11: RNA prediction error at common dimensions and a larger spectral dimension.

Collection	Method	$p = 5$	$p = 10$	$p = 20$
CITE-seq	PALE	0.5928 (0.0024)	0.5711 (0.0010)	0.5643 (0.0020)
	PALE without selection	0.5960 (0.0002)	0.5710 (0.00005)	0.5643 (0.0001)
	Isotropic LE	0.5994	0.5734	0.5667
	MLKR spectral	0.5904	0.5655	0.5602
	PCA	0.6304	0.6042	–
	PLS	0.6100	0.5984	–
PBMC	PALE	0.5177 (0.0002)	0.4917 (0.0002)	0.4890 (0.0013)
	PALE without selection	0.5177 (0.0001)	0.4917 (0.0001)	0.4875 (0.0003)
	Isotropic LE	0.5336	0.5092	0.4987
	MLKR spectral	0.5154	0.4882	0.4827
	PCA	0.5816	0.5404	–
	PLS	0.5424	0.5199	–

Table 11 shows lower prediction error for PALE than for isotropic and linear representations at both common dimensions in both collections. Numbers in parentheses represent standard deviations. They describe variation over the 100 block assignments; the other fitted representations are fixed. At ten coordinates, PBMC error is 0.4917 for PALE, 0.5092 for isotropic eigenmaps, and 0.5199 for PLS. The corresponding CITE-seq errors are 0.5711, 0.5734, and 0.5984. Full-feature ridge errors are 0.5934 in CITE-seq and 0.5155 in PBMC. At ten coordinates, PALE also improves on full-feature ridge in both collections.

Table 12 gives the error of each comparison minus the PALE error, with pointwise 95% paired intervals from 1000 cell resamples. Cellwise losses are averaged over block assignments before resampling. At ten coordinates, the isotropic difference is 0.0023 with interval $[0.0010, 0.0034]$ in CITE-seq and 0.0175 with interval $[0.0134, 0.0218]$ in PBMC. At five coordinates, selection reduces CITE-seq error by 0.0032 relative to using all weights, with interval $[0.0023, 0.0042]$. The other all-

Table 12: Paired differences in RNA prediction error from PALE.

Collection	Comparison	$p = 5$	$p = 10$
CITE-seq	PALE without selection	0.0032 [0.0023, 0.0042]	-0.0001 [-0.0007, 0.0004]
	Isotropic LE	0.0066 [0.0025, 0.0108]	0.0023 [0.0010, 0.0034]
	MLKR spectral	-0.0024 [-0.0046, -0.0005]	-0.0057 [-0.0080, -0.0036]
	PCA	0.0376 [0.0336, 0.0415]	0.0331 [0.0301, 0.0358]
	PLS	0.0172 [0.0135, 0.0206]	0.0272 [0.0244, 0.0298]
PBMC	PALE without selection	-0.00003 [-0.0003, 0.0003]	-0.00003 [-0.0004, 0.0004]
	Isotropic LE	0.0159 [0.0116, 0.0204]	0.0175 [0.0134, 0.0218]
	MLKR spectral	-0.0023 [-0.0037, -0.0010]	-0.0035 [-0.0051, -0.0019]
	PCA	0.0639 [0.0590, 0.0685]	0.0487 [0.0453, 0.0518]
	PLS	0.0247 [0.0218, 0.0275]	0.0282 [0.0253, 0.0310]

weight intervals include zero. The repeated assignments and cell resamples quantify variation within each collection, not variation between donors.

MLKR spectral has lower error than PALE at both common dimensions in both collections. The comparison does not establish a prediction advantage over all response-dependent metrics. Reducing PALE from twenty to ten coordinates increases error by 0.0068 in CITE-seq and 0.0027 in PBMC. Table 11 also shows lower variation over block assignments at ten coordinates in both collections.

I.10 Complete computational cost

The cost of learning a representation includes more than metric fitting. We measure the complete calculation from metric fitting to prediction for new observations in the sphere model of Section I.6. The observed dimension is 192, the response has three coordinates, and the noise amplitude is 0.4. The metric uses the same prediction scale, ridge, screening rule, and global certificate as the identified model. The metric and training graph use the same $n = N$ observations, with $n \in \{1024, 4096, 8192\}$. Training blocks contain 64, 128, and 128 observations, respectively. The representation has three coordinates and $k_n = \lceil n^{3/4} \rceil$ neighbors. An independent sample of 1024 observations assesses prediction.

PALE is computed by both the dense and exact block implementations. Isotropic eigenmaps use the same block implementation, and PLS provides a linear comparison. Each method receives the same generated observations and fitting responses when its criterion uses responses. All methods use a linear ridge decoder with penalty one; coordinate scaling uses the fitting observations. Spectral predictions use (23) with the training graph fixed. The prediction error is measured from the known conditional mean. This calculation evaluates the implementation used for new observations; the joint-graph least-squares experiment in Section I.6 separately assesses the prediction argument of Theorem 8.

We generate 100 independent samples at each sample size. Each sample and method is evaluated in a fresh process with one numerical thread. Method order is randomized within each sample, and benchmark processes run sequentially. Wall time starts before training-block construction and ends after all new-point predictions. It includes the global metric certificate, graph construction, eigensolution, and fitting the prediction coefficients. Data generation, library imports, and saving results are excluded from time. Peak resident memory is recorded at the end of prediction for the full process, including its numerical libraries and the generated data. Measurements use an Apple M4 Max processor with 128 GiB of memory, macOS 26.2, Python 3.9.6, NumPy 2.0.2, and SciPy 1.13.1.

Table 13 reports mean time, peak resident memory, and conditional-mean prediction error. Numbers in parentheses represent standard deviations. The two PALE implementations give identical metric

Table 13: Complete fitting and new-point prediction cost in the continuous model.

n	Method	Time in seconds	Peak memory in MiB	Prediction error
1024	PALE dense	0.200 (0.063)	229.6 (1.9)	0.0399 (0.0794)
	PALE exact blocks	0.093 (0.030)	175.6 (2.5)	0.0399 (0.0794)
	Isotropic LE	0.227 (0.075)	176.0 (3.2)	0.9889 (0.0076)
	PLS	0.006 (0.003)	164.2 (1.0)	0.1830 (0.0097)
4096	PALE dense	1.870 (0.465)	1221.5 (8.7)	0.0034 (0.0017)
	PALE exact blocks	0.762 (0.227)	309.6 (1.9)	0.0034 (0.0017)
	Isotropic LE	2.720 (0.791)	305.7 (1.3)	0.9868 (0.0066)
	PLS	0.020 (0.010)	181.0 (0.7)	0.0517 (0.0032)
8192	PALE dense	6.678 (1.768)	3864.0 (27.3)	0.0016 (0.0009)
	PALE exact blocks	2.357 (0.642)	595.3 (4.6)	0.0016 (0.0009)
	Isotropic LE	9.522 (2.211)	599.4 (2.8)	0.9857 (0.0082)
	PLS	0.061 (0.033)	203.3 (0.7)	0.0268 (0.0018)

weights, neighbor radii, affinities, and predictions in each of the 300 samples. All 600 metric fits meet the prescribed global tolerance $1/(N \log N)$.

Table 14: Mean time in seconds for each computational step.

n	Method	Metric	Graph	Eigenvectors	Decoder	New points
1024	PALE dense	0.041	0.079	0.013	0.001	0.066
1024	PALE exact blocks	0.039	0.024	0.013	0.001	0.015
1024	Isotropic LE	0.000	0.127	0.032	0.001	0.068
4096	PALE dense	0.189	1.273	0.149	0.001	0.258
4096	PALE exact blocks	0.194	0.352	0.162	0.001	0.054
4096	Isotropic LE	0.000	2.000	0.464	0.001	0.255
8192	PALE dense	0.371	5.195	0.556	0.001	0.555
8192	PALE exact blocks	0.365	1.316	0.574	0.001	0.102
8192	Isotropic LE	0.000	7.386	1.665	0.001	0.470

At $n = 8192$, mean total time decreases from 6.678 to 2.357 seconds and mean peak memory from 3864.0 to 595.3 MiB. Table 14 shows that graph construction decreases from 5.195 to 1.316 seconds and new-point prediction from 0.555 to 0.102 seconds. These calculations use the selected coordinates without storing a full distance matrix. The metric objective and eigenproblem are unchanged.

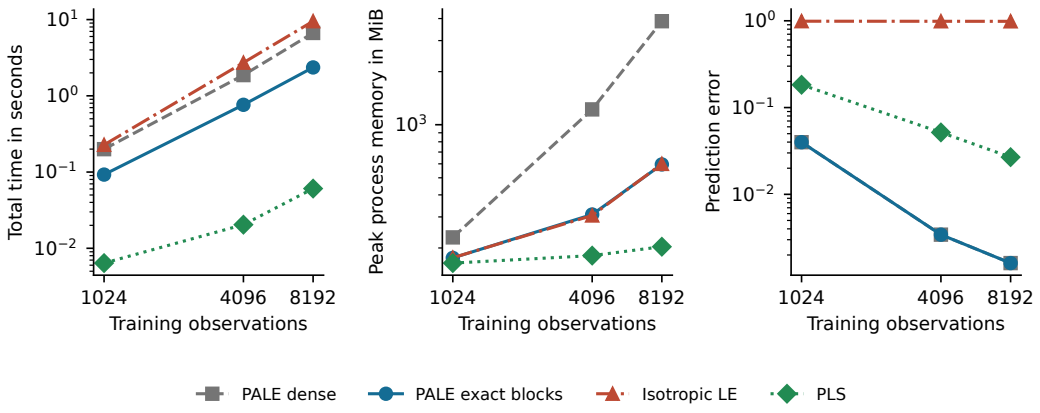

Figure 4: Mean computation time, peak memory, and prediction error.

Table 15: Macro-averaged classification error. Parentheses give standard deviations across 100 runs.

Method	Synthetic	Outdoor allocations	Outdoor evaluation	Prokaryotic confirm.
Posterior-moment PALE	**0.190 (0.009)**	**0.239 (0.021)**	**0.238**	0.091 (0.027)
Posterior-score eigenmap	–	–	–	**0.087 (0.024)**
Extra Trees	0.367 (0.059)	0.374 (0.033)	0.422	0.097 (0.023)
Privileged pseudo-label LDA	0.284 (0.017)	0.269 (0.025)	0.278	0.102 (0.022)
Extra Trees score distillation	0.227 (0.026)	0.274 (0.026)	0.322	0.105 (0.035)
Kernel score distillation	0.401 (0.029)	0.317 (0.032)	0.328	0.100 (0.029)
Isotropic eigenmap	–	–	–	0.123 (0.034)
Original posterior-response PALE	–	0.413 (0.046)	0.475	0.119 (0.034)

Figure 4 compares the decrease in prediction error with the increase in computation as the sample size grows. At $n = 8192$, PALE has conditional-mean error 0.0016, compared with 0.0268 for PLS. PLS uses less time and memory. PALE is more accurate than isotropic eigenmaps and requires less total time at each sample size, including metric fitting. The exact search remains quadratic in the number of observations; Section R.1 derives its storage bound.

J COMPLETE POSTERIOR-MOMENT NUMERICAL RESULTS

The synthetic classification study has four classes, five predictive coordinates, and 95 independent Gaussian nuisance coordinates; each of 100 data sets has eight labels per class, 1200 paired unlabeled observations, and 2000 evaluation observations.

Outdoor Scene has 2688 images from eight classes and four feature views (Dong et al., 2018); the deployable vector adds three label-free nuisance GIST blocks to the current-image GIST block, and the teacher uses all four clean views. The 100 fixed allocations of 80 labeled, 1968 paired unlabeled, and 320 evaluation images are followed by an evaluation of all 24 fixed methods on a separate balanced set of 320 images; this is a controlled real-feature analysis, not the unmodified benchmark. The public Prokaryotic benchmark contains 551 species in four classes with gene-repertoire, proteome-composition, and text views of 393, 3, and 438 features (Brbić et al., 2016; Niu et al., 2023); we use the gene-repertoire view at evaluation and all three views for the teacher. Each of 100 fixed allocations selects ten labeled observations per class, 431 paired unlabeled observations enter the graph, and all 20 prediction arrays for a separate balanced set of 40 observations were fixed before confirmation-label access.

Table 15 shows synthetic errors 0.190 and 0.227 for posterior-moment PALE and its strongest deployable comparison; the proposed error is smaller in all 100 replications, with all five predictive coordinates selected. Across the Outdoor allocations, the proposed and closest errors are 0.239 and 0.269, with paired 95 percent difference interval $[0.025, 0.035]$, and all 18 selected coordinates come from the clean GIST block in every allocation.

On the separate Outdoor evaluation set, the proposed, closest, and Extra Trees distillation errors are 0.238, 0.278, and 0.322, with paired intervals $[0.003, 0.078]$ and $[0.044, 0.125]$ and adjusted exact values 0.0298 and 0.000306; original label and posterior-response PALE have errors 0.438 and 0.475.

On Prokaryotic, the posterior-score eigenmap has the smallest mean error among all 20 methods: 0.087, compared with 0.091 for its native classifier and 0.097 for Extra Trees. The allocation-resampling intervals for comparison minus proposed error are $[0.001, 0.006]$ and $[0.004, 0.016]$, respectively; the observation-resampling interval against the native classifier is $[-0.002, 0.010]$. The comparisons below include all methods and both uncertainty calculations.

The comparisons assess coordinate selection, native classification, and prediction from the class-score affinity. They use (13) and (14), with the fixed covariance regularization specified below. Paired uncertainty calculations distinguish variation across training allocations from variation across confirmation observations.

J.1 SYNTHETIC PRODUCTION STUDY

We use four class means: $\mu_1 = (1, 1, 1, 1, 1)^\top$, $\mu_2 = (1, -1, 1, -1, 1)^\top$, $\mu_3 = -\mu_2$, and $\mu_4 = -\mu_1$. Conditional on class C, the first five deployable coordinates satisfy $X_{1:5} = \mu_C + 1.2Z$, and the auxiliary view is $V = \mu_C + 0.35E$. The vectors Z and E are independent $N(0, I_5)$ variables, independent of C. The remaining 95 deployable coordinates are independent $N(0, 1)$ variables, independent of C, Z, and E. We sample eight labeled observations per class, 1200 paired unlabeled observations, and 2000 evaluation observations. Classes in the latter two samples are uniform. Each view is standardized using its 32 labeled observations.

The synthetic development phase uses ten seeds that do not appear in the production phase. The production seeds are 958000 through 958099. The RBF teacher has penalty 10. Its class scores are converted by a fixed softmax temperature of one, and the half with largest maximum score is selected. For each selected observation, we use the one-hot vector of the teacher's largest score as its responsibility. The labeled responsibility weight is 0.5. Posterior-moment PALE selects five coordinates and uses covariance ridge 0.001 without diagonal shrinkage. The prediction prior is uniform.

Every comparison setting is fixed in the development phase. The comparison set includes direct pseudo-label LDA on all coordinates; Fisher and ANOVA selection followed by LDA; RBF support vector classification; logistic regression; Extra Trees; random forest; Extra Trees score distillation; a multilayer perceptron; LightGBM; neighborhood component analysis followed by nearest neighbors; kernel score distillation; and a partial least-squares representation of the paired auxiliary view. The all-view teacher is a reference that observes the auxiliary view at prediction time. For each comparison, Table 16 gives a paired 95 percent Student interval for comparison error minus proposed error. The win rate is the fraction of replications with strictly smaller proposed error.

Table 16: Complete 100-replication posterior-moment simulation. The interval is for comparison error minus proposed error.

Method	Error	SD	Difference interval	Win rate
Posterior PALE	0.190	0.009	[0.000, 0.000]	0.00
All-view teacher reference	0.218	0.020	[0.023, 0.032]	0.91
Extra Trees distillation	0.227	0.026	[0.032, 0.042]	1.00
Observed-view self-training	0.237	0.048	[0.037, 0.056]	0.99
PLS privileged representation	0.270	0.025	[0.075, 0.084]	1.00
Privileged pseudo-label LDA	0.284	0.017	[0.090, 0.096]	1.00
Label Fisher selection with LDA	0.298	0.075	[0.093, 0.123]	1.00
ANOVA selection with LDA	0.298	0.075	[0.093, 0.123]	1.00
Extra Trees	0.367	0.059	[0.165, 0.189]	1.00
Kernel score distillation	0.401	0.029	[0.206, 0.217]	1.00
Random forest	0.409	0.066	[0.205, 0.232]	1.00
LightGBM	0.538	0.067	[0.334, 0.361]	1.00
Multilayer perceptron	0.549	0.022	[0.355, 0.364]	1.00
Logistic regression	0.552	0.020	[0.357, 0.365]	1.00
RBF SVM	0.557	0.020	[0.363, 0.371]	1.00
NCA nearest neighbors	0.627	0.024	[0.431, 0.442]	1.00

The proposed method recovers the five-coordinate support in every replication. Its mean error is smaller than that of every deployable comparison in Table 16. The all-view teacher is nondeployable and has smaller error than the proposed method in nine replications; its mean error remains larger.

J.2 OUTDOOR SCENE ALLOCATIONS

The source has GIST, color-moment, HOG, and local-binary-pattern views with dimensions 512, 432, 256, and 48. Each view uses at most 40 principal components fitted from the 80 labeled observations in each allocation. We remove components whose singular values do not exceed machine precision times the larger training matrix dimension times the largest singular value. Standardization uses the remaining components and the same labeled observations. Standardized coordinates

are clipped to $[-5, 5]$. Three independent label-free index arrays select nuisance GIST rows from the full empirical collection. Each selected row is transformed through the fitted GIST map and appended to the deployable GIST representation. The resulting deployable dimension is 160, of which the first 40 coordinates belong to the current image.

The open-data set contains 50 labeled images per class. Each of 100 seeds selects ten per class for training and leaves 40 per class for evaluation. The 1968 paired unlabeled images are common to these allocations. Posterior-moment PALE selects half of the teacher scores, selects 18 coordinates, and uses covariance ridge 0.001. The selected unlabeled observations have hard responsibilities, with labeled responsibility weight 0.5 and no diagonal covariance shrinkage. The complete comparison set includes the synthetic-study families, supervised Fisher and mutual-information selection, the original label and posterior-response versions of PALE, observed-view self-training, and two nondeployable references. Table 17 gives paired 95 percent Student intervals across allocations.

Table 17: Outdoor Scene controlled-nuisance results over 100 open-data allocations. The interval is for comparison error minus proposed error.

Method	Error	SD	Difference interval	Win rate
Posterior PALE	0.239	0.021	[0.000, 0.000]	0.00
Privileged pseudo-label LDA	0.269	0.025	[0.025, 0.035]	0.90
Extra Trees distillation	0.274	0.026	[0.030, 0.040]	0.89
Kernel score distillation	0.317	0.032	[0.072, 0.083]	1.00
Clean-view Extra Trees reference†	0.319	0.023	[0.074, 0.085]	1.00
PLS privileged representation	0.370	0.031	[0.123, 0.137]	1.00
Supervised Fisher selection with LDA	0.373	0.047	[0.124, 0.143]	1.00
Extra Trees	0.374	0.033	[0.128, 0.142]	1.00
Original label PALE	0.376	0.041	[0.128, 0.145]	1.00
Observed-view self-training	0.392	0.057	[0.141, 0.163]	1.00
Original posterior PALE	0.413	0.046	[0.163, 0.183]	1.00
All-view teacher reference†	0.426	0.033	[0.179, 0.194]	1.00
Mutual-information selection with LDA	0.472	0.042	[0.224, 0.242]	1.00

†The reference uses information excluded from the deployable observation scheme.

Posterior-moment PALE selects only coordinates from the current-image GIST block in all 100 allocations. Its error is smaller than the closest comparison in 90 of the allocations and smaller than Extra Trees distillation in 89.

J.3 SEPARATE OUTDOOR EVALUATION

The evaluation uses 80 labeled images, the same 1968 paired unlabeled images, and a separate balanced set of 320 images. All 24 methods use the settings selected on the development observations and the preceding rule for excluding numerical null components. The predictions were recomputed after evaluation labels had been accessed; these comparisons describe a reanalysis of this fixed sample and do not constitute a new blinded confirmation. Paired uncertainty uses 10000 stratified bootstrap samples. For each deployable comparison, the last column of Table 18 gives a Holm-adjusted one-sided exact McNemar value as a descriptive comparison.

Posterior-moment PALE has the smallest observed error among all 24 methods. Its error difference from the closest method is 0.040625, with a positive paired bootstrap interval and adjusted exact value 0.0298. The differences from Extra Trees distillation and the two original PALE procedures also have positive intervals.

J.4 PUBLIC PROKARYOTIC BENCHMARK

The Prokaryotic data contain 551 species in four classes and three aligned views: 393 gene-repertoire features, three proteome-composition features, and 438 text features (Brbić et al., 2016; Niu et al., 2023). We use the gene-repertoire view as the deployable measurement. The teacher uses

Table 18: Outdoor Scene evaluation on the separate 320-image sample. The interval is for comparison error minus proposed error. The last column gives the Holm-adjusted one-sided exact McNemar value for deployable comparisons.

Method	Error	Difference	Bootstrap interval	Adjusted value
Posterior PALE	0.238	0.000	[0.000, 0.000]	–
Privileged pseudo-label LDA	0.278	0.041	[0.003, 0.078]	0.02979
Extra Trees distillation	0.322	0.084	[0.044, 0.125]	3.06×10^{-4}
Kernel score distillation	0.328	0.091	[0.047, 0.134]	3.06×10^{-4}
Clean-view Extra Trees reference†	0.356	0.119	[0.075, 0.166]	–
ANOVA selection with LDA	0.400	0.163	[0.109, 0.216]	3.37×10^{-8}
Supervised Fisher selection with LDA	0.400	0.163	[0.109, 0.216]	3.37×10^{-8}
PLS privileged representation	0.403	0.166	[0.119, 0.213]	1.78×10^{-9}
All-view teacher reference†	0.409	0.172	[0.119, 0.222]	–
Extra Trees	0.422	0.184	[0.134, 0.234]	1.62×10^{-10}
Observed-view self-training	0.438	0.200	[0.150, 0.250]	3.69×10^{-12}
Original label PALE	0.438	0.200	[0.147, 0.253]	5.18×10^{-11}
Random forest	0.447	0.209	[0.159, 0.259]	3.42×10^{-12}
Mutual-information selection with LDA	0.447	0.209	[0.153, 0.266]	5.45×10^{-11}
ANOVA selection with logistic regression	0.469	0.231	[0.181, 0.281]	6.71×10^{-15}
Supervised Fisher selection with logistic regression	0.469	0.231	[0.181, 0.281]	6.71×10^{-15}
Original posterior PALE	0.475	0.238	[0.191, 0.287]	8.38×10^{-16}
Mutual-information selection with logistic regression	0.491	0.253	[0.200, 0.306]	1.36×10^{-16}
L1 logistic regression	0.491	0.253	[0.200, 0.306]	8.26×10^{-16}
LightGBM	0.516	0.278	[0.222, 0.334]	3.47×10^{-18}
L2 logistic regression	0.619	0.381	[0.322, 0.441]	8.98×10^{-26}
Multilayer perceptron	0.619	0.381	[0.322, 0.441]	8.98×10^{-26}
RBF SVM	0.653	0.416	[0.356, 0.475]	3.63×10^{-29}
NCA nearest neighbors	0.709	0.472	[0.412, 0.531]	2.73×10^{-37}

†The reference uses auxiliary measurements at prediction time or prespecified knowledge of the clean feature block.

all three views for labeled and paired unlabeled observations. No artificial feature or nuisance block is added.

The fixed roles contain 80 development observations, 40 balanced confirmation observations, and 431 paired unlabeled observations. Each of 100 seeds selects ten labeled development observations per class. The other development observations are not used by that confirmation fit. The teacher probabilities provide soft responsibilities for the class moments. The spectral predictor uses hard teacher targets for the paired unlabeled observations. The graph contains the 40 labeled observations, the 431 paired unlabeled observations, and the 40 confirmation covariates. Every preprocessing map, tuning value, and prediction array was fixed before the confirmation labels were accessed.

For each allocation, principal component analysis and standardization are fitted separately in each view using the 40 labeled observations. We use 39, 3, and 39 components in the respective views and truncate standardized coordinates to $[-5, 5]$. The deployable view has 39 coordinates, all of which enter the posterior-moment scores. The RBF teacher uses penalty one and softmax temperature 0.5. We select the half of the paired unlabeled observations with the largest maximum class probability and assign labeled responsibility weight one. The covariance uses $\tau_{\mathrm{cov}} = 0.5$ and ridge 0.1. The Gaussian score bandwidth is 16 times the median pairwise score distance among labeled observations. The spectral predictor uses 25 nonconstant eigenvectors, labeled weight 100, and ridge parameter 10.

The 20 registered methods include the native posterior-moment classifier, direct and eigenmap versions of tree and kernel score distillation, privileged pseudo-label LDA, isotropic and self-training eigenmaps, label spreading, supervised neighborhood component analysis, the original posterior-response PALE procedure, linear and kernel classifiers, boosting methods, and a multilayer perceptron. Table 19 gives both allocation and observation resampling. The allocations reuse the same confirmation observations and do not represent independent biological samples.

The posterior-score eigenmap has the smallest mean error among all registered methods. Its error is 0.08725, compared with 0.09075 for the native posterior-moment classifier and 0.09700 for Extra Trees, the strongest external comparison. The allocation-resampling intervals for these two differences are $[0.0010, 0.0060]$ and $[0.0040, 0.0157]$. The corresponding observation-resampling intervals include zero. The result establishes the prediction ranking under the registered protocol and separates this empirical conclusion from sampling across new species collections.

K PROOFS OF THE FINITE-GRAPH RESULTS

The finite-graph identities justify the symmetric prediction weights and their exact derivatives. The Gaussian construction also verifies graph degeneracy within the metric family.

Proof of Proposition 9. Because $W\mathbf{1} = \mathbf{1}$, the degree matrix is I_n and $L = I_n - W$. Connectedness implies that the null space of L is spanned by $\mathbf{1}$. The constraints in (19) restrict Y to p orthonormal vectors in $\mathbf{1}^{\perp}$. The Ky Fan variational principle gives

$$\min_{Y^{\top}Y = I_p,\, Y^{\top}\mathbf{1} = 0} \operatorname{tr}(Y^{\top} L Y) = \sum_{r=2}^{p+1} \lambda_r(L).$$

The equality condition in the same principle gives the minimizing eigenspace. This completes the proof. □

The equal-block construction in the proof of Theorem 1 gives an explicit spectrum and row entropy. We next consider the Gaussian metric family.

Proof of Proposition 10. For each group, permutation invariance makes the scaling factor constant within that group. Write this factor as $u_g(t)$. The row equation for group g is

$$u_g(t)\Big\{(s-1)u_g(t) + s\sum_{r \neq g} e^{-t(c_g - c_r)^2} u_r(t)\Big\} = 1,$$

Table 19: Prokaryotic confirmation over 100 fixed training allocations. Differences subtract the posterior-score eigenmap error; intervals resample allocations and confirmation observations, and adjusted values use one-sided sign tests across allocations.

Method	Error	Difference	Allocation interval	Observation interval	Adjusted value
Posterior-score eigenmap	**0.087 (0.024)**	0.000	[0.000, 0.000]	[0.000, 0.000]	–
Posterior-moment native classifier	0.091 (0.027)	0.003	[0.001, 0.006]	[-0.002, 0.010]	0.004678
Extra Trees	0.097 (0.023)	0.010	[0.004, 0.016]	[-0.005, 0.030]	8.70×10^{-6}
Kernel score distillation	0.100 (0.029)	0.013	[0.009, 0.017]	[0.002, 0.033]	2.73×10^{-9}
Privileged pseudo-label LDA	0.102 (0.022)	0.014	[0.011, 0.018]	[0.001, 0.034]	4.38×10^{-10}
Extra Trees score distillation	0.105 (0.035)	0.018	[0.013, 0.024]	[0.003, 0.042]	6.32×10^{-10}
Kernel distillation eigenmap	0.112 (0.027)	0.024	[0.019, 0.029]	[0.003, 0.053]	7.33×10^{-16}
Extra Trees distillation eigenmap	0.119 (0.035)	0.031	[0.025, 0.038]	[0.011, 0.058]	5.24×10^{-16}
Original posterior-response PALE	0.119 (0.034)	0.031	[0.025, 0.038]	[0.013, 0.054]	5.16×10^{-16}
Multinomial logistic regression	0.121 (0.033)	0.034	[0.028, 0.040]	[0.012, 0.060]	7.05×10^{-20}
Isotropic eigenmap	0.123 (0.034)	0.036	[0.030, 0.042]	[0.014, 0.062]	1.12×10^{-20}
Observed self-training isotropic eigenmap	0.128 (0.041)	0.040	[0.033, 0.048]	[0.017, 0.067]	2.68×10^{-20}
RBF SVM	0.133 (0.040)	0.046	[0.039, 0.053]	[0.021, 0.074]	1.13×10^{-22}
Observed self-training score eigenmap	0.141 (0.042)	0.054	[0.046, 0.062]	[0.026, 0.083]	9.05×10^{-26}
Supervised NCA eigenmap	0.158 (0.052)	0.071	[0.062, 0.080]	[0.043, 0.101]	3.79×10^{-28}
LightGBM	0.171 (0.060)	0.084	[0.072, 0.096]	[0.056, 0.113]	1.63×10^{-22}
Label-only score eigenmap	0.175 (0.051)	0.088	[0.079, 0.097]	[0.048, 0.130]	2.02×10^{-28}
Histogram gradient boosting	0.188 (0.059)	0.100	[0.089, 0.112]	[0.069, 0.133]	5.68×10^{-29}
Label spreading	0.213 (0.079)	0.126	[0.111, 0.141]	[0.091, 0.161]	5.68×10^{-29}
Multilayer perceptron	0.720 (0.111)	0.633	[0.611, 0.654]	[0.556, 0.699]	1.50×10^{-29}

where c_g is the value of the separating feature in group g. The scaling factors remain bounded above and away from zero. Because the c_g are distinct, every exponential in the second term converges to zero. Hence $u_g(t) \to (s-1)^{-1/2}$ for every g.

It follows that $W(t)$ converges entrywise to the block matrix B in the proof of Theorem 1. Eigenvalue continuity gives $\Phi_p\{W(t)\} \to \Phi_p(B) = 0$. Continuity of $-x \log x$ at zero gives the row-entropy limit $\log(s-1)$. This completes the proof. □

The construction also gives a quantitative statement for a degree barrier. It is useful when comparing normalized and unnormalized formulations.

Corollary 22. *For every function $g : (0,\infty)^n \to \mathbb{R}$, the penalty $g(W\mathbf{1})$ is constant along the sequence in Theorem 1. In particular, no penalty depending only on degrees can keep Φ_p away from zero.*

Corollary 22 formalizes the difference between isolated vertices and weak cuts. A positive degree at every vertex gives no lower bound for graph conductance.

Proof of Corollary 22. Theorem 1 gives $W_\varepsilon \mathbf{1} = \mathbf{1}$ for every ε. Thus $g(W_\varepsilon \mathbf{1}) = g(\mathbf{1})$, while $\Phi_p(W_\varepsilon) \to 0$. This completes the proof. □

The row-entropy limit in Theorem 1 identifies entropy constraints that allow the eigenmap loss to approach zero. Their range depends on the block size.

Corollary 23. *For every $h_0 < \log(s-1)$, the sequence in Theorem 1 satisfies*

$$\min_{1 \le i \le n} \Big\{ -\sum_{j \ne i} W_{\varepsilon,ij} \log W_{\varepsilon,ij} \Big\} \ge h_0$$

for all sufficiently small $\varepsilon > 0$, while $\Phi_p(W_\varepsilon) \to 0$.

Corollary 23 shows that an entropy lower bound below $\log(s-1)$ does not bound the spectral gap away from zero.

Proof of Corollary 23. Every row of W_ε has the same entropy, and Theorem 1 gives its limit as $\log(s-1)$. For $h_0 < \log(s-1)$, convergence gives the entropy inequality for all sufficiently small ε. The profile-loss conclusion is part of the same theorem. This completes the proof. □

Proof of Proposition 12. The support graph of K is the complete graph without loops. For $n \ge 3$, this support is fully indecomposable and has total support. The Sinkhorn–Knopp theorem gives positive diagonal matrices D_r, D_c such that $D_r K D_c$ is bistochastic, and the two scalings are unique up to reciprocal scalar multiplication (Sinkhorn & Knopp, 1967).

Because K is symmetric, transposition shows that $D_c K D_r$ is also bistochastic. Uniqueness gives $D_c = \alpha D_r$ for a positive scalar α. Set $D_u = \sqrt{\alpha} D_r$. Then $D_u K D_u$ is bistochastic. Any other symmetric scaling is a scalar multiple of D_u by two-sided uniqueness; unit row sums fix the scalar at one.

Positive edge weights give irreducibility, and cycles of lengths two and three give aperiodicity. Perron–Frobenius gives $|\lambda| < 1$ outside the constant eigenspace. Symmetry makes the eigenvalues real and greater than -1; hence $I_n + W$ is positive definite. This completes the proof. □

Proof of Proposition 11. For a simple random sample of k coordinates,

$$d^2_{a,\xi}(x_i, x_j) = \frac{m}{k} \sum_{\ell : \xi_\ell = 1} z_\ell.$$

The expectation and finite-population variance of this total estimator are

$$\mathbb{E}_\xi d^2_{a,\xi} = \sum_{\ell=1}^{m} z_\ell, \qquad \operatorname{var}_\xi(d^2_{a,\xi}) = \frac{m^2(1-k/m)}{k(m-1)} \sum_{\ell=1}^{m} (z_\ell - \bar{z})^2.$$

For $r \neq s$, sequential sampling gives

$$\mathbb{P}(\xi_r = 1, \xi_s = 0) = \frac{k}{m}\frac{m-k}{m-1}.$$

The probability that this ordered pair is not separated by any of J independent masks is at most the exponential of minus J times this probability. A union bound over the $m(m-1)$ ordered pairs proves the final assertion. This completes the proof. □

The permutation result follows from the distance, scaling, and spectral equations with the observation blocks relabeled.

Proof of Proposition 13. For every pair of transformed observations,

$$d^2_{S^\top a, S^\top \xi}\{(PXS)_i, (PXS)_j\} = d^2_{a,\xi}(x_{P^{-1}i}, x_{P^{-1}j}).$$

Thus $\widetilde{K} = PKP^\top$. If u scales K, then Pu scales $\widetilde{K}$. Uniqueness in Proposition 12 gives $\widetilde{W} = PWP^\top$. The held-out matrix is transformed by a row permutation and a column permutation. The Frobenius norm is invariant under both, which proves equality of the losses.

Apply the block calculation to the rows of I_r and their images under P, using the mask attached to I_r. Summing over the relabeled blocks gives equality of the empirical objectives after the feature weights are permuted. The ridge term is unchanged because S is orthogonal and $S\mathbf{1} = \mathbf{1}$. It also gives equality of their infima, which preserves the approximation inequality and maps $(\widehat{a}, \widehat{h})$ to $(S^\top \widehat{a}, \widehat{h})$. Equation (8) maps $\widehat{S}$ to the permuted support and $\widehat{a}^{\mathrm{rep}}$ to $S^\top \widehat{a}^{\mathrm{rep}}$. The nearest-neighbor radii are permuted in the same way. Hence the final Laplacians satisfy

$$\widehat{\widetilde{L}} = P\widehat{L}P^\top.$$

Conjugation by P maps the entire minimizer set. A unique selected subspace gives orthogonally related bases. At a boundary tie, only the set identity is guaranteed for independently selected minimizers. This completes the proof. □

We next derive the implicit differential used by the algorithm. The proof keeps the scaling contribution, which is the term omitted by a gradient that treats normalization as fixed.

Proof of Proposition 14. Write $u_i = e^{z_i}$ and $K_{ij} = e^{\eta_{ij}}$ for $i \neq j$. The row equations are

$$\sum_j \exp(z_i + \eta_{ij} + z_j) = 1.$$

Their differential is

$$\dot{z}_i + \sum_j W_{ij}\dot{z}_j + \sum_j W_{ij}\dot{\eta}_{ij} = 0,$$

which is (20). Differentiating W_{ij} gives (21).

The loss differential is $\langle G, \dot{W}\rangle$. Put $H = G \circ W$ and $c = H\mathbf{1}$. Symmetry gives

$$\langle G, \dot{W}\rangle = \sum_{i,j} H_{ij}\dot{\eta}_{ij} + 2c^\top \dot{z}.$$

Let $b = (I_n + W)^{-1}c$. Equation (20) and symmetry of $I_n + W$ give

$$2c^\top \dot{z} = -2b^\top (W \circ \dot{\eta})\mathbf{1} = -\sum_{i,j} W_{ij}(b_i + b_j)\dot{\eta}_{ij}.$$

Thus

$$\dot{\ell}_n = \sum_{i,j} C_{ij}\dot{\eta}_{ij}.$$

For a feature weight,

$$\frac{\partial \eta_{ij}}{\partial a_\ell} = -\frac{\xi_\ell (x_{i\ell} - x_{j\ell})^2}{2\pi h^2}.$$

For $\log h$,

$$\frac{\partial \eta_{ij}}{\partial \log h} = \frac{d_{a,\xi}^2(x_i, x_j)}{h^2}.$$

Substitution proves both formulas. This completes the proof. □

Proof of Proposition 19. Compactness and Lemma 4 give finite Lipschitz constants for the criterion and its gradient. For every cell C, bandwidth h, and $a \in C$,

$$\widehat{\mathcal{R}}_\rho(a, h) \geq \widehat{\mathcal{R}}_\rho(c_C, h) - L_R\|a - c_C\|_2 \geq \widehat{\mathcal{R}}_\rho(c_C, h) - L_R r_C.$$

Hence B_t is a lower bound for the global infimum, while U_t is the value of a feasible point. This proves

$$0 \leq U_t - \inf_{a,h} \widehat{\mathcal{R}}_\rho(a, h) \leq U_t - B_t.$$

Every remaining center has criterion at least the global infimum, which gives

$$B_t \geq \inf_{a,h} \widehat{\mathcal{R}}_\rho(a, h) - L_R \overline{r}_t.$$

The remaining cell containing a global minimizer has a center whose criterion is at most the global infimum plus $L_R \overline{r}_t$. Hence

$$U_t - B_t \leq 2L_R \overline{r}_t.$$

Let (C_t, h_t) attain B_t before C_t is divided. Because its center has been evaluated, $U_t \leq \widehat{\mathcal{R}}_\rho(c_{C_t}, h_t)$, and hence $0 \leq U_t - B_t \leq L_R r_{C_t}$. Selected depths diverge because no cell is selected twice and each bounded depth contains finitely many cells. After d midpoint subdivisions of a widest coordinate interval, the largest coordinate width is at most $m(1-\delta)2^{-\lfloor d/m \rfloor}$. This implies $r_{C_t} \to 0$ and proves $U_t - B_t \to 0$. Pruning cannot remove a cell containing a global minimizer because its lower bound does not exceed the global minimum.

The ridge Hessian equals $(\rho/m)I_m$. The definition of M_R gives, for every $w \in \mathbf{1}^\perp$,

$$w^\top \nabla_a^2 \widehat{\mathcal{R}}_\rho(a, h) w \geq (\rho/m - M_R)\|w\|_2^2.$$

Hence $\rho/m > M_R$ implies strong convexity on $\mathcal{A}_\delta$ with constant $\kappa_R = \rho/m - M_R$.

For the curvature certificate, let a^* minimize the fixed-bandwidth criterion and choose $v \in N_{\mathcal{A}_\delta}(a)$. Strong convexity and $v^\top(a^* - a) \leq 0$ give

$$\widehat{\mathcal{R}}_\rho(a, h) - \widehat{\mathcal{R}}_\rho(a^*, h) \leq \{\nabla_a \widehat{\mathcal{R}}_\rho(a, h) + v\}^\top (a - a^*) - \frac{\kappa_R}{2}\|a - a^*\|_2^2 \leq \frac{\|\nabla_a \widehat{\mathcal{R}}_\rho(a, h) + v\|_2^2}{2\kappa_R}.$$

Taking the infimum over v proves the residual bound.

Let L_g be a gradient Lipschitz constant. For any step $\gamma \leq L_g^{-1}$, the projection inequality and the descent lemma give

$$\widehat{\mathcal{R}}_\rho(a^{(t+1)}, h) \leq \widehat{\mathcal{R}}_\rho(a^{(t)}, h) - \frac{1}{2\gamma}\|a^{(t+1)} - a^{(t)}\|_2^2.$$

Under strong convexity, put $\theta = \kappa_R / L_g$ and $y = (1-\theta)a^{(t)} + \theta a^*$. The projection step minimizes its quadratic upper model over $\mathcal{A}_\delta$. Smoothness and strong convexity give

$$\begin{aligned}\widehat{\mathcal{R}}_\rho(a^{(t+1)}, h) &\leq \widehat{\mathcal{R}}_\rho(a^{(t)}, h) + \theta \nabla \widehat{\mathcal{R}}_\rho(a^{(t)}, h)^\top (a^* - a^{(t)}) + \frac{L_g \theta^2}{2}\|a^* - a^{(t)}\|_2^2 \\ &\leq (1-\theta)\widehat{\mathcal{R}}_\rho(a^{(t)}, h) + \theta \widehat{\mathcal{R}}_\rho(a^*, h).\end{aligned}$$

Iteration proves the stated linear rate.

An Armijo line search accepts a positive step whenever the projected gradient is nonzero. Summation of the decrease and continuity of the projected-gradient map show that every accumulation point is stationary. This completes the proof. □

Proof of Proposition 15. For a fixed h and coordinate j, Lemma 4 gives

$$\Big|m\frac{\partial}{\partial a_j}q_b(\mathbf{1},h;\zeta)\Big|\le C.$$

Hoeffding's inequality and a union bound over $m|\mathcal{H}|$ coordinate-bandwidth pairs give, with probability at least $1-\alpha$,

$$\max_{h,j}\Big|\widehat{d}_j(h)-\mathbb{E}\widehat{d}_j(h)\Big|\le C\sqrt{\frac{\log(2m|\mathcal{H}|/\alpha)}{R}}.$$

The coordinatewise proof of Proposition 31 contains one log-kernel derivative, whose bound is C/m. Multiplication by m gives

$$\max_{h,j}|\mathbb{E}\widehat{d}_j(h)-d_j(h)|\le C\sqrt{\frac{m\log(mb)}{b}}.$$

The triangle inequality proves the first assertion.

On the stated event, $j\in S_0$ implies $\widehat{d}_j(h)>\lambda_R$, while $j\notin S_0$ implies $\widehat{d}_j(h)\le\lambda_R$. Thus the screened set equals S_0. This completes the proof. □

Proof of Proposition 16. The restricted polytope is nonempty for every nonempty S: assign $\delta+m(1-\delta)/s$ to its coordinates and δ to the remaining coordinates. If $a_j>\delta$ for an excluded coordinate and every coordinate in S equals A_m, the trace exceeds m. Hence the inner set in the definition of $\widehat{\Gamma}_j(S,h)$ is nonempty.

Suppose the exchange certificate is negative and a global minimizer a^* has $a_j^*>\delta$ for some $j\notin S$. The definition of $\widehat{\Gamma}_j(S,h)$ gives an $\ell\in S$ with $a_\ell^*<A_m$ and

$$\nabla_a\widehat{\mathcal{R}}_\rho(a^*,h)^\top(\mathsf{e}_\ell-\mathsf{e}_j)<0.$$

For small $\epsilon>0$, the feasible point $a^*+\epsilon(\mathsf{e}_\ell-\mathsf{e}_j)$ has smaller criterion, a contradiction. Every global minimizer over $\mathcal{A}_\delta$ thus lies on the restricted polytope.

This polytope has affine dimension $s-1$. Partition it into cells with Euclidean radius at most $\eta/(2L_R)$. A regular covering uses at most $C_s\left\{\frac{L_R\,\mathrm{diam}(\mathcal{A}_\delta)}{\eta}\right\}^{s-1}$ cells. Because its infimum equals the full infimum, Proposition 19 bounds the original global objective gap by $2L_R\overline{r}_t\le\eta$.

Under the population exchange condition, the difference between an empirical exchange derivative and its population value is at most twice the uniform gradient error. An error smaller than $\chi/4$ gives $\widehat{\Gamma}_j(S,h)\le-\chi/2$ for every excluded coordinate. This proves the final assertion and completes the proof. □

Proof of Proposition 17. Lemma 4 gives

$$\max_{r,t}\Big|\frac{\partial^2\widehat{\mathcal{R}}_0(a,h)}{\partial a_r\partial a_t}\Big|\le C_Hm^{-2}.$$

The Euclidean norm of the difference between two rows of this Hessian is at most $2C_Hm^{-3/2}$. The corresponding row difference for the ridge Hessian $(\rho/m)I_m$ has norm $\sqrt{2}\,\rho/m$. Hence

$$\widehat{H}_{\ell j}(h)\le H_m$$

uniformly over a, the observations, and the masks.

Fix one block and write $k_-\!>0$ for the uniform lower bound on its off-diagonal kernel entries. Let u be its exact scaling vector. The sum of the entries of W outside row and column i is $b-2$. The row equations and $k_-\le K_{ij}\le1$ give

$$b-2\le\Big(\sum_{j\ne i}u_j\Big)^2,\qquad\frac{k_-}{\sqrt{3b}}\le u_i\le\frac{1}{k_-\sqrt{b-2}}\le\frac{\sqrt{3}}{k_-\sqrt{b}}.$$

For the lower bound, use $1 \le u_i(b-1)\max_j u_j$ and the upper bound just proved. Put $A = \log(2\sqrt{3}/k_-)$ and $z_i = \log(\sqrt{b}\,u_i)$. The scaling root lies in the interior of $[-A, A]^b$ and minimizes

$$F(z) = \frac{1}{2b}\sum_{i\neq j} K_{ij} e^{z_i+z_j} - \sum_i z_i.$$

For every z in this box and $v \in \mathbb{R}^b$,

$$v^\top \nabla^2 F(z) v = \frac{1}{b}\sum_{i<j} K_{ij} e^{z_i+z_j}(v_i+v_j)^2, \qquad c\|v\|_2^2 \le v^\top \nabla^2 F(z) v \le C\|v\|_2^2.$$

Here the constants are independent of b, because $\sum_{i<j}(v_i+v_j)^2 = (b-2)\|v\|_2^2 + (\mathbf{1}^\top v)^2$ and $b \ge 3$. Projected gradient descent on this box contracts Euclidean distance to the root by a fixed factor. Starting at zero, $C\log(b/\epsilon)$ iterations give maximum-norm error at most $c\epsilon$. Each iteration costs $O(b^2)$ operations.

The same lower kernel bound gives $W_{ij} \ge k_-^3/(3b)$, and hence $\lambda_{\min}(I_b + W) \ge k_-^3/9$. The upper eigenvalue is at most two. In Proposition 14, the adjoint and its right side have maximum norm $O(b^{-1})$. Multiplying the system by b gives a bounded right side; contracting linear iteration attains accuracy $c\epsilon$ in $C\log(b/\epsilon)$ steps. Lemma 4, the differentiated row equations, and $|m\partial_{a_j}\log K_{i\ell}| \le C$ bound loss and scaled-score errors by $C\epsilon$ when log-scaling and scaled-adjoint errors are at most ϵ. Decreasing the tolerance by this constant and using residual bounds gives certified evaluation intervals.

Pairwise distances and accumulation of all coordinate gradients cost $O(b^2 m)$ per block under coordinate masking. A prescribed response adds $O(b^2 q_V)$ for response matrix products, bounded by the stated factor $1 + q_V/m$. A dense linear-system step costs at most $O(b^3)$. One such step over all blocks and bandwidths has cost $O\{|\mathcal{H}|R(b^3 + b^2 m)\}$. The complete evaluation multiplies this bound by at most $C\log(b/\epsilon)$. Checking all candidate and excluded coordinate pairs costs at most $O(|\mathcal{H}|m^2)$. Use outward objective intervals and restrict cell radii to $\eta/(4L_R)$. The certified gap is at most $2L_R\overline{r}_t + 4\epsilon \le \eta$. Proposition 16 then requires at most $C_s\left\{\frac{L_R\,\mathrm{diam}(\mathcal{A}_\delta)}{\eta}\right\}^{s-1}$ restricted cells for each bandwidth, after increasing C_s. The cost $CRC_{\mathrm{ev}}(b,m)\log(b/\epsilon)$ per cell and the one-box calculation give the operation bound. Numerical score errors tending to zero preserve fixed strict population margins.

Under the population margin, an empirical gradient error smaller than $\chi/4$ changes each derivative difference by at most $\chi/2$. Substitution of the deterministic Hessian enclosure gives

$$\max_{j\notin S}\max_{\ell\in S}\Big[\frac{\partial\widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_\ell} - \frac{\partial\widehat{\mathcal{R}}_\rho(\mathbf{1},h)}{\partial a_j} + H_m\,\mathrm{diam}(\mathcal{A}_\delta)\Big] \le -\chi/2.$$

The assumed candidate-set and gradient convergence probabilities prove the final assertion. This completes the proof. □

Proof of Proposition 25. Fix a mask and write its observed variables as

$$X_\xi = (X_{\xi\cap\mathcal{S}}, X_{\xi\cap\mathcal{N}}).$$

The product distribution and the diagonal Gaussian kernel give

$$P_\xi = P_{\xi,\mathcal{S}} \otimes \bigotimes_{j\in\xi\cap\mathcal{N}} P_j, \qquad k_{a,h,\xi} = k^{\mathcal{S}}_{a,h,\xi}\prod_{j\in\xi\cap\mathcal{N}} k_{a_j,h,j}.$$

The second equality denotes multiplication of the signal kernel and the coordinatewise nuisance kernels. Uniqueness of each positive symmetric scaling gives

$$T_{a,h,\xi} = T^{\mathcal{S}}_{a,h,\xi} \otimes \bigotimes_{j\in\xi\cap\mathcal{N}} T_{a_j,h,j}.$$

Every factor preserves the constant function. The conditional mean of a held-out signal coordinate is measurable with respect to $X_{\xi\cap\mathcal{S}}$, while the conditional mean of a held-out nuisance coordinate is its marginal mean. It follows that

$$T_{a,h,\xi} m_{\xi,\mathcal{S}} = T^{\mathcal{S}}_{a,h,\xi} m_{\xi,\mathcal{S}}, \qquad T_{a,h,\xi}\mathbb{E}X_j = \mathbb{E}X_j, \quad j \in \xi^c\cap\mathcal{N}.$$

Proposition 21, followed by averaging over the masks, proves

$$\mathcal{R}_{\infty,\rho}(a,h) = \mathcal{V} + Q_h(a_{\mathcal{S}}) + \frac{\rho}{2m}\|a - \mathbf{1}\|_2^2.$$

Put $a^\circ = (u^\circ, \delta \mathbf{1}_q)$. The signal stationarity equations hold with equality. The lower-bound multiplier of each nuisance coordinate equals

$$\lambda^\circ + \frac{\rho}{m}(\delta - 1) > 0.$$

Thus a° satisfies the Karush–Kuhn–Tucker equations of the convex objective. Strong convexity gives uniqueness when $\rho > 0$. When $\rho = 0$, signal curvature fixes the signal vector, whose coordinates sum to $m - q\delta$; the trace and lower bounds fix all nuisance coordinates at δ. Hence a° is unique.

Strict complementarity makes every nuisance component of a critical direction equal zero. The remaining signal direction lies in $\mathbf{1}_s^\perp$, where the Hessian is bounded below by $\kappa_{\mathcal{S}}$. This proves the critical-cone assertion. Differentiating the decomposition at $a = \mathbf{1}$ gives the screening scores. This completes the proof. □

Proof of Proposition 26. For a fixed paired mask, each held-out signal coordinate has its partner in the observed view. Mutual independence of the signal pairs and nuisance coordinates factorizes the population operator over all observed coordinates. If j is observed and $\bar{j}$ is held out, every operator except T_{j,a_j} acts on a constant factor. Hence

$$\|m_{\xi,\mathcal{S}} - T^{\mathcal{S}}_{a,h,\xi} m_{\xi,\mathcal{S}}\|^2_{L^2(P)} = \sum_{\substack{j \in \mathcal{S} \\ j \text{ observed}}} r_j(a_j).$$

Each orientation has probability $1/2$, and $|\xi^c| = m/2$. Averaging gives

$$Q_h(a_{\mathcal{S}}) = \sum_{j \in \mathcal{S}} \omega_j r_j(a_j).$$

The curvature and multiplier inequalities now verify Proposition 25. At $a = \mathbf{1}$, direct differentiation gives

$$d_j(h) = -m\omega_j r_j'(1) > 0, \quad j \in \mathcal{S}, \qquad d_j(h) = 0, \quad j \in \mathcal{N}.$$

Let $a_j > \delta$ for a nuisance coordinate. The trace constraint gives

$$\sum_{\ell \in \mathcal{S}} a_\ell < m - q\delta = \sum_{\ell \in \mathcal{S}} u^\circ_\ell.$$

Some $\ell \in \mathcal{S}$ satisfies $a_\ell < u^\circ_\ell$. The coordinatewise curvature bound and the stationarity equation at u° imply

$$\omega_\ell r_\ell'(a_\ell) + \frac{\rho}{m}(a_\ell - 1) < -\lambda^\circ.$$

The nuisance derivative is at least $\rho(\delta - 1)/m$. Subtracting the two derivatives gives the stated bound with $-\chi$. This completes the proof. □

Proof of Corollary 10. Throughout this proof, the coordinate kernels use $\pi = k/m$ in their exponent. The functions r_j and their derivatives are evaluated for that design. For a signal coordinate j, its risk term appears exactly when j is observed and $\bar{j}$ is held out. Proposition 11 gives

$$\mathbb{P}(\xi_j = 1, \xi_{\bar{j}} = 0) = \frac{k(m-k)}{m(m-1)}.$$

Every mask has $m - k$ held-out coordinates. The product-operator calculation in Proposition 26 then gives

$$Q_h(a_{\mathcal{S}}) = \sum_{j \in \mathcal{S}} \frac{k(m-k)}{m(m-1)(m-k)} r_j(a_j) = \sum_{j \in \mathcal{S}} \omega_k r_j(a_j).$$

The Karush–Kuhn–Tucker and exchange calculations in the proof of Proposition 26 apply with the stated replacements. This completes the proof. □

Proof of Proposition 27. For $h > h_0$ and every feasible a, the mean value theorem gives

$$Q_h(a_{\mathcal{S}}) - Q_{h_0}(a_{\mathcal{S}}) = \sum_{j\in\mathcal{S}} \omega_j \left\{\widetilde{r}_j(a_j/h^2) - \widetilde{r}_j(a_j/h_0^2)\right\}$$
$$\geq c_h \sum_{j\in\mathcal{S}} \omega_j a_j \left(h_0^{-2} - h^{-2}\right) \geq c_h \delta \sum_{j\in\mathcal{S}} \omega_j \left(h_0^{-2} - h^{-2}\right).$$

The ridge and irreducible variance do not depend on h. Taking the infimum over a and the minimum over $h > h_0$ proves the stated bandwidth gap. It also shows that a joint minimizer must use h_0. Proposition 25 gives the unique metric minimizer $a_0 = (u^\circ, \delta \mathbf{1}_q)$ at this bandwidth.

For local quadratic growth, write $t = a_{\mathcal{N}} - \delta \mathbf{1}_q \geq 0$ and $T = \mathbf{1}^\top t$. Feasibility gives

$$\mathbf{1}^\top (a_{\mathcal{S}} - u^\circ) = -T.$$

Strong convexity of F_{h_0}, its stationarity equation, and the nuisance ridge terms give

$$\mathcal{R}_{\infty,\rho}(a, h_0) - \mathcal{R}_{\infty,\rho}(a_0, h_0)$$
$$\geq \frac{\kappa_{\mathcal{S}}}{2}\|a_{\mathcal{S}} - u^\circ\|_2^2 + \left\{\lambda^\circ + \frac{\rho}{m}(\delta - 1)\right\} T + \frac{\rho}{2m}\|t\|_2^2.$$

The multiplier in braces is positive. On a ball of radius r_0, $T \geq \|t\|_2 \geq r_0^{-1}\|t\|_2^2$. Decreasing r_0 if needed yields the asserted quadratic inequality for some $\kappa > 0$. This completes the proof. □

Proof of Theorem 10. Fix $\xi \in \Xi_1$. The factorization in the proof of Proposition 25 gives

$$T_{a,h,\xi}\psi_{\xi,\mathcal{S}} = T^{\mathcal{S}}_{a,h,\xi}\psi_{\xi,\mathcal{S}}.$$

The nuisance factors do not enter the right side. For a held-out nuisance coordinate, its conditional mean is constant and is fixed by $T_{a,h,\xi}$. For $\xi \notin \Xi_1$, the complete conditional-mean vector is constant after the held-out nuisance coordinates are included, and its approximation error equals zero.

For $\xi \in \Xi_1$, Lemma 5 applied to the embedded signal view $\phi_{\xi,\mathcal{S}}$ gives

$$T^{\mathcal{S}}_{a,h,\xi} f = f - h^2 \mathcal{B}^{\mathcal{S}}_{a,\xi} f + O(h^4)$$

in $L^2(P)$, uniformly over $a \in \mathcal{A}_\delta$. The mask support is finite, and the metric coefficients, density derivatives, and ellipticity constants are uniform on the compact parameter set. Proposition 21 now gives

$$\mathcal{R}_{\infty,0}(a, h) - \mathcal{V} = \sum_{\xi\in\Xi_1} \frac{p_\xi}{|\xi^c|}\|\psi_{\xi,\mathcal{S}} - T^{\mathcal{S}}_{a,h,\xi}\psi_{\xi,\mathcal{S}}\|^2_{L^2(P)} = h^4 \mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) + O(h^6),$$

uniformly over $\mathcal{A}_\delta$. This proves the first assertion. Division by h^4 gives uniform convergence of the penalized criterion to

$$\mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) + \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2.$$

Compactness and the argmin theorem give the remaining assertions. This completes the proof. □

Proof of Corollary 11. The proof of Theorem 10 gives

$$\sup_{a\in\mathcal{A}_\delta} \left| h^{-4}\{\mathcal{R}_{\infty,\rho_h}(a, h) - \mathcal{V}\} - \mathcal{J}_{\mathrm{mask}}(a_{\mathcal{S}}) - \frac{\rho_\star}{2m}\|a - \mathbf{1}\|_2^2 \right| \leq C\left\{h^2 + \left|\frac{\rho_h}{h^4} - \rho_\star\right|\right\}.$$

The minimizing property of a_h and quadratic growth imply

$$\kappa_{\mathrm{mask}}\|a_h - a_\star^{\mathrm{mask}}\|_2^2 \leq C\left\{h^2 + \left|\frac{\rho_h}{h^4} - \rho_\star\right|\right\}.$$

This proves the parameter bound. Compactness of $\mathcal{M}$ gives

$$\sup_z \|g^{\mathcal{S}}_{a_h}(z) - g^{\mathcal{S}}_{a_\star^{\mathrm{mask}}}(z)\|_{\mathrm{op}} \leq C\|a_h - a_\star^{\mathrm{mask}}\|_2.$$

Uniform ellipticity bounds the corresponding curve lengths in both directions. Taking the infimum over curves proves the geodesic bound. This completes the proof. □

The binary model proof verifies the role of repeated information without a local approximation. Its tensor product form is a direct consequence of the product distribution.

Proof of Proposition 28. The signal kernel has value one when two latent states agree and $r = \exp(-2a_1/h^2)$ when they differ. Under the uniform distribution on the two states, its bistochastic Markov transition probabilities are

$$p(z,z) = \frac{1}{1+r}, \qquad p(z,-z) = \frac{r}{1+r}.$$

The product measure factorizes the scaling and normalized operator. The nuisance operator preserves constants, giving

$$T_a Z = \frac{1-r}{1+r} Z = \tanh(a_1/h^2) Z.$$

The independent noise in V has conditional mean zero. The graph predictor of V is the same expression. Squaring $V - T_a V$ and taking expectations gives $\{1 - \tanh(a_1/h^2)\}^2 + \sigma^2$. This quantity is strictly decreasing in a_1. Under $a_1 + a_2 = 2$ and the lower bounds, its minimum is attained at $a_1 = 2 - \delta$ and $a_2 = \delta$. This completes the proof. □

K.1 Interior and boundary inference in a bounded model

A fixed bounded distribution permits direct comparison of an interior metric, a weakly active boundary, and a strictly active boundary. Let Z, N, and ε be independent Rademacher variables, let the observed view be (Z, N), and let $V = Z + \sigma\varepsilon$, where $\sigma > 0$. Use a fixed bandwidth $h > 0$ and $0 < \delta < 1$. For $t \in [\delta, 2-\delta]$, define

$$r(t) = \tanh(t/h^2), \qquad f_\rho(t) = \{1 - r(t)\}^2 + \sigma^2 + \frac{\rho}{2}(t-1)^2, \qquad \rho_c = -\frac{f_0'(2-\delta)}{1-\delta}.$$

The function f_ρ is the population risk on the trace constraint. The following proposition verifies identification, curvature, and the observation covariance without an additional uniqueness assumption.

Proposition 43. *In the bounded binary response model, for every $\rho \geq 0$,*

$$f_\rho''(t) = \frac{2}{h^4}\{1 - r(t)\}\{1 - r(t)^2\}\{1 + 3r(t)\} + \rho > 0.$$

The unique population metric satisfies

$$a_{01} = \begin{cases} 2-\delta, & 0 \leq \rho \leq \rho_c, \\ \textit{the unique zero of } f_\rho' \textit{ in } (\delta, 2-\delta), & \rho > \rho_c, \end{cases} \qquad a_{02} = 2 - a_{01}.$$

At $\rho \leq \rho_c$, the nuisance multiplier is $(\rho_c - \rho)(1-\delta)$. The observation influence and covariance are

$$\psi_0(Z,N,V) = -4\sigma r'(a_{01})\{1 - r(a_{01})\} Z\varepsilon \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \Omega_0 = \begin{pmatrix} 16\sigma^2 r'(a_{01})^2 \{1 - r(a_{01})\}^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Proposition 43 includes a weak boundary with singular score covariance. Curvature holds on the entire trace tangent space, which verifies the curvature premise of Theorem 16. Put $\tau = \sqrt{\Omega_{0,11}}/f_\rho''(a_{01})$ and let $G \sim N(0,1)$. Under the sample-size and optimization conditions of that theorem, the first coordinate has limit

$$\sqrt{Rb_R}(\widehat{a}_{\mathrm{obs},1} - a_{01}) \Longrightarrow \begin{cases} 0, & 0 \leq \rho < \rho_c, \\ \min(0, \tau G), & \rho = \rho_c, \\ \tau G, & \rho > \rho_c. \end{cases}$$

At the weak boundary, half the limiting probability is at zero. At a strictly active boundary, the first-order limit is a point mass; a zero-width interval is not a nominal confidence interval. The numerical analysis evaluates coverage only for the interior and weak-boundary regimes and reports selection frequency in the strictly active regime.

Proof of Proposition 43. Proposition 28 gives f_ρ. Differentiation gives

$$f'_\rho(t) = -\frac{2}{h^2}\{1 - r(t)\}\{1 - r(t)^2\} + \rho(t-1).$$

A second derivative gives the asserted positive expression. Also, $f'_\rho(\delta) < 0$ and $f'_\rho(2-\delta) = (\rho - \rho_c)(1-\delta)$. Strict convexity proves the minimizer and multiplier assertions. For a trace tangent vector $v = (s, -s)$,

$$v^\top H_0 v = f''_\rho(a_{01})s^2 = \frac{f''_\rho(a_{01})}{2}\|v\|_2^2.$$

At the product distribution, $m(Z, N) = Z$ and $Tm = r(a_{01})m$. Sign changes of Z and N act transitively on the four states; the corresponding change of V preserves squared loss. The four centered score derivatives are equal and sum to zero. Thus perturbing the distribution of (Z, N) with the conditional response distribution fixed gives zero derivative.

For a perturbation $\dot{m}$ of the conditional mean with the observed distribution fixed, the part of the risk that depends on the metric is $-2\langle m, Tm\rangle + \langle Tm, Tm\rangle$. The operator and its metric derivatives are self-adjoint. On m, the first-coordinate derivative equals $r'(a_{01})m$, and the second-coordinate derivative equals zero. Differentiating first in m and then in the metric gives

$$D_m \nabla_a \mathcal{R}_{\infty,\rho}[\dot{m}] = -4r'(a_{01})\{1 - r(a_{01})\}\langle Z, \dot{m}\rangle \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

An observation (z, n, v) contributes $z(v - z) = \sigma z\varepsilon$ to this inner product. Conditional variance contributes only to the metric-independent term $\mathbb{E}V^2$. This proves the influence formula. Its mean is zero, and $\mathbb{E}(Z\varepsilon)^2 = 1$ gives the covariance. This completes the proof. □

K.2 Exact computation for repeated observations

Repeated observed states reduce the dimension of the scaling equations without changing the estimator. In a block of size b, let n_g be the number of observations at the gth value of (Z, N), for $1 \le g \le 4$, and let S_{1g} and S_{2g} be the sums of V and V^2 in that group. Write $K_{g\ell}$ for the Gaussian kernel between the two observed states. Uniqueness of symmetric scaling makes its factor constant within each nonempty group. Set $u_g = 0$ for an empty group; its count and response sums are zero. For each nonempty group, the row equation is

$$u_g\Big(\sum_{\ell=1}^{4} n_\ell K_{g\ell} u_\ell - u_g\Big) = 1.$$

The subtraction of u_g deletes each observation's own diagonal. Put $Q_g = u_g \sum_\ell K_{g\ell} u_\ell S_{1\ell}$. Then the exact block loss is

$$\ell_b^V(a, h) = \frac{1}{b}\sum_{g=1}^{4}\left\{(1 + u_g^2)^2 S_{2g} - 2(1 + u_g^2)Q_g S_{1g} + n_g Q_g^2\right\}.$$

This expression equals the loss obtained from all b individual observations. Deleting one observation changes only its group's count and response sums. The block-jackknife calculation requires at most eight distinct deletions per block in the Rademacher model.

The same representation gives a global curvature certificate. Suppose all four observed groups contain at least three observations. Put $n_* = \min_g n_g$, $p_* = n_*/b$, and $\kappa_b = 1 - 1/(n_* - 1)$. Let B_V bound the centered responses and let $F_b(t) = \ell_b^V((t, 2-t), h) + \rho(t-1)^2/2$. For R blocks, put $F_R = R^{-1}\sum_r F_{b,r}$ and write M_3 for the average of the block-specific third-derivative bounds given below. Let Δ be the maximum spacing of a grid that includes both interval ends.

Proposition 44. *For every $t \in [\delta, 2-\delta]$,*

$$|F'''_b(t)| \le B_V^2\left(\frac{2}{h^2}\right)^3\left(1 + \frac{2}{\kappa_b\sqrt{p_*}}\right)\left(10 + \frac{32}{\kappa_b\sqrt{p_*}}\right).$$

The empirical curvature satisfies

$$\inf_{t\in[\delta,2-\delta]} F''_R(t) \ge \min_{t \text{ on the grid}} F''_R(t) - \frac{1}{2}M_3\Delta.$$

Proposition 44 controls curvature between the evaluated points. A positive lower bound makes the scalar score equation sufficient for global optimization. Proposition 19 then gives the objective gap from the normal-cone residual. This calculation supplies the optimization tolerance used by the inference experiment. The group-count condition is a computational check, not an additional population assumption. In the bounded binary model, each observed-state count is binomial with success probability $1/4$. For $b \geq 16$, the Chernoff bound and a union bound give $\mathbb{P}(\min_{r,g} n_{rg} < 3) \leq 4R\exp(-b/32)$. Thus the sampling model controls the probability that this check fails; global subdivision remains available for any block whose counts do not pass it.

Proof of Proposition 44. The row equation gives $u_g^2 \leq 1/(n_g - 1)$. The full Gaussian matrix is positive semidefinite. After symmetric scaling and diagonal deletion,

$$\lambda_{\min}(I_b + W) \geq 1 - \max_g u_g^2 \geq \kappa_b.$$

Put $c = 2/h^2$ and $C_1 = (\kappa_b\sqrt{p_*})^{-1}$. For differentiation in t, the absolute log-kernel derivative is at most c. The differentiated row equations and constancy within groups give $\|\dot{z}\|_\infty \leq cC_1$, where $z = \log u$. Put $A_{ij} = \partial_t \log K_{ij} + \dot{z}_i + \dot{z}_j$ and $C_2 = c(1+2C_1)$. Then $\|A\|_{\max} \leq C_2$. The weighted projection calculation in Proposition 20 gives

$$\sum_{i \neq j} W_{ij} A_{ij}^2 \leq bc^2.$$

The second differentiated row equation is $(I_b + W)\ddot{z} = -(W \circ A^2)\mathbf{1}$. Jensen's inequality gives $\|\ddot{z}\|_\infty \leq C_2 c C_1$, where $\circ$ denotes entrywise multiplication. The row-constrained derivatives are weighted orthogonal projections. Their weighted norms give

$$\|\dot{W}V\|_2 \leq B_V\sqrt{b}\,c, \qquad \|\ddot{W}V\|_2 \leq B_V\sqrt{b}\,C_2 c,$$

and

$$\|W^{(3)}V\|_2 \leq B_V\sqrt{b}\,c(C_2^2 + 6C_2 c C_1).$$

The final inequality follows by projecting $A^3 + 3A \circ (\ddot{z}\mathbf{1}^\top + \mathbf{1}\ddot{z}^\top)$ and using the two preceding bounds. Differentiating the squared residual three times now yields

$$|F_b'''(t)| \leq 6B_V^2 C_2 c^2 + 4B_V^2 c(C_2^2 + 6C_2 c C_1) = B_V^2 c^3(1 + 2C_1)(10 + 32C_1).$$

The ridge has zero third derivative. Each point of the interval is within $\Delta/2$ of a grid point. The mean-value theorem proves the second assertion. This completes the proof. □

When every observed group has at least three observations, the observed sum of response squares gives a global objective bound without a population response bound. Use c, C_1, and C_2 from the preceding proof. The differentiated row equations give $\|z'\|_\infty \leq cC_1$ and $\|z''\|_\infty \leq C_1 C_2^2$. Symmetry and the unit row sums of W imply

$$\|W'\|_{\mathrm{op}} \leq C_2, \qquad \|W''\|_{\mathrm{op}} \leq C_2^2(1 + 2C_1).$$

Because $\|I_b - W\|_{\mathrm{op}} \leq 2$, differentiation of the squared prediction error yields

$$|F_b''(t)| \leq \frac{2\|V\|_2^2}{b} C_2^2(3 + 4C_1) + \rho =: M_{2,b}.$$

This bound holds throughout $[\delta, 2 - \delta]$ for every realized response vector. Let $\widetilde{F}_b(s)$ and $\widetilde{F}_b'(s)$ have numerical error bounds e_0 and e_1. On an interval with center s and radius r, Taylor's theorem gives

$$\inf_{|t-s| \leq r} F_b(t) \geq \widetilde{F}_b(s) - e_0 - \big\{|\widetilde{F}_b'(s)| + e_1\big\}r - \frac{1}{2}M_{2,b}r^2.$$

Taking the smallest lower bound over a partition of the feasible interval gives a global lower bound. The smallest evaluated value plus its numerical error gives an upper bound. Subdivision continues until their difference is at most the prescribed optimization tolerance. This calculation applies when the sample curvature is not positive and uses no population bound on the responses.

For numerical certification, errors in the scaling and differentiated row equations are propagated before applying the grid bound. Write $J(z) = \mathrm{diag}\{W(z)\mathbf{1}\} + W(z)$. If $\widetilde{z}$ is a computed log

scaling and $J(\widetilde{z})$ has lower eigenvalue $\lambda > 0$, a radius $r = 2\|W(\widetilde{z})\mathbf{1} - \mathbf{1}\|_2/\lambda \le \log(2)/2$ contains the exact scaling. Indeed, $J(\widetilde{z}+s) \succeq e^{-2r}J(\widetilde{z})$ for $\|s\|_2 \le r$, and the radial derivative is nonnegative on the boundary of this ball. Strict convexity places the unique root inside it. Thus each affinity entry lies between its computed value multiplied by e^{-2r} and by e^{2r}. The first and second differentiated equations use the inverse bound κ_b^{-1} to propagate this error into the score and Hessian. Subtracting the resulting Hessian error at each grid point preserves the lower bound in Proposition 44.

The error propagation can be evaluated using four-dimensional group quantities. Write $\widetilde{W} = W(\widetilde{z})$, $s_W = \max_i \sum_j \widetilde{W}_{ij}$, $c = 2/h^2$, and $e_W = s_W(e^{2r} - 1)$. Then $\max_i \sum_j |W_{ij} - \widetilde{W}_{ij}| \le e_W$. Let $\widetilde{z}_1$ and $\widetilde{z}_2$ be the computed derivatives and put $q_j = \|\widetilde{z}_j\|_\infty$. Define $H_{ij} = \partial_t \log K_{ij}$ off the diagonal and $H_{ii} = 0$, and let $\widetilde{A}_{ij} = H_{ij} + \widetilde{z}_{1i} + \widetilde{z}_{1j}$. The residual bounds used here are

$$d_1 \ge \|(I + \widetilde{W})\widetilde{z}_1 + (\widetilde{W} \circ H)\mathbf{1}\|_\infty, \quad d_2 \ge \|(I + \widetilde{W})\widetilde{z}_2 + (\widetilde{W} \circ \widetilde{A}^2)\mathbf{1}\|_\infty.$$

If $\widetilde{d}_j$ bounds the residual with $\operatorname{diag}(\widetilde{W}\mathbf{1}) + \widetilde{W}$ in place of $I + \widetilde{W}$, one may use $d_j = \widetilde{d}_j + \|\widetilde{W}\mathbf{1} - \mathbf{1}\|_\infty q_j$. Put $A_1 = c + 2q_1$ and $A_2 = A_1^2 + 2q_2$. Subtracting the exact and computed equations gives

$$\|z' - \widetilde{z}_1\|_\infty \le e_1 = \frac{e_W(c + q_1) + d_1}{\kappa_b\sqrt{p_*}},$$

and

$$\|z'' - \widetilde{z}_2\|_\infty \le e_2 = \frac{e_W(A_1^2 + q_2) + 4A_1e_1 + 4e_1^2 + d_2}{\kappa_b\sqrt{p_*}}.$$

The conversion from the weighted group norm to maximum norm contributes $p_*^{-1/2}$. Define

$$E_0 = B_V e_W, \qquad E_1 = B_V(e_W A_1 + 2e_1), \qquad E_2 = B_V(e_W A_2 + 4A_1e_1 + 4e_1^2 + 2e_2),$$

and put $R_0 = B_V(1 + s_W)$, $R_1 = B_V s_W A_1$, and $R_2 = B_V s_W A_2$. These quantities bound the errors and computed magnitudes of the residual and its first two derivatives. Expanding the two products in the differentiated squared loss yields

$$|F_b' - \widetilde{F}_b'| \le 2(R_0E_1 + R_1E_0 + E_0E_1),$$
$$|F_b'' - \widetilde{F}_b''| \le 2(2R_1E_1 + E_1^2 + R_0E_2 + R_2E_0 + E_0E_2).$$

The ridge derivatives are exact analytic terms. Averaging these bounds over the independent blocks gives the score and Hessian error bounds for the empirical criterion.

L A Gaussian repeated-measurement model

The binary calculation in Proposition 28 isolates the identification mechanism without integration. A Gaussian model gives a continuous counterpart and shows how measurement error changes the strength, but not the direction, of metric learning.

Use the same prescribed view split as in Proposition 28. Let

$$Z \sim N(0, 1), \qquad X = Z + \varepsilon_1, \qquad V = Z + \varepsilon_2,$$

where $\varepsilon_1 \sim N(0, \sigma_1^2)$ and $\varepsilon_2 \sim N(0, \sigma_2^2)$ are independent of each other and of Z. Let N be an independent nuisance coordinate. The observed view is (X, N) and V is held out. Use a product Gaussian kernel with weights $a_1 + a_2 = 2$, where a_1 multiplies $(X - X')^2$ and a_2 multiplies $(N - N')^2$.

Put $v_1 = 1 + \sigma_1^2$, $c = a_1/h^2$, and

$$r(c) = \frac{\sqrt{1 + 4c^2v_1^2} - 1}{2cv_1} = \frac{2cv_1}{1 + \sqrt{1 + 4c^2v_1^2}}.$$

The value at $c = 0$ is defined by continuity as zero.

Proposition 45. *In the Gaussian model above, the population graph predictor and its risk satisfy*

$$T_aV(x, n) = \frac{r(c)}{v_1}x, \qquad \mathbb{E}\{V - T_aV(X, N)\}^2 = 1 + \sigma_2^2 - \frac{2r(c) - r(c)^2}{v_1}.$$

The risk is strictly decreasing in $a_1 > 0$. Its unique minimizer over $\mathcal{A}_\delta$ is $(a_1, a_2) = (2 - \delta, \delta)$.

Proposition 45 gives an exact continuous model in which prediction removes an independent nuisance coordinate. A larger σ_1^2 weakens the conditional mean of V given X through the factor v_1^{-1}. A larger σ_2^2 raises only the irreducible error and does not change the optimal metric.

Proof of Proposition 45. The conditional mean of the held-out measurement is

$$\mathbb{E}(V \mid X = x, N = n) = \mathbb{E}(Z \mid X = x) = \frac{x}{v_1}.$$

For $r \in (0, 1)$, let

$$q_r(y \mid x) = \frac{1}{\sqrt{2\pi v_1(1 - r^2)}} \exp\left\{ -\frac{(y - rx)^2}{2v_1(1 - r^2)} \right\}.$$

This transition is reversible with invariant distribution $N(0, v_1)$. If p_1 denotes its density, then

$$\log \frac{q_r(y \mid x)}{p_1(y)} = -\frac{r^2(x^2 + y^2)}{2v_1(1 - r^2)} + \frac{rxy}{v_1(1 - r^2)} - \frac{1}{2}\log(1 - r^2).$$

The log of the unnormalized signal kernel is

$$-\frac{c}{2}(x - y)^2 = -\frac{c}{2}(x^2 + y^2) + cxy.$$

Choose r to satisfy

$$c = \frac{r}{v_1(1 - r^2)}.$$

The two log kernels then differ by a function of x, the same function of y, and a constant. Thus $q_r(y \mid x)/p_1(y)$ is a symmetric scaling of the signal kernel. Uniqueness identifies it as the population bistochastic kernel. Solving the quadratic equation for r gives the expression stated above. Its Markov operator maps x to rx.

The nuisance kernel factorizes from the signal kernel. Proposition 1 gives

$$T_a V(x, n) = T_{a_1}\{x/v_1\} = r(c)x/v_1.$$

Because $\mathrm{var}(V) = 1 + \sigma_2^2$, $\mathrm{cov}(V, X) = 1$, and $\mathrm{var}(X) = v_1$,

$$\mathbb{E}\{V - r(c)X/v_1\}^2 = 1 + \sigma_2^2 - \frac{2r(c) - r(c)^2}{v_1}.$$

The map $r \mapsto r/(1 - r^2)$ is strictly increasing on $(0, 1)$, and hence $r(c)$ is strictly increasing in c. The risk derivative with respect to r equals $-2(1 - r)/v_1 < 0$. Thus the risk is strictly decreasing in a_1. The trace and lower-bound constraints give the stated minimizer. This completes the proof. □

Proof of Proposition 29. The product distribution and kernel factorize the scaling and Markov operator. Each coordinatewise operator preserves constants, giving

$$T_a \mu_\ell(U_\ell) = T_{\ell, a_\ell} \mu_\ell(U_\ell).$$

The regularized risk equals

$$\sum_{\ell=1}^{s} \omega_\ell r_\ell(a_\ell) + \frac{\rho}{2m} \sum_{j=1}^{m} (a_j - 1)^2.$$

It has no other dependence on the nuisance coordinates.

At $a_0 = (a_S^\circ, \delta \mathbf{1}_q)$, the signal-coordinate Karush–Kuhn–Tucker equations hold with equality by the definition of λ°. The lower-bound multiplier for every nuisance coordinate equals

$$\lambda^\circ + \frac{\rho}{m}(\delta - 1) > 0.$$

The Karush–Kuhn–Tucker equations and convexity make a_0 a global minimizer with strictly complementary nuisance constraints. Signal curvature fixes the signal vector at every minimizer; the trace and lower bounds then give $a_j = \delta$ on all nuisance coordinates. Hence the minimizer is unique.

Strict complementarity sets every nuisance direction in the critical cone equal to zero. On the remaining signal tangent space, the Hessian is diagonal with entries at least κ_S. This proves positive definiteness on the critical cone. At $a = \mathbf{1}$, the unpenalized derivative equals $\omega_\ell r'_\ell(1)$ on a signal coordinate and zero on a nuisance coordinate. Multiplication by $-m$ gives the stated screening scores.

If $a_j > \delta$ for a nuisance coordinate, then

$$\sum_{\ell=1}^{s} a_\ell < m - q\delta = \sum_{\ell=1}^{s} a^\circ_{S\ell}.$$

Some signal coordinate satisfies $a_\ell < a^\circ_{S\ell} < A_m$. The curvature condition makes its derivative strictly smaller than $-\lambda^\circ$, while the nuisance derivative is at least $\rho(\delta - 1)/m$. Hence

$$\frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_\ell} - \frac{\partial \mathcal{R}_{\infty,\rho}(a,h)}{\partial a_j} \le -\lambda^\circ - \frac{\rho}{m}(\delta - 1) = -\chi.$$

This completes the proof. □

Proof of Proposition 18. Lemma 4 gives

$$\max_{r,t} \left| \frac{\partial^2 \widehat{\mathcal{R}}_0(a,h)}{\partial a_r \partial a_t} \right| \le C_H m^{-2}.$$

For $g_{\ell j}(a) = \partial_\ell \widehat{\mathcal{R}}_\rho(a,h) - \partial_j \widehat{\mathcal{R}}_\rho(a,h)$, the data part of every coordinate derivative is bounded by $2C_H m^{-2}$. The ridge Hessian is $(\rho/m)I_m$, which gives

$$\|\nabla g_{\ell j}(a)\|_\infty \le 2C_H m^{-2} + \rho/m = H^{(1)}_m.$$

For every $a \in \mathcal{A}_\delta$,

$$g_{\ell j}(a) \le g_{\ell j}(\mathbf{1}) + H^{(1)}_m \|a - \mathbf{1}\|_1 \le g_{\ell j}(\mathbf{1}) + 2m(1-\delta)H^{(1)}_m.$$

The first condition makes the last quantity negative for every $j \notin S$ and $\ell \in S$. It bounds the minimum in $\widehat{\Gamma}_j(S,h)$ from above and proves the deterministic certificate.

Let $\epsilon_{R,b,m}$ denote the right side of Proposition 15. On its probability-$1 - o(1)$ event, $\epsilon_{R,b,m} < \lambda_R$ and $\lambda_R + \epsilon_{R,b,m} < c_d$ give $\widehat{S}_{\mathrm{cand}}(h) = S$ for every bandwidth. At $a = \mathbf{1}$, the ridge gradient equals zero and

$$m\{g_{\ell j}(\mathbf{1}) - g^\infty_{\ell j}(\mathbf{1})\} = -\{\widehat{d}_\ell(h) - d_\ell(h)\} + \{\widehat{d}_j(h) - d_j(h)\}.$$

Its absolute value is at most $2\epsilon_{R,b,m}$. Under $\rho_m = \bar{\rho}/m$,

$$2m^2(1-\delta)H^{(1)}_m = (4C_H + 2\bar{\rho})(1-\delta).$$

The population inequality and $2\epsilon_{R,b,m} < \chi$ prove the empirical one-box inequality for every bandwidth. Proposition 17 gives the stated operation bound. This completes the proof. □

Proof of Proposition 30. The observed signal and nuisance coordinates are independent, and their kernels factorize. Conditional centering of e gives $\mathbb{E}(V_\ell \mid X, N) = X_\ell / v_X$ and $\mathbb{E}V_\ell^2 = 1 + \sigma_V^2$. No Gaussian assumption on the response errors is needed. Applying the Gaussian scaling calculation in Proposition 45 to each observed signal coordinate gives the population risk

$$\frac{1}{s}\sum_{\ell=1}^{s} f(a_\ell) + \frac{\rho}{2m}\sum_{j=1}^{m}(a_j - 1)^2.$$

It has no further dependence on the nuisance weights.

Put $t = a v_X / h^2$ and $v = (1 + 4t^2)^{1/2}$. As a function of t,

$$r(t) = \frac{2t}{1+v}, \qquad r'(t) = \frac{2}{v(1+v)} > 0, \qquad r''(t) = -\frac{8t(1+2v)}{v^3(1+v)^2} < 0.$$

It follows that

$$f'(a) < 0, \qquad f''(a) = \frac{2v_X}{h^4}\{r'(t)^2 - (1 - r(t))r''(t)\} > 0.$$

At the symmetric signal vector, the signal-only multiplier and exchange margin are

$$\lambda^\circ = -\frac{f'(a_S)}{s} - \frac{\rho}{m}(a_S - 1), \qquad \chi = -\frac{f'(a_S)}{s} - \frac{\rho}{m}(a_S - \delta) > 0.$$

If a nuisance weight satisfies $a_j > \delta$, signal weights sum to less than sa_S, giving some $a_\ell < a_S$. Transferring $\epsilon > 0$ from j to ℓ has directional derivative at most

$$\frac{f'(a_S)}{s} + \frac{\rho}{m}(a_S - \delta) < 0.$$

Such a point cannot minimize the risk. Every nuisance weight thus equals δ. Strict convexity of f and permutation symmetry of the signal coordinates then give the unique signal weights a_S.

The strict directional inequality gives positive lower-bound multipliers for all nuisance coordinates. The critical cone fixes these coordinates at zero. On the remaining signal tangent space, the Hessian has diagonal entries $s^{-1}f''(a_S) + \rho/m > 0$, which proves positive definiteness. At $a = \mathbf{1}$, the unpenalized derivative is $s^{-1}f'(1)$ for every signal coordinate and zero for every nuisance coordinate. Multiplication by $-m$ proves the screening-score formulas. This completes the proof. □

M A FULL SYMMETRIC METRIC

A full symmetric metric permits predictive information to depend on cross-coordinate distance terms within a prespecified measurement basis. Coordinate masking remains basis-dependent.

Let

$$\mathcal{M}_\delta = \{M \in \mathbb{R}^{m\times m} : M = M^\top,\ M \succeq \delta I_m,\ \mathrm{tr}(M) = m\}.$$

For a mask ξ, let $P_\xi = \mathrm{diag}(\xi)$ and replace (2) by

$$d^2_{M,\xi}(x_i, x_j) = \frac{1}{\pi}(x_i - x_j)^\top P_\xi M P_\xi (x_i - x_j). \tag{22}$$

The symmetric scaling and prediction loss use this distance. The prediction bandwidth is parameterized as $h = m\tau$ with $\tau \in [\tau_-, \tau_+]$. Under a trace constraint, a full metric can concentrate on a direction whose squared coordinate difference is of order m. The scale $h \asymp m$ gives a dimension-free kernel lower bound in this larger class. The trace and eigenvalue lower bound separate metric shape from bandwidth and prevent rank collapse.

The full-matrix gradient follows from the same score coefficient in Proposition 14.

Proposition 46. *Let C be defined as in Proposition 14 for the distance in* (22)*. Then*

$$\nabla_M \ell_n(M, h; \xi) = -\frac{1}{2\pi h^2}\sum_{i,j=1}^{n} C_{ij} P_\xi (x_i - x_j)(x_i - x_j)^\top P_\xi.$$

The gradient is symmetric.

Proposition 46 requires one scaling derivative solve per mask, as in the diagonal method. The difference is the $m \times m$ gradient accumulation.

Proof of Proposition 46. For a symmetric direction H,

$$\dot\eta_{ij}(H) = -\frac{1}{2\pi h^2}\,\mathrm{tr}\{H P_\xi (x_i - x_j)(x_i - x_j)^\top P_\xi\}.$$

The proof of Proposition 14 gives $\dot\ell_n = \sum_{i,j} C_{ij}\dot\eta_{ij}(H)$. Collecting the coefficient of H proves the formula. Every summand is symmetric. This completes the proof. □

Projected gradient descent on $\mathcal{M}_\delta$ has a simple spectral projection. If $S = S^\top$ has eigendecomposition $S = V \operatorname{diag}(s_1, \dots, s_m) V^\top$, project $(s_1, \dots, s_m)$ onto the simplex with lower bound δ and total mass m. Recombine the projected eigenvalues with V. This is the Euclidean projection because the Frobenius norm is unitarily invariant.

The parameter dimension changes from $m - 1$ to

$$q_m = \frac{m(m+1)}{2} - 1.$$

Let $E_1, \dots, E_{q_m}$ be a Frobenius-orthonormal basis of the symmetric trace-zero matrices, and write

$$\theta(M) = \{\operatorname{tr}(E_1 M), \dots, \operatorname{tr}(E_{q_m} M)\}^\top.$$

Let $\widehat{M}$ satisfy the full-matrix version of (7), with penalty $\rho \|M - I_m\|_F^2/(2m)$ and error η_R. Let $(M_{0,b}, h_{0,b})$ minimize the finite-block risk. The next theorem states the statistical cost of the larger metric without asserting an unverified increasing-dimensional population-smoother limit.

Theorem 20. *Suppose $\|X\|_\infty \le B$, $\pi \ge \pi_- > 0$, $b \ge 3$, every mask includes and omits at least one coordinate, $h/m \in [\tau_-, \tau_+]$, and m is fixed. Assume $(M_{0,b}, h_{0,b})$ is the unique finite-block minimizer, $M_{0,b} \succ \delta I_m$, its bandwidth gap is positive, and its tangent Hessian has smallest eigenvalue at least $\kappa > 0$. With probability at least $1 - \alpha$,*

$$\sup_{M \in \mathcal{M}_\delta,\, h \in \mathcal{H}} |\widehat{\mathcal{R}}_\rho(M, h) - \mathcal{R}_{b,\rho}(M, h)| \le C_5 \sqrt{\frac{q_m \log(C_6 R) + \log(|\mathcal{H}|/\alpha)}{R}}.$$

If b is fixed and $\eta_R = o_{\mathbb{P}}(R^{-1})$, then

$$\sqrt{R}\, \theta(\widehat{M} - M_{0,b}) \Longrightarrow N(0, H_M^{-1} \Sigma_M H_M^{-1}),$$

where H_M and Σ_M are the finite-block tangent Hessian and score covariance.

Theorem 20 shows that the unrestricted symmetric metric has q_m free parameters even for its finite-block target. A population-smoother result follows from the argument in Proposition 31 when $b \to \infty$ and m is fixed. No growing-m claim is made for the full metric.

Proof of Theorem 20. For $M \in \mathcal{M}_\delta$, $\lambda_{\max}(M) \le m - (m-1)\delta$ and

$$d_{M,\xi}^2(x_i, x_j) \le \frac{4B^2 m^2}{\pi_-}.$$

Because $h^2 \ge m^2 \tau_-^2$, the off-diagonal kernel entries have a positive lower bound independent of m. The compactness and implicit-function argument in Lemma 4 gives uniform bounds for fixed m.

The set $\mathcal{M}_\delta$ is contained in a Frobenius ball of radius m. Its ε-covering number satisfies

$$N(\varepsilon, \mathcal{M}_\delta, \|\cdot\|_F) \le (Cm/\varepsilon)^{q_m}.$$

Hoeffding's inequality on this net, a union bound over $\mathcal{H}$, and the uniform Lipschitz bound give the first assertion after setting $\varepsilon = R^{-1}$.

Uniform convergence, uniqueness, and the bandwidth gap give $\widehat{M} \to M_{0,b}$ and $\mathbb{P}(\widehat{h} = h_{0,b}) \to 1$. Let $\widetilde{M}$ be the exact empirical minimizer at $h_{0,b}$. The eigenvalue constraint is inactive with probability tending to one. In the tangent coordinates, its empirical score equation is

$$0 = \nabla_\theta \widehat{\mathcal{R}}_\rho(M_{0,b}, h_{0,b}) + \overline{H}_M \theta(\widetilde{M} - M_{0,b}).$$

Here $\overline{H}_M = \int_0^1 \widehat{H}_M\{M_{0,b} + t(\widetilde{M} - M_{0,b})\}\, dt$ is the integrated tangent Hessian. The multivariate central limit theorem gives a centered normal limit with covariance Σ_M for the first term after multiplication by $\sqrt{R}$. Uniform convergence of the empirical Hessian gives $\overline{H}_M \to H_M$. Local strong convexity and the objective gap give $\|\widehat{M} - \widetilde{M}\|_F \le C\eta_R^{1/2} = o_{\mathbb{P}}(R^{-1/2})$. Solving the score equation and transferring the limit from $\widetilde{M}$ to $\widehat{M}$ proves the result. This completes the proof. □

Remark 3. A symmetric perturbation orthogonal in the Frobenius inner product to $P_\xi(x-x')(x-x')^\top P_\xi$ for every admissible mask and $P \otimes P$-almost every pair leaves the population affinities unchanged. The unpenalized predictive risk cannot identify that perturbation. Shrinkage toward I_m can select a representative but does not identify the unpenalized target. Full-matrix identification must be checked after coordinate masking.

Remark 3 is the matrix analogue of Remark 2. Parameterization by $M = (B+B^\top)/2$ also contains an unidentifiable skew-symmetric part of B. Direct optimization over a symmetric M avoids this redundant parameter.

N Continuous prediction-bandwidth estimation

The finite prediction-bandwidth set is the default procedure. A continuous normalized bandwidth is useful when the sample size supports one additional parameter.

Let $h = \sqrt{m}\,\tau$ with $\tau \in [\tau_-, \tau_+]$ and use the derivative with respect to $\log h$ in Proposition 14. Near an interior finite-block minimizer $(a_{0,b}, \tau_{0,b})$, use the local coordinates

$$\vartheta = \begin{pmatrix} Q^\top(a - a_{0,b}) \\ \log\tau - \log\tau_{0,b} \end{pmatrix}.$$

Let H_ϑ be the Hessian of the expected finite-block criterion $\mathcal{R}_{b,\rho}(a, \sqrt{m}\,\tau)$ at $\vartheta = 0$ in these coordinates, and let Σ_ϑ be the covariance of the corresponding block score.

Theorem 21. *Suppose m and b are fixed, Condition 1 holds, $(a_{0,b}, \tau_{0,b})$ is the unique interior finite-block minimizer, and $\lambda_{\min}(H_\vartheta) \geq \kappa > 0$. If the computed criterion is within $o_{\mathbb{P}}(R^{-1})$ of its infimum, then*

$$\sqrt{R}\,\widehat{\vartheta} \Longrightarrow N(0, H_\vartheta^{-1}\Sigma_\vartheta H_\vartheta^{-1}).$$

For fixed m, $\|\widehat{\vartheta}\|_2 = O_{\mathbb{P}}(R^{-1/2})$.

Theorem 21 treats prediction-bandwidth uncertainty as part of finite-block metric estimation. It does not concern the nearest-neighbor order in the representation graph.

Proof of Theorem 21. The interval for $\log\tau$ is compact. Off the diagonal, $\partial^r_{\log\tau} \log K_{ij} = (-2)^r \log K_{ij}$ for $r = 1, 2, 3$. These derivatives and their mixed metric derivatives are uniformly bounded. The differentiated scaling equations in the proof of Lemma 4 then bound the first three derivatives of the scaling and loss with respect to $(a, \log\tau)$. Concentration over a covering of the parameter set gives uniform convergence of the criterion and its first two derivatives. Uniqueness gives consistency for $(a_{0,b}, \tau_{0,b})$.

Both constraints are inactive near the population minimizer. Let $\widetilde{\vartheta}$ be the exact empirical minimizer. Taylor expansion of its empirical score gives

$$0 = \nabla_\vartheta \widehat{\mathcal{R}}_\rho(0) + \overline{H}_\vartheta \widetilde{\vartheta}.$$

Here $\overline{H}_\vartheta = \int_0^1 \widehat{H}_\vartheta(t\widetilde{\vartheta})\,dt$. The block central limit theorem gives

$$\sqrt{R}\,\nabla_\vartheta \widehat{\mathcal{R}}_\rho(0) \Longrightarrow N(0, \Sigma_\vartheta),$$

and the uniform derivative result gives $\overline{H}_\vartheta \to H_\vartheta$. The population lower bound makes the limit invertible. The objective gap and local strong convexity give $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_2 = o_{\mathbb{P}}(R^{-1/2})$. Solving the score equation and applying Slutsky's theorem proves both assertions. This completes the proof. □

O Observation-level cross-fitting

The loss in (5) excludes a held-out feature from the distance and deletes the diagonal of the graph. An observation-level split gives a stronger separation: the evaluation observation does not enter either the training kernel or its symmetric scaling. This version is useful when prediction risk itself is the reported inferential target.

For one fold, let I_{tr} and I_{ev} denote the training and evaluation indices. Construct the masked training kernel on I_{tr} and let u_j, $j \in I_{\mathrm{tr}}$, be its symmetric scaling factors. For $i \in I_{\mathrm{ev}}$ and $j \in I_{\mathrm{tr}}$, define

$$p_{ij}(a,h;\xi) = \frac{u_j K_{ij}(a,h;\xi)}{\sum_{q\in I_{\mathrm{tr}}} u_q K_{iq}(a,h;\xi)}.$$

The cross-fitted prediction and loss are

$$\widehat{x}_{i,\xi^c} = \sum_{j\in I_{\mathrm{tr}}} p_{ij} x_{j,\xi^c}, \qquad \ell_{\mathrm{cf}}(a,h;\xi) = \frac{1}{|I_{\mathrm{ev}}||\xi^c|} \sum_{i\in I_{\mathrm{ev}}} \|x_{i,\xi^c} - \widehat{x}_{i,\xi^c}\|_2^2.$$

The factors u_j preserve the density correction learned from the training graph. The normalization over j makes each evaluation prediction a convex combination of training targets.

The derivative requires the change of the training scaling and the change of the evaluation row. Let $z_j = \log u_j$. For a scalar perturbation, compute $\dot{z}$ from (20) on the training graph and put

$$s_{ij} = \dot{\eta}_{ij} + \dot{z}_j, \qquad \overline{s}_i = \sum_{j\in I_{\mathrm{tr}}} p_{ij} s_{ij}.$$

Proposition 47. *For every $i \in I_{\mathrm{ev}}$ and $j \in I_{\mathrm{tr}}$,*

$$\dot{p}_{ij} = p_{ij}(s_{ij} - \overline{s}_i).$$

If $e_i = x_{i,\xi^c} - \widehat{x}_{i,\xi^c}$, then

$$\dot{\ell}_{\mathrm{cf}} = -\frac{2}{|I_{\mathrm{ev}}||\xi^c|} \sum_{i\in I_{\mathrm{ev}}} e_i^\top \sum_{j\in I_{\mathrm{tr}}} p_{ij}(s_{ij} - \overline{s}_i) x_{j,\xi^c}.$$

Proposition 47 permits the same projected metric update used by the block estimator. The only additional calculation is the normalized kernel row for each evaluation observation. No eigendecomposition is needed during cross-fitted metric estimation.

Proof of Proposition 47. Write

$$p_{ij} = \frac{\exp(z_j + \eta_{ij})}{\sum_{q\in I_{\mathrm{tr}}} \exp(z_q + \eta_{iq})}.$$

Differentiating the log probability gives

$$\frac{\dot{p}_{ij}}{p_{ij}} = \dot{z}_j + \dot{\eta}_{ij} - \sum_{q\in I_{\mathrm{tr}}} p_{iq}(\dot{z}_q + \dot{\eta}_{iq}) = s_{ij} - \overline{s}_i.$$

This proves the first formula. The prediction differential is

$$\dot{\widehat{x}}_{i,\xi^c} = \sum_{j\in I_{\mathrm{tr}}} \dot{p}_{ij} x_{j,\xi^c}.$$

Because $\dot{e}_i = -\dot{\widehat{x}}_{i,\xi^c}$, differentiating the sum of squared residuals proves the second formula. This completes the proof. □

Cross-fitting also gives a direct validation bound because the fitted training graph is independent of the evaluation observations. Let $\mathcal{C}$ be a finite collection of candidates fitted on I_{tr}. A candidate may specify a bandwidth, a ridge value, and an optimization start. Let $\widehat{R}_{\mathrm{ev}}(c)$ be its average evaluation loss and let $R_{\mathrm{ev}}(c)$ be the conditional expectation of that loss given the training data and masks.

Theorem 22. *Suppose $\|X\|_\infty \le B$ and $|I_{\mathrm{ev}}| = n_{\mathrm{ev}}$. Conditionally on the training data and the prespecified masks, with probability at least $1-\alpha$,*

$$\max_{c\in\mathcal{C}} |\widehat{R}_{\mathrm{ev}}(c) - R_{\mathrm{ev}}(c)| \le 4B^2 \sqrt{\frac{\log(2|\mathcal{C}|/\alpha)}{2n_{\mathrm{ev}}}}.$$

If $\widehat{c}$ minimizes $\widehat{R}_{\mathrm{ev}}$ and c_0 minimizes R_{ev}, then

$$R_{\mathrm{ev}}(\widehat{c}) - R_{\mathrm{ev}}(c_0) \le 8B^2 \sqrt{\frac{\log(2|\mathcal{C}|/\alpha)}{2n_{\mathrm{ev}}}}.$$

Theorem 22 gives a finite-sample guarantee for selecting among prespecified candidates. Its sample size is the number of independent evaluation observations, not the number of pairwise affinities in their kernel rows. The candidate collection must be fixed using the training data and must not be enlarged after examining the evaluation losses.

Proof of Theorem 22. Predictions are convex combinations, and both predictions and observations lie in $[-B, B]$, giving loss at most $4B^2$. Conditional on training data and masks, the n_{ev} losses are independent. Hoeffding's inequality and a union bound over $\mathcal{C}$ give the first assertion.

Let the right side of the first inequality be t. On the event in that inequality,

$$R_{\mathrm{ev}}(\widehat{c}) \le \widehat{R}_{\mathrm{ev}}(\widehat{c}) + t \le \widehat{R}_{\mathrm{ev}}(c_0) + t \le R_{\mathrm{ev}}(c_0) + 2t.$$

This proves the second assertion. This completes the proof. □

Independent validation also selects candidates inside the geometric regime. The candidate set must shrink toward one geometric target; a fixed stable-bandwidth candidate and a shrinking-bandwidth candidate estimate different population objects.

Let $\mathcal{C}_N^{\mathrm{geo}}$ index a finite grid of bandwidth and ridge pairs fixed before the training observations are used. The metric for candidate c is fitted on those observations at (h_c, ρ_c), with certified optimization error η_c. The fitted metric and this error may depend on the training sample. Put

$$\Delta_c = \sqrt{\frac{m \log\{mR|\mathcal{C}_N^{\mathrm{geo}}|/h_c\}}{R}} + \sqrt{\frac{m \log\{mRJ|\mathcal{C}_N^{\mathrm{geo}}|/h_c\}}{RJ}} + q_b(h_c) + \eta_c$$

and define

$$e_N = \max_{c \in \mathcal{C}_N^{\mathrm{geo}}} \Big[h_c + \Big| \frac{\rho_c}{h_c^4} - \rho_\star \Big|^{1/2} + \Big\{ \frac{\Delta_c}{h_c^4} \Big\}^{1/2} \Big].$$

Under the product-mask model, replace $q_b(h_c)$ by $q_b^{\mathrm{prod}}(h_c)$ in Δ_c, and replace $a_\star$ and $\beta_\star$ below by $a_\star^{\mathrm{mask}}$ and

$$\beta_\star^{\mathrm{mask}} = \min_{j : a_{\star j}^{\mathrm{mask}} > \delta} (a_{\star j}^{\mathrm{mask}} - \delta).$$

Corollary 24. *Suppose the assumptions of Theorem 12, or the assumptions of Corollary 14 with the stated replacement, hold uniformly over $\mathcal{C}_N^{\mathrm{geo}}$,*

$$\max_{c \in \mathcal{C}_N^{\mathrm{geo}}} h_c \to 0, \qquad e_N \to 0 \quad \textit{in probability},$$

and the evaluation observations are independent of the fitted candidates. If $\widehat{c}$ minimizes the evaluation loss, then

$$\|\widehat{a}_{\widehat{c}} - a_\star\|_2 = O_{\mathbb{P}}(e_N),$$

and, conditionally on the training observations, with probability at least $1 - \alpha$,

$$R_{\mathrm{ev}}(\widehat{c}) - \min_{c \in \mathcal{C}_N^{\mathrm{geo}}} R_{\mathrm{ev}}(c) \le 8B^2 \sqrt{\frac{\log\{2|\mathcal{C}_N^{\mathrm{geo}}|/\alpha\}}{2n_{\mathrm{ev}}}}.$$

The tensor and geodesic errors are $O_{\mathbb{P}}(e_N)$. If

$$e_N = o_{\mathbb{P}}(t_N), \qquad t_N = o(\beta_\star), \qquad \beta_\star = \min_{j : a_{\star j} > \delta} (a_{\star j} - \delta),$$

then the selected representation support equals $\{j : a_{\star j} > \delta\}$ with probability tending to one.

Corollary 24 permits validation to select the bandwidth, ridge, and optimization start without changing the limiting geometric target. For the stable mode, Theorem 22 directly gives the stated oracle-risk bound for any prespecified finite collection of bandwidth and ridge candidates. Corollary 11 gives the same geometry conclusion under the product-mask model.

Proof of Corollary 24. The covering and concentration arguments in Theorem 12 apply to all candidates after a union bound. In the product case, Proposition 33 gives the stated replacement. The two logarithms in Δ_c include this union bound. Hence

$$\max_{c \in \mathcal{C}_N^{\mathrm{geo}}} \|\widehat{a}_c - a_\star\|_2 = O_{\mathbb{P}}(e_N).$$

The assertion holds at the random index $\widehat{c}$ because it holds simultaneously over the candidate set. The tensor and geodesic calculations in Proposition 24 are Lipschitz in the metric and give their stated rates.

Theorem 22, applied conditionally on the training observations, proves the oracle-risk inequality. On the simultaneous metric event,

$$\|\widehat{a}_{\widehat{c}} - a_\star\|_\infty \le \|\widehat{a}_{\widehat{c}} - a_\star\|_2 = O_{\mathbb{P}}(e_N).$$

The two threshold orders and the argument in Corollary 19 prove support recovery. This completes the proof. □

A single independent evaluation split gives the finite-sample bound above. The feature masks are generated independently in the training and evaluation collections. The selected candidate may then be refitted on all observations because its full-data rate holds simultaneously over the prespecified candidate set. This validation result is separate from observation-level metric inference in Theorem 16. The simultaneous spectral event in Lemma 13 permits the refitted full-data metric and graph to use the same observations.

P Out-of-sample representation

A new observation is represented using the fitted metric and the fixed nearest-neighbor training graph. The extension below agrees with the training eigenvector equation under its diagonal-deletion convention.

Let r_j be the training k_nth-neighbor radii, let $\widehat{L} y_r = \nu_r y_r$, and let x be a new standardized observation. Denote its k_nth-neighbor distance among the training observations by r_x and define

$$a_j(x) = \varphi\left\{ \frac{d_{\widehat{a}^{\mathrm{rep}}}(x, x_j)}{r_x \vee r_j} \right\}, \qquad d(x) = \sum_{j=1}^{n} a_j(x).$$

If $r_x = r_j = 0$, the same positive-radius limit gives $a_j(x) = \varphi(0)$ for coincident observations and $a_j(x) = 0$ for distinct observations. With the fitted training graph held fixed, the eigen-equation for an added graph row gives the extension

$$\widehat{y}_r(x) = \frac{\sum_{j=1}^{n} a_j(x) y_{jr}}{d(x) - \nu_r}. \tag{23}$$

When $x = x_i$ is a training observation, compute r_x after excluding x_i and set $a_i(x_i) = 0$. Under this diagonal-deletion convention and the same neighbor radii, (23) equals the corresponding eigenvector coordinate.

Proposition 48. *Suppose $|d(x) - \nu_r| \ge c_d k_n$. Under perturbations δa, δy_r, and $\delta \nu_r$, the first-order change in* (23) *is bounded by*

$$\frac{\|\delta a\|_2 \|y_r\|_2 + \|a(x)\|_2 \|\delta y_r\|_2}{c_d k_n} + \frac{|\delta d - \delta \nu_r|\, |a(x)^\top y_r|}{c_d^2 k_n^2},$$

where $\delta d = \mathbf{1}^\top \delta a$.

Proposition 48 separates instability of a new neighbor row from instability of the selected eigenspace. The denominator is observable and is checked before a coordinate is reported.

Proof of Proposition 48. Differentiate

$$\{d(x) - \nu_r\}^{-1} a(x)^\top y_r$$

with respect to its four arguments and apply Cauchy–Schwarz to the two vector terms. The assumed denominator bound proves the result. This completes the proof. □

Q SELECTION OF THE REPRESENTATION DIMENSION

The representation dimension determines the boundary of the selected spectral subspace. A known scientific dimension should be fixed before fitting. When this information is unavailable, a separated population eigengap gives a label-free selection rule.

For fixed $p_{\max} \geq 2$, let $\mathcal{P} = \{1, \ldots, p_{\max}\}$ and define

$$\Delta_p^0 = \lambda_{p+2}(a_0^{\mathrm{rep}}) - \lambda_{p+1}(a_0^{\mathrm{rep}}), \qquad p \in \mathcal{P}.$$

Let p_0 maximize Δ_p^0. For the scaled fitted graph Laplacian $L_{n,k}(\widehat{a}^{\mathrm{rep}})$, define

$$\widehat{\Delta}_p = \widehat{\lambda}_{p+2} - \widehat{\lambda}_{p+1}, \qquad \widehat{p} = \arg\max_{p \in \mathcal{P}} \widehat{\Delta}_p.$$

Ties are resolved by taking the smaller dimension.

Proposition 49. *Suppose*

$$\eta = \Delta_{p_0}^0 - \max_{p \in \mathcal{P},\, p \neq p_0} \Delta_p^0 > 0.$$

If

$$\max_{1 \leq j \leq p_{\max}+2} |\widehat{\lambda}_j - \lambda_j(a_0^{\mathrm{rep}})| < \eta/4,$$

then $\widehat{p} = p_0$. *Under Theorem 19,*

$$(k_n/n)^{1/d} + \|\widehat{a}^{\mathrm{rep}} - a_0^{\mathrm{rep}}\|_2 = o_{\mathbb{P}}(\eta) \quad \Longrightarrow \quad \mathbb{P}(\widehat{p} = p_0) \to 1.$$

Proposition 49 separates dimension selection from the existence of a stable dimension. The rule is consistent only when one candidate population eigengap exceeds the others by a nonzero margin.

Proof of Proposition 49. Let

$$e_n = \max_{1 \leq j \leq p_{\max}+2} |\widehat{\lambda}_j - \lambda_j(a_0^{\mathrm{rep}})|.$$

For every $p \in \mathcal{P}$,

$$|\widehat{\Delta}_p - \Delta_p^0| \leq 2e_n.$$

Hence, for $p \neq p_0$,

$$\widehat{\Delta}_{p_0} - \widehat{\Delta}_p \geq \eta - 4e_n > 0.$$

For the stochastic assertion, apply Lemma 13 with $p = p_{\max}$; its larger spectral cluster needs no internal gaps. Proposition 37 controls metric changes at these indices by min–max. The stated order gives $e_n = o_{\mathbb{P}}(\eta)$. This completes the proof. □

The same argument permits $p_{\max}$ to increase when the spectral bound and gap margin hold uniformly over the enlarged candidate set.

R IMPLEMENTATION DETAILS

Symmetric scaling determines both the predictor and its derivative. Numerical error in this scaling is included in the objective bounds before a fitted metric is used to construct the graph.

For a log kernel η_{ij} with $\eta_{ii} = -\infty$, the scaling solves the convex problem

$$\min_{z \in \mathbb{R}^n} \Big[\frac{1}{2} \sum_{i,j} \exp(z_i + \eta_{ij} + z_j) - \sum_i z_i \Big].$$

Its gradient is $W\mathbf{1} - \mathbf{1}$, and its Hessian is $\mathrm{diag}(W\mathbf{1}) + W$. At the solution, the Hessian is $I_n + W$. Damped log-domain fixed-point iterations provide an initial solution. After Newton refinement, it satisfies the residual criterion

$$\|W\mathbf{1} - \mathbf{1}\|_\infty \leq 10^{-8}.$$

If the initial convex objective cannot resolve further decreases, an initial residual of 10^{-6} is sufficient for subsequent refinement. Newton refinement uses a decrease in the maximum row residual and targets 5×10^{-14}. A residual above 10^{-11} is rejected. The remaining residual is propagated through the root enclosure described after Proposition 44; it is not treated as zero. The diagonal remains exactly zero throughout.

Two quantities describe prediction concentration and numerical conditioning. The effective neighborhood size of row i is

$$\nu_i(W) = \Big(\sum_{j=1}^{n} W_{ij}^2 \Big)^{-1}.$$

It lies between one and $n-1$. A value near one means that a single observation supplies most of the prediction weight. A value near $n-1$ means that the row is nearly uniform. This quantity describes the distribution of prediction weights and is not itself an estimate of prediction accuracy.

The conditioning quantity is $s(W) = \lambda_{\min}(I_n + W)$. Proposition 12 gives $s(W) > 0$, and (20) gives

$$\|\dot{z}\|_2 \le s(W)^{-1} \|(W \circ \dot{\eta})\mathbf{1}\|_2.$$

A small value of $s(W)$ permits a larger derivative bound and requires a more accurate scaling calculation. The global objective certificate uses the actual row residual and a lower bound on the computed scaling Hessian. Finite-difference and high-precision derivative checks are independent of evaluation labels.

Projection onto the metric polytope is obtained by finding t such that

$$\sum_{j=1}^{m} \max(\delta, v_j - t) = m.$$

The projected coordinate is $\max(\delta, v_j - t)$. Sorting the breakpoints gives the projection in $O(m \log m)$ operations. A parameter cell adds coordinatewise upper bounds, and the same scalar projection calculation applies after truncation at both ends.

Sequential quadratic programming supplies the candidates in Section I. The initial values and iteration budgets are fixed before each comparison. With $\widehat{\mathcal{R}}$ denoting the penalized criterion, the descent result in Proposition 19 also covers projected-gradient steps satisfying the Armijo test

$$\widehat{\mathcal{R}}(a^+, h) \le \widehat{\mathcal{R}}(a, h) + 10^{-4} \nabla \widehat{\mathcal{R}}(a, h)^\top (a^+ - a).$$

A stationary point does not provide a global objective gap. Proposition 20 bounds each block loss over a parameter cell through extrema of linear log-kernel functions. These extrema are computed exactly by allocation under a box and a trace constraint. The lower bound includes the smallest ridge penalty on the cell. Longest-edge subdivision refines the cell with the smallest lower bound, while any feasible candidate can update the upper bound. The calculation returns the achieved difference between the global upper and lower bounds even when a fixed budget is exhausted.

Restriction to a proposed support is valid only after the exchange inequalities in Proposition 16 are verified. Gradient-ranked initial values alone do not justify this restriction. The higher-dimensional numerical fits search the full metric polytope. The binary inference fits instead verify global tangent curvature by Proposition 44 and obtain the objective gap from the scalar score residual. After fitting, (8) is applied before neighbor search.

Feature centering and scaling constants are fixed by the population design or an independent pilot sample. The same constants are used for metric fitting and representation. Constant pilot coordinates are removed before blocks are formed. The final graph is symmetric with a zero diagonal and uses the triangular nearest-neighbor affinity. Its eigenvectors are centered and orthonormal, as required by (9). If the graph is disconnected, the solver excludes only the global constant vector and includes any other zero-eigenvalue directions required by that variational problem.

R.1 Exact distance computation

Feature selection reduces the number of coordinates needed to construct the representation graph. A separate storage calculation is needed because a sparse graph can still require a dense intermediate distance matrix. We compute the same graph without storing that matrix.

The weights in (8) equal zero outside $\widehat{S}$. Restricting each observation to $\widehat{S}$ leaves every weighted distance unchanged. We compute distances for at most 128 observations at a time. The first pass finds their k_nth-neighbor radii after excluding the diagonal, and stores only these radii. The second pass recomputes the distances and stores the nonzero affinities. Each entry uses the same distance and the same pair of radii as the full-matrix calculation. Thus the affinity, Laplacian, and minimizers in (9) are unchanged.

For positive neighbor radii, the number of nonzero entries is bounded directly. Continuity and compact support give $\varphi(1) = 0$. A positive entry requires the distance to be strictly less than at least one of its two radii. Hence

$$\sum_{i \neq j} \mathbf{1}\{\widehat{A}_{ij} > 0\} \leq 2 \sum_{i} \sum_{j \neq i} \mathbf{1}\{d_{\widehat{a}^{\mathrm{rep}}}(x_i, x_j) < \widehat{r}_i\} \leq 2n(k_n - 1).$$

This bound does not require distinct distances. Each strict neighbor set contains at most $k_n - 1$ observations. The independent observations in the sphere model of Theorem 8 have continuous coordinate distributions. Two observations agree on a nonempty coordinate subset with probability zero. A finite union over the possible subsets shows that the selected radii are positive almost surely, although selection uses the observations. For arbitrary discrete data with zero radii, the storage bound depends on the actual number of nonzero entries.

The two distance passes require $O(n^2|\widehat{S}|)$ operations. Their storage, including the selected covariates and the sparse affinity, is $O\{n(|\widehat{S}| + k_n + 1)\}$. This is a storage improvement; exact all-pairs search remains quadratic in n. The original measurements and the metric-fitting calculations are separate from this graph workspace. For graphs with more than 800 vertices, the eigensolver uses sparse Laplacian products and a rank-one shift of the constant direction. Its starting vector and convergence tolerance are the same as in the full-matrix implementation.

New observations are processed in blocks of the same size. For each observation, the algorithm obtains its exact neighbor radius, degree, and weighted sums in (23) before discarding its distance row. Training radii and eigenvectors stay fixed. The denominator check is unchanged. Thus processing the queries separately does not change their coordinates or the fitted prediction rule. For n_{new} queries, distance and coordinate calculations require $O\{n_{\mathrm{new}} n(|\widehat{S}| + p)\}$ operations, while the temporary distance storage remains $O(n)$ for the fixed block size. Section I.10 measures the complete calculation, including metric fitting and prediction.

S Sample-size sequences and population examples

The asymptotic conditions distinguish increasing graph resolution, accurate metric estimation, and removal of mask-integration error. Explicit sequences show that these conditions are compatible. The finite experiments in Section I use the separately specified sample sizes and tuning values.

S.1 Graph resolution and observation-level inference

We give sample-size sequences that satisfy the spectral and observation-level inference conditions. For a fixed intrinsic dimension $d \geq 2$ and constant $C_k > 0$, consider

$$k_n = \left\lceil C_k n^{4/(d+4)} (\log n)^{d/(d+4)} \log\log n \right\rceil.$$

Then $k_n/n \to 0$ and $k_n^{1+4/d}/\{n^{4/d} \log n\}$ is of order $(\log\log n)^{1+4/d}$, which tends to infinity. This verifies the graph conditions in Proposition 39. In dimension two, the same choice can be written as

$$k_n = \left\lceil C_k n^{2/3} (\log n)^{1/3} \log\log n \right\rceil \wedge (n-1),$$

where the truncation is inactive for sufficiently large n. The triangular kernel may be multiplied by its integral-normalizing constant without changing the representation eigenvectors.

For observation-level inference, take $b_R = \lceil R^{3/2} \rceil$, $J_R = \lceil b_R \log b_R \rceil$, and $\eta_R = \{Rb_R \log(Rb_R)\}^{-1}$. Then $R(\log b_R)^2/b_R = O(R^{-1/2} \log^2 R) \to 0$, $b_R/J_R \to 0$, and $Rb_R\eta_R \to 0$. Thus the three sample-size conditions in Theorem 16 hold. An exact mask average, including a deterministic scientific view, needs no condition on J_R. Corollary 16 also permits $b_R = \lceil R \log R \rceil$ with exact mask averaging and the same optimization order. Then $R/b_R \to 0$, whereas $R(\log b_R)^2/b_R \to \infty$. This sequence satisfies the weaker condition but not the earlier logarithmic condition.

S.2 Stable prediction and support selection

A finite normalized-bandwidth family keeps the prediction kernel bounded away from zero. One admissible family is

$$\tau \in \{1/4, 1/2, 1, 2\}, \qquad \rho \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}, \qquad h = \sqrt{m}\,\tau.$$

with optimization tolerance $\eta_c = R^{-2}$. This is a parameter example for the stable-bandwidth theory. For fixed m and population identification and support-separation bounds that are uniform over the candidates, define

$$\overline{\omega}_{R,J,b} = \left\{ \frac{\log(2m|\mathcal{H}|)}{R} \right\}^{1/4} + \left\{ \frac{\log(2m|\mathcal{H}|)}{RJ} \right\}^{1/4} + \left\{ \frac{m \log(mb)}{b} \right\}^{1/4} + R^{-1}.$$

and use

$$t_N^{\mathrm{stab}} = \min\{(1-\delta)/2, \overline{\omega}_{R,J,b}^{1/2}\}.$$

If $R, b \to \infty$, then $\overline{\omega}_{R,J,b} \to 0$ for every $J \geq 1$. The parameter-error bound in Corollary 13 is $O_{\mathbb{P}}(\overline{\omega}_{R,J,b})$. Hence its ratio to the threshold tends to zero, while the threshold eventually lies below every fixed positive beta-min value. Corollary 19 then gives support consistency.

S.3 Decreasing prediction bandwidths

The local-geometry results require a different sequence. Let m be fixed, let $N = Rb$, and let R and b grow as positive powers of N. Let J be at most polynomial in N. Put $d_{\mathrm{fit}} = d_{\max}$ for product-view masks and $d_{\mathrm{fit}} = d$ when every observed coordinate lies on the signal manifold. With $\ell_N = \log\log(N + 16)$, define

$$h_{\min,N} = \ell_N \max\left[\left\{ \frac{\log R}{R} \right\}^{1/8}, \left\{ \frac{\log(RJ)}{RJ} \right\}^{1/8}, \left\{ \frac{\log b}{b} \right\}^{1/(5d_{\mathrm{fit}}+8)} \right].$$

The finite set $\mathcal{C}_N^{\mathrm{geo}}$ consists of the bandwidth and ridge pairs

$$h_c \in h_{\min,N}\{1, 3/2, 2\}, \qquad \rho_c/h_c^4 \in \{0, (2\ell_N)^{-1}, \ell_N^{-1}\}, \qquad \eta_c = h_c^4/\ell_N^2.$$

An upper bound $h_{\max} = 0.25$ is eventually satisfied by every candidate because $h_{\min,N} \to 0$. The definition gives $Rh_c^8/\log R \geq \ell_N^8$, $RJh_c^8/\log(RJ) \geq \ell_N^8$, and $bh_c^{5d_{\mathrm{fit}}+8}/\log b \geq \ell_N^{5d_{\mathrm{fit}}+8}$. These lower bounds tend to infinity. Also, $\rho_c/h_c^4 \to 0$ and $\eta_c/h_c^4 = \ell_N^{-2} \to 0$. Thus the candidates satisfy the primitive orders following Theorem 12 and Corollary 14, with limiting ridge value zero.

To make the support threshold explicit, define

$$\overline{q}_b(h) = \sqrt{\frac{\log b}{bh^{5d_{\mathrm{fit}}}}} + \frac{1}{bh^{3d_{\mathrm{fit}}}},$$

$$\overline{\Delta}_c = \sqrt{\frac{m \log\{R|\mathcal{C}_N^{\mathrm{geo}}|/h_c\}}{R}} + \sqrt{\frac{m \log\{RJ|\mathcal{C}_N^{\mathrm{geo}}|/h_c\}}{RJ}} + \overline{q}_b(h_c) + \eta_c$$

and

$$\overline{e}_N = \max_c \left[h_c + \left(\frac{\rho_c}{h_c^4} \right)^{1/2} + \left\{ \frac{\overline{\Delta}_c}{h_c^4} \right\}^{1/2} \right].$$

Then the preceding orders imply $\overline{\Delta}_c / h_c^4 \to 0$ uniformly over the finite candidate set and $\overline{e}_N \to 0$. The threshold

$$t_N^{\mathrm{geo}} = \min\{(1-\delta)/2, \overline{e}_N^{1/2}\}$$

separates an $O_{\mathbb{P}}(\overline{e}_N)$ metric error from a fixed positive support margin. This conclusion uses the population identification conditions in the cited results; a bandwidth sequence alone does not establish identification.

S.4 Observed-view examples

Smooth observed views clarify the embedding premise in the local-geometry theorem. Let the latent angles be independent, with

$$U_i, V_i \sim \mathrm{Unif}(0, 2\pi)$$

and consider the eight-coordinate map

$$\cos U_i,\ \sin U_i,\ \cos V_i,\ \sin V_i,\ \cos(U_i + V_i),\ \sin(U_i + V_i),\ \cos(U_i - V_i),\ \sin(U_i - V_i).$$

The first four coordinates already identify a point of the two-dimensional torus. Their derivative has rank two at every point. The remaining four coordinates preserve smoothness and injectivity. Every positive diagonal metric thus gives an embedding of a compact manifold without boundary.

A paired scientific view can be obtained by replacing each component ϕ_j of $(\cos U, \sin U, \cos V, \sin V)$ by $(\phi_j, \phi_j + 0.1\phi_j^3)$. Each coordinate transformation has derivative at least one. Observing either member of every pair still determines the four original coordinates and hence gives an embedding. Independent nuisance coordinates may be added to the observed view under Theorem 10. That theorem compares the functions

$$h^{-4}\{\mathcal{R}_{\infty,0}(a,h) - \mathcal{V}\} \quad \text{and} \quad \mathcal{J}(a)$$

with $\mathcal{J}_{\mathrm{mask}}$ replacing $\mathcal{J}$ for the paired-view construction. The comparison is uniform over the metric class and concerns the population conditional means.

A bounded surface map can instead be defined from independent uniform U, V on $[-1, 1]$ by

$$U_i, \quad V_i, \quad U_i V_i, \quad U_i^2 - 1/3, \quad V_i^2 - 1/3, \quad \sin(\pi U_i)\cos(\pi V_i).$$

Its first two coordinates give an injective map of a square, but the square has a boundary. Bounded-prediction results apply when their identification conditions hold; the closed-manifold theorem does not apply to this surface.

Heterogeneous product views need not have identical marginal distributions. Let B_ℓ have a beta distribution with parameters $(2, 5)$ for odd ℓ and $(5, 2)$ for even ℓ, independently across ℓ, and standardize $U_\ell = 2B_\ell - 1$ by its population mean and standard deviation. For independent centered bounded errors e_ℓ, put

$$V_\ell = \begin{cases} U_\ell + e_\ell, & \ell \text{ odd}, \\ U_\ell^2 - \mathbb{E}(U_\ell^2) + e_\ell, & \ell \text{ even}, \end{cases} \qquad N_j \sim \mathrm{Unif}(-\sqrt{3}, \sqrt{3}).$$

Both members of each signal pair are bounded, and independent nuisance kernels factor from their conditional means by Proposition 1. The curvature and multiplier inequalities in Proposition 26 remain separate population conditions. Product factorization alone does not imply them. Proposition 43 supplies a bounded model in which these identification and curvature calculations are explicit.

T RELATION TO NEARBY METHODS

Laplacian Eigenmaps estimate coordinates for a fixed affinity (Belkin & Niyogi, 2003). Adaptive windows alter neighborhoods (Xu et al., 2024), and Soft Laplacian Eigenmap uses a feature-weighted distance to regularize its graph (Fathinezhad et al., 2025). Smooth graph learning estimates a graph from the variation of observed signals (Kalofolias, 2016; Dong et al., 2016), while self-smoothing and self-diffusion transform an existing similarity matrix (Jiang et al., 2011; Wang & Tu, 2012). The label-free PALE criterion assesses a metric by predicting measurements excluded from the distance. This exclusion gives the conditional-mean target in Proposition 21; the fitted metric then determines the spectral graph.

Leave-one-out kernel-regression error is also used for supervised metric estimation (Weinberger & Tesauro, 2007). Split-brain autoencoders predict across feature groups to learn neural representations (Zhang et al., 2017), while Noise2Self estimates denoising risk by excluding coordinates with independent measurement noise (Batson & Royer, 2019). The shared prediction principle does not determine the same object: PALE estimates a regularized metric through a symmetric bistochastic predictor and quantifies its effect on spectral coordinates.

SLAPS learns an adjacency matrix using classification and feature-denoising losses (Fatemi et al., 2021). This makes it a direct graph-learning comparison; our numerical study uses only its self-supervised reconstruction loss. SUBLIME estimates graph structure by contrastive agreement (Liu et al., 2022). GraphMAE2 uses repeated masking and latent representation prediction on an observed graph (Hou et al., 2023); it does not estimate an affinity for observations given without edges. The label-free PALE criterion uses neither a message-passing predictor nor class labels.

Unsupervised multiple-kernel learning selects a kernel through localized reconstruction and agreement with a prescribed geometry (Zhuang et al., 2011). Laplacian Score ranks features by their smoothness on a fixed neighborhood graph (He et al., 2005). Gated Laplacian scores select low-frequency features on a learned graph (Lindenbaum et al., 2021), and multi-modal scores distinguish shared and modality-specific structure (Yang et al., 2023). Spectral self-supervised feature selection predicts pseudo-labels obtained from graph eigenvectors (Segal et al., 2024); our comparison applies its selected features to the common eigensolver. These criteria use reconstruction or spectral structure to assess features. PALE assesses the metric before computing representation eigenvectors, using the conditional mean of a specified measurement view.

Correlated gates use partial feature sets in self-supervised pretraining but fit their final gates to a labeled prediction target (Lee et al., 2022); they are excluded from the label-free comparison. Supervised metric methods use class labels or similar and dissimilar pairs (Bach & Jordan, 2006; Ying & Li, 2012). Graph attention networks train attention for a prediction task (Veličković et al., 2018), and bilevel graph learning jointly estimates a graph distribution and neural-network parameters (Franceschi et al., 2019). The label-free loss studied here uses cross-coordinate prediction, with the scope stated in Remark 2.

Wasserstein singular vectors estimate feature ground costs and sample distances for nonnegative histograms (Huizing et al., 2022), a different domain from our signed standardized coordinates. Independent coordinate selection for product manifolds removes redundant spectral coordinates before factorizing latent components (He et al., 2023). Augmentation invariant manifold learning uses copies that preserve a specified latent component (Wang, 2025); our random masks define prediction tasks without asserting this invariance.

The observed-space similarities in t-SNE and UMAP are specified from feature distances (van der Maaten & Hinton, 2008; McInnes et al., 2018). SNEkhorn matches an observed-space symmetric entropic affinity with a reduced-space bistochastic affinity (Van Assel et al., 2023). These methods estimate reduced coordinates for a prescribed observed-space cost. PALE estimates that cost before computing Laplacian eigenvectors.

Alternating diffusion composes diffusion operators from paired views to estimate common latent variation under a model with conditionally independent sensor-specific variation (Lederman & Talmon, 2018). Talmon & Wu (2019) analyze the corresponding common-manifold operator and describe a Nyström extension that uses one sensor at a new observation. These methods estimate shared geometry through view-specific kernels. We estimate coordinate weights from prediction risk and use them to construct the spectral graph. For single-cell data, weighted nearest neighbors learn cell-

specific weights for RNA and protein similarities (Hao et al., 2021), and totalVI estimates a joint probabilistic representation of their counts, with explicit models for protein background and batch effects (Gayoso et al., 2021). Our molecular analysis predicts RNA from a protein-only representation. Joint embeddings that use evaluation-cell RNA have access to the quantity being predicted. Comparisons under our observation scheme require single-view extensions learned only from the fitting data.

T.1 Further comparisons with graph and classification methods

Training-only information and alternative definitions of similarity lead to different statistical targets. The following comparisons place the posterior-moment estimator and the predictive metric among these approaches.

Training-only measurements have been studied as learning using privileged information (Vapnik & Vashist, 2009; Vapnik & Izmailov, 2015): generalized and privileged-feature distillation train a predictor from teacher scores, with the latter teacher using both measurement sets (Lopez-Paz et al., 2016; Yang et al., 2022). Metric imitation transfers privileged geometry to prediction features, privileged metric learning uses the auxiliary metric as a local decision threshold (Dai et al., 2015; Yang et al., 2018), and learning using unselected features assigns a training role to excluded features (Taylor et al., 2016). Confidence-based pseudo-labeling selects reliably predicted observations (Sohn et al., 2020), soft-label discriminant analysis estimates class structure from labels inferred using prediction features (Zhao et al., 2014; Wang et al., 2016), and moment-constrained discriminant analysis relates label-dependent and label-independent parameters (Loog, 2014). Unlike distillation, which fits the deployed predictor to teacher scores, posterior-moment PALE uses the teacher scores to screen observations and to estimate class moments of the prediction measurements, which determine the classifier and the affinity; only the subsequent spectral predictor fits teacher targets.

Other dimension-reduction and feature-selection methods define similarity differently: Isomap, locally linear embedding, Hessian eigenmaps, and local tangent space alignment preserve geometric properties of the observed features (Tenenbaum et al., 2000; Roweis & Saul, 2000; Donoho & Grimes, 2003; Zhang & Zha, 2004), and gated-Laplacian and spectral self-supervised selection use graph frequency or labels from an initial representation (Lindenbaum et al., 2021; Segal et al., 2024). Group discovery uses smoothness on sample and feature graphs (Lifshitz et al., 2026), variable priority uses local two-class tasks and sparse regression without a graph metric (Zhou et al., 2026). Multi-modal selection uses shared and view-specific Laplacian scores, SNEkhorn matches observed and reduced affinities (Yang et al., 2023; Van Assel et al., 2023), and augmentation invariant manifold learning establishes spectral convergence and classification guarantees from augmented observations (Wang, 2025). We estimate the coordinate weights from prediction without requiring augmented copies of each observation.

U FINITE-BANDWIDTH PREDICTION AND UNLABELLED MEASUREMENTS

U.1 A NONLINEAR RESPONSE AT A FINITE PREDICTION BANDWIDTH

The large-bandwidth condition in Theorem 8 permits identification over the full diagonal metric family. A direct calculation gives a different result at any fixed positive bandwidth when the metric is constant within prespecified sphere blocks. In this model, the response has zero covariance with every observed coordinate. We also obtain a prediction-risk comparison for the full diagonal family.

Let $X = (X^{(1)}, \dots, X^{(K)})$, where $K \geq 2$ and the blocks are independent and uniform on the sphere of radius $\sqrt{3}$ in $\mathbb{R}^3$. Put $m = 3K$. Let $H_2 = (H_{21}, \dots, H_{25})^\top$ be an orthonormal basis, under the uniform probability distribution, of the restrictions of homogeneous harmonic polynomials of degree two. The prescribed response used to fit the metric is

$$V = H_2(X^{(1)}) + \sigma e,$$

where $\sigma \geq 0$ and the five coordinates of e are independent Rademacher variables, independent of X. Consider the prespecified family

$$\mathcal{G}_\delta = \left\{ (c_1 \mathbf{1}_3, \dots, c_K \mathbf{1}_3) : c_k \geq \delta, \quad \sum_{k=1}^{K} c_k = K \right\} \subset \mathcal{A}_\delta.$$

For $\kappa > 0$, define

$$\lambda_2(\kappa) = 1 - \frac{3 \coth \kappa}{\kappa} + \frac{3}{\kappa^2}, \qquad A = K - (K-1)\delta, \qquad \kappa_A = \frac{3A}{h^2}.$$

Here $h > 0$ and $\rho \geq 0$ are fixed; no lower bound on $h/\sqrt{m}$ is imposed. Put

$$d_h = \frac{12\{1 - \lambda_2(\kappa_A)\}}{5h^4 \cosh^2(\kappa_A)}, \qquad \rho_{\max} = \frac{A d_h}{1 - \delta}.$$

Proposition 50. *Suppose $0 < \delta < 1$. For every $a = (c_1 \mathbf{1}_3, \dots, c_K \mathbf{1}_3) \in \mathcal{G}_\delta$,*

$$\operatorname{cov}(X, V) = 0, \qquad \mathcal{R}_{\infty,\rho}(a, h) = \sigma^2 + \{1 - \lambda_2(3c_1/h^2)\}^2 + \frac{\rho}{2K} \sum_{k=1}^{K} (c_k - 1)^2.$$

The function λ_2 takes values in $(0, 1)$ and is strictly increasing. If $0 \leq \rho < \rho_{\max}$, the unique minimizer over $\mathcal{G}_\delta$ is

$$a^{\mathrm{G}} = (A \mathbf{1}_3, \delta \mathbf{1}_{m-3}).$$

For $\nu_h = \delta\{d_h - \rho(1-\delta)/A\} > 0$, it satisfies

$$\mathcal{R}_{\infty,\rho}(a, h) - \mathcal{R}_{\infty,\rho}(a^{\mathrm{G}}, h) \geq \frac{\nu_h}{6} \|a - a^{\mathrm{G}}\|_1, \qquad a \in \mathcal{G}_\delta.$$

For any minimizer a^{D} of $\mathcal{R}_{\infty,\rho}(\cdot, h)$ over the full family $\mathcal{A}_\delta$,

$$\mathcal{R}_{\infty,0}(a^{\mathrm{D}}, h) \leq \mathcal{R}_{\infty,0}(\mathbf{1}, h) - \Delta_h,$$

where

$$\Delta_h = \{1 - \lambda_2(3/h^2)\}^2 - \{1 - \lambda_2(\kappa_A)\}^2 - \frac{\rho}{2}(K-1)(1-\delta)^2 > 0.$$

The full-family risk inequality also holds for any $\rho \geq 0$ with $\Delta_h > 0$, without the condition $\rho < \rho_{\max}$.

The proposition separates two conclusions. It identifies the metric within $\mathcal{G}_\delta$ and bounds the prediction risk of a minimizer over $\mathcal{A}_\delta$. It does not identify the support of the latter minimizer. All constants are explicit for each K and h; the bound on ρ need not be uniform as h or K changes.

For a^{G}, selection with any threshold in $(0, A - \delta)$ gives the first sphere with unit representation weights. Its first eight nonconstant population eigenfunctions contain the three degree-one and five degree-two harmonics, and hence its conditional-mean approximation error is $A_8 = 0$. The isotropic product-sphere representation has $A_8 = 1$. Thus this response distinguishes the two population

spectral representations even though every linear predictor based on X has normalized response risk at least $1 + \sigma^2$. The isotropic eight-dimensional eigenspace has no external eigengap; this approximation comparison does not apply a finite-sample eigenspace convergence result to it. On the block-isotropic family, the population Gaussian row integrals are constant, and row normalization gives the same population operator. The example establishes nonlinear predictive content and a representation benefit from selection, without comparing the two normalization rules.

Proof of Proposition 50. Every degree-two harmonic has mean zero, variance one under the chosen normalization, and is an even function of its argument. Central symmetry of the first sphere gives $\mathbb{E}\{X^{(1)} H_2(X^{(1)})^\top\} = 0$. Independence gives the same identity for the other blocks, and the response errors have mean zero. This proves the covariance assertion.

Write $x = \sqrt{3}u$ and $x' = \sqrt{3}u'$ in one sphere block. For its weight c, the Gaussian kernel equals

$$\exp\{-c\|x - x'\|_2^2/(2h^2)\} = e^{-\kappa} e^{\kappa u^\top u'}, \qquad \kappa = 3c/h^2.$$

Its integral with respect to u' is $e^{-\kappa}\sinh(\kappa)/\kappa$, independently of u. Its symmetric bistochastic scaling is constant. The normalized kernel on the unit sphere is $\kappa e^{\kappa u^\top u'}/\sinh\kappa$. Under the uniform probability distribution, $t = u^\top u'$ is uniform on $[-1, 1]$ before this change of density. Put $Z(\kappa) = \sinh(\kappa)/\kappa$. Rotational invariance and the trace of the second-moment matrix give

$$\mathbb{E}_\kappa(u'u'^\top \mid u) = \frac{1 - Z''(\kappa)/Z(\kappa)}{2} I_3 + \frac{3Z''(\kappa)/Z(\kappa) - 1}{2} uu^\top.$$

Each degree-two harmonic is the restriction of $x^\top Bx$ for a symmetric matrix B with $\mathrm{tr}(B) = 0$. Its image under the operator is multiplied by

$$\frac{3Z''(\kappa)/Z(\kappa) - 1}{2} = 1 - \frac{3\coth\kappa}{\kappa} + \frac{3}{\kappa^2} = \lambda_2(\kappa).$$

The full Gaussian kernel and its scaling factor over the independent blocks. Each nuisance factor integrates to one. Thus $T_{a,h} H_2(X^{(1)}) = \lambda_2(3c_1/h^2) H_2(X^{(1)})$. Orthogonality and independence of the response error prove the risk identity.

Direct differentiation gives

$$\lambda_2'(\kappa) = \frac{3\{\kappa^2 + \kappa\sinh\kappa\cosh\kappa - 2\sinh^2\kappa\}}{\kappa^3\sinh^2\kappa}.$$

The numerator inside braces has the convergent expansion

$$\kappa^2 + \frac{\kappa}{2}\sinh(2\kappa) - \cosh(2\kappa) + 1 = \sum_{n=3}^{\infty} \frac{2^{2n-1}(n-2)}{(2n)!}\kappa^{2n} \geq \frac{2\kappa^6}{45}.$$

Because $\sinh\kappa \leq \kappa\cosh\kappa$,

$$\lambda_2'(\kappa) \geq \frac{2\kappa}{15\cosh^2\kappa} > 0.$$

The power series gives $\lambda_2(\kappa) \to 0$ as $\kappa \downarrow 0$. The strict inequality $\mathbb{E}_\kappa(t^2) < 1$ gives $\lambda_2(\kappa) < 1$. Thus $0 < \lambda_2(\kappa) < 1$.

For a fixed $c_1 = c$, the remaining weights sum to $K - c$. Convexity of the squared ridge terms gives

$$\mathcal{R}_{\infty,\rho}(a, h) - \sigma^2 \geq F(c), \qquad F(c) = \{1 - \lambda_2(3c/h^2)\}^2 + \frac{\rho}{2(K-1)}(c-1)^2, \quad \delta \leq c \leq A.$$

Equality holds when $c_2 = \cdots = c_K = (K - c)/(K - 1)$, and also at $c = A$. For $\kappa = 3c/h^2 \leq \kappa_A$, the derivative bound gives

$$\frac{6}{h^2}\{1 - \lambda_2(\kappa)\}\lambda_2'(\kappa) \geq d_h c, \qquad F'(c) \leq -d_h c + \frac{\rho}{K-1}(c-1).$$

For $c \leq 1$, the right side is at most $-\delta d_h$. For $1 \leq c \leq A$, use $(c-1)/c \leq (A-1)/A = (K-1)(1-\delta)/A$ to bound it above by $-c\{d_h - \rho(1-\delta)/A\}$. Under the assumed ridge bound, these two inequalities give $F'(c) \leq -\nu_h < 0$ throughout $[\delta, A]$. Integration proves

$$\mathcal{R}_{\infty,\rho}(a, h) - \mathcal{R}_{\infty,\rho}(a^{\mathrm{G}}, h) \geq F(c_1) - F(A) \geq \nu_h(A - c_1).$$

Feasibility gives $\sum_{k=2}^{K}(c_k-\delta) = A-c_1$ and hence $\|a-a^{\mathrm{G}}\|_1 = 6(A-c_1)$. This proves uniqueness and the stated bound.

The metric a^{G} is feasible for the full diagonal problem. Nonnegativity of the ridge penalty and global optimality give

$$\mathcal{R}_{\infty,0}(a^{\mathrm{D}},h) \le \mathcal{R}_{\infty,\rho}(a^{\mathrm{D}},h) \le \mathcal{R}_{\infty,\rho}(a^{\mathrm{G}},h) = \sigma^2 + \{1-\lambda_2(\kappa_A)\}^2 + \frac{\rho}{2}(K-1)(1-\delta)^2.$$

Subtract this expression from $\mathcal{R}_{\infty,0}(\mathbf{1},h) = \sigma^2 + \{1-\lambda_2(3/h^2)\}^2$. This proves the risk inequality whenever $\Delta_h > 0$. Under $\rho < \rho_{\max}$, strict decrease of F gives $\Delta_h = F(1) - F(A) > 0$.

For the spectral comparison, the uniform density makes the continuum operator a positive scalar multiple of the negative Laplace–Beltrami operator. The polar-coordinate calculation in the proof of Theorem 8 gives eigenvalue $\ell(\ell+1)/3$ and multiplicity $2\ell+1$ for degree-ℓ harmonics on the sphere of radius $\sqrt{3}$. The first eight nonconstant harmonics have degrees one and two. They contain H_2, giving $A_8 = 0$ after selection under a^{G}. On the isotropic product of K spheres, eigenfunctions are products of the block harmonics and eigenvalues add. The $3K$ single-block degree-one functions have eigenvalue $2/3$, and the $9\binom{K}{2}$ products of degree-one functions from two distinct blocks have eigenvalue $4/3$. Their combined dimension is at least eight for $K \ge 2$. Both spaces are orthogonal to $H_2(X^{(1)})$: the first by parity and independence, and the second by the zero mean of a factor from another block. Every choice of the first eight nonconstant isotropic eigenfunctions has $A_8 = 1$. Finally, $\mathrm{cov}(X,V) = 0$ and $\mathbb{E}(V) = 0$ imply that the best affine predictor from X is zero, with normalized risk $5^{-1}\mathbb{E}\|V\|_2^2 = 1+\sigma^2$. This completes the proof. □

U.2 Nonlinear prediction and sensitivity to the weight lower bound

We assess whether a response-based metric can identify predictive coordinates when each raw coordinate has zero Pearson correlation with the response. We generate 16 independent vectors uniformly on the sphere of radius $\sqrt{3}$ in $\mathbb{R}^3$, yielding 48 observed covariates. The response contains the five normalized degree-2 spherical harmonics of the first sphere and independent sign noise with standard deviation 0.4. Each conditional response mean has variance one. The conditional means are even under an antipodal change of the signal vector, which makes their covariance with every raw coordinate equal to zero. The estimator receives the 48 individual coordinates and their responses without the signal identities or sphere grouping.

Each of 100 independent data sets contains 512 training observations, 256 validation observations, and 512 test observations. We fit the response metric on eight disjoint training blocks of size 64, with $\tau = 0.5$, $h = \sqrt{48}\tau$, ridge coefficient $\rho = 10^{-4}$, and lower bound $\delta = 0.02$. The selection threshold is $\delta + 0.5$. Paired fits at $\delta = 0.01$ and 0.1 use the same observations and blocks. All spectral methods use the 1,280 observed covariate vectors, $k = \lceil 1280^{3/4} \rceil$, and eight nonconstant eigenvectors. The population sphere eigenspace containing degrees one and two has dimension $3+5=8$, allowing a linear decoder to represent the quadratic response. Each method selects its ridge-head penalty from $\{0.01, 1, 100\}$ using noisy validation responses. Metrics and decoder coefficients use training responses only. The primary error is the average squared difference between the five predicted responses and their conditional means at the 512 test observations.

The correlation comparison selects the three largest sums of squared Pearson correlations with the five responses. It is supplied the support cardinality three but no feature identities. We also compare isotropic Laplacian eigenmaps, linear and spectral partial least squares with three or eight components, and diagonal metric learning for kernel regression (MLKR) followed by the same eight-dimensional spectral decoder. Diagonal MLKR selects among six metric candidates, with initialization scales $\{0.5, 1, 2\}$ and penalties $\{0, 0.001\}$, using the same validation observations. An oracle-support representation and a PALE representation retaining all fitted weights provide additional comparisons. PALE uses four prescribed initializations; each PALE and diagonal-MLKR optimization has a limit of 120 iterations. The initializations and total numbers of metric candidates differ, and no computational advantage is claimed for either method.

Table 20 reports conditional-mean prediction errors. Numbers in parentheses represent standard deviations across the 100 data sets. The paired differences subtract each method's error from the error of PALE at $\delta = 0.02$; the normal-approximation intervals equal the average paired difference plus or minus 1.96 times its empirical standard error. PALE has mean error 0.014020, compared

Table 20: Conditional-mean prediction errors under nonlinear sphere responses.

Method	MSE (SD)	Paired difference	95% paired interval
PALE, $\delta = 0.01$	0.0140 (0.0034)	3.27×10^{-7}	$[-1.03,\ 7.57] \times 10^{-7}$
PALE, $\delta = 0.02$	0.0140 (0.0034)	—	—
PALE, $\delta = 0.1$	0.0140 (0.0034)	-3.42×10^{-6}	$[-7.01,\ 0.18] \times 10^{-6}$
PALE, all weights	0.0140 (0.0033)	-5.89×10^{-6}	$[-3.20,\ 2.02] \times 10^{-5}$
Correlation, top three	1.0012 (0.0698)	-0.9872	$[-1.0010,\ -0.9734]$
Isotropic LE	1.0139 (0.0062)	-0.9999	$[-1.0014,\ -0.9985]$
Oracle support	0.0140 (0.0033)	4.08×10^{-5}	$[2.07,\ 6.09] \times 10^{-5}$
PLS linear	1.0480 (0.0102)	-1.0340	$[-1.0361,\ -1.0318]$
PLS spectral	1.0529 (0.0123)	-1.0389	$[-1.0415,\ -1.0363]$
Diagonal MLKR spectral	0.0144 (0.0033)	-4.22×10^{-4}	$[-5.37,\ -3.08] \times 10^{-4}$

with 1.001237 for correlation screening and 1.013950 for the isotropic graph. Its error is close to the oracle-support error of 0.013979 and the diagonal-MLKR error of 0.014442. PALE selects exactly the three signal coordinates in all 100 data sets at each of the three lower bounds. The correlation comparison recovers the complete support in none of the 100 data sets. Retaining all PALE weights gives mean error 0.014026. This comparison shows little predictive effect from the final threshold in this design, after the metric has concentrated on the signal coordinates.

The fitted kernels are evaluated at a finite bandwidth. Across the 100 primary fits, the median of the pre-scaling exponent $-\log K_{ij}$ ranges from 3.895 to 4.066; the averages of its within-sample 10th and 90th percentiles are 0.860 and 7.137. All 1,200 PALE starts report successful optimizer termination. Across the 300 selected fits, the unit-step simplex gradient-mapping norm is at most 1.584×10^{-5}. Among the 600 diagonal-MLKR candidates, one reaches the iteration limit; validation selects a converged candidate in every data set. Proposition 50 applies to the population group-isotropic family at these settings: for $\delta = 0.02$, its sufficient ridge upper bound is $\rho_{\max} = 2.372 \times 10^{-4}$ and its normalized predictive-risk improvement bound is $\Delta_h = 0.6659$. These population values do not certify the finite-sample optimizations. These fits have no global objective-gap certificate. The support guarantee in the block-isotropic population family does not establish a global identification result for unrestricted diagonal metrics.

U.3 Label-free prediction and classification of handwritten digits

We examine whether predicting omitted measurements improves a representation of unmodified public data. The UCI Optical Recognition of Handwritten Digits data (Alpaydin & Kaynak, 1998), distributed as the test-set copy in scikit-learn, contain 1,797 images with 64 integer pixel values between 0 and 16. We add no nuisance coordinates. We center each coordinate using a fixed cohort of 512 images and divide every coordinate by the square root of the average of its 64 training variances. This common scale preserves the relative pixel variances.

The 512 metric observations form four disjoint blocks of size 128. Within each block, $J = 16$ random masks retain 32 of the 64 coordinates. We use $\delta = 0.02$, $\rho = 0.0005$, threshold $\delta + 0.5$, and prediction scales $h/\sqrt{64} \in \{0.6, 0.9, 1.2\}$. Each fit selects the bandwidth by its training criterion and uses the same fixed limit of 400 optimization iterations per bandwidth. We repeat this fit under 100 independently generated block and mask allocations of the same cohort. These are not 100 independent image data sets. We assess the unpenalized masked-prediction loss on a separate fixed set of 512 images with fixed assessment masks. The isotropic comparison uses the same bandwidth selected by each fitted metric; it does not separately optimize its bandwidth.

For classification, the learned metric defines an eight-dimensional spectral representation of all 1,797 observed covariate vectors, with 24 neighbors. We compare thresholded and all-weight PALE, isotropic LE, PCA fitted on the metric cohort, and Laplacian Score and SSFS feature selection followed by the same spectral graph. Laplacian Score uses a 24-neighbor graph and heat scale equal to the median positive squared edge distance. SSFS uses 500 subsamples. Both rankings are computed once on the metric cohort without digit labels, and each retains the same number of features

Table 21: Label-free pixel prediction and digit classification.

Masked pixel prediction		
Method	Error	PALE minus isotropic
PALE metric	0.6301 (0.0020)	-0.1222 (0.0020)
Isotropic metric	0.7523 (0.0000)	—
Transductive classification		
Method	Accuracy (%)	PALE minus comparator (percentage points)
PALE selected	97.603 (0.178)	—
PALE all weights	97.639 (0.160)	-0.036 (0.084)
Isotropic LE	98.275 (0.000)	-0.672 (0.178)
PCA	95.604 (0.000)	1.999 (0.178)
Laplacian Score + LE	97.559 (0.089)	0.045 (0.194)
SSFS + LE	95.842 (0.111)	1.761 (0.209)

as the corresponding PALE fit. Digit labels are used only to fit and assess a seven-nearest-neighbor classifier in five fixed non-stratified folds. All image covariates enter each spectral graph, making this a transductive assessment.

We selected $J = 16$ from $J \in \{1, 2, 16\}$ using exploratory omitted-pixel prediction errors before examining the pilot classification results. Those classification results were available before the 100-allocation study; no subsequent result changed its settings. The final pixel-assessment rows exclude the first pilot's fitting and tuning rows, but 150 and 160 overlap the second pilot's fitting and tuning rows, respectively. The same image covariates also enter the exploratory transductive graphs. This is a fixed-data assessment after exploratory design selection, not an independent external confirmation.

Table 21 reports prediction loss and classification accuracy. Numbers in parentheses represent standard deviations over the 100 block and mask allocations. Differences subtract each comparison from PALE. These standard deviations measure variation within the fixed data and assessment folds, not uncertainty in population classification accuracy. The zero standard deviations for isotropic LE and PCA arise from their fixed representations and assessment folds.

All 100 fits select $h/\sqrt{64} = 0.6$ and retain an average of 30.50 coordinates, with standard deviation 1.04 and range 28–33. Mean omitted-pixel prediction error decreases from 0.752312 for the isotropic metric to 0.630110 for PALE, a reduction of 16.24%; every allocation reduces this error. This comparison evaluates the fitted weights before representation thresholding. The thresholded and all-weight graphs share the same fitted prediction rule.

Classification gives a different comparison. PALE has mean accuracy 97.603%, compared with 95.604% for PCA, 95.842% for SSFS, and 97.559% for Laplacian Score. Isotropic LE has accuracy 98.275% and exceeds PALE in every allocation. Retaining all fitted weights gives mean accuracy 97.639%. Thus the lower omitted-pixel prediction loss does not establish improved digit classification, and the difference from Laplacian Score is small in this fixed-data assessment.

All 300 bandwidth candidates terminate successfully under the local optimization criterion. The largest unit-step simplex gradient-mapping norm among the 100 selected fits is 5.035×10^{-6}. Their global objective-gap bounds range from 0.540929 to 0.625495; none meets the requested tolerance 0.001. These diagnostics do not certify the shrinking optimization error assumed in the asymptotic results.

References for the Supplementary Materials

10x Genomics. 10k PBMCs from a healthy donor: Gene expression with a panel of TotalSeq-B antibodies. Public dataset, Cell Ranger 3.0.0, 2018. URL `https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-gene-expression-and-cell-surface-protein-3-standard-3-0-0`.

Published November 19, 2018. Accessed August 31, 2026.

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50P49. Dataset.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. doi: 10.1214/009053606000001217.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533. PMLR, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Research*, 44(21): 10074–10090, 2016. doi: 10.1093/nar/gkw964.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs. *Applied and Computational Harmonic Analysis*, 60:123–175, 2022. doi: 10.1016/j.acha.2022.02.004.

Xiuyuan Cheng, Yixuan Tan, and Nan Wu. Improved convergence rate of kNN graph Laplacians: Differentiable self-tuned affinity. *Applied and Computational Harmonic Analysis*, 86:101919, 2027. doi: 10.1016/j.acha.2026.101919.

Dengxin Dai, Till Kroeger, Radu Timofte, and Luc Van Gool. Metric imitation by manifold transfer for efficient vision applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3527–3536, 2015. doi: 10.1109/CVPR.2015.7298975.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Xiao Dong, Lei Zhu, Xuemeng Song, Jingjing Li, and Zhiyong Cheng. Adaptive collaborative similarity learning for unsupervised multi-view feature selection. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2064–2070, 2018. doi: 10.24963/ijcai.2018/285.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning Laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016. doi: 10.1109/TSP.2016.2602809.

David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003. doi: 10.1073/pnas.1031596100.

Bahare Fatemi, Layla El Asri, and Seyed Mehran Kazemi. SLAPS: Self-supervision improves structure learning for graph neural networks. In *Advances in Neural Information Processing Systems 34*, pp. 22667–22681. Curran Associates, Inc., 2021.

Fatemeh Fathinezhad, Peyman Adibi, and Bijan Shoushtarian. Soft laplacian eigenmap: A novel regularized graph-based method for dimensionality reduction with local geometry preservation. *Advances in Data Analysis and Classification*, 2025. doi: 10.1007/s11634-025-00658-1.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1972–1982, 2019.

Nicolás García Trillos. Variational limits of k-NN graph-based functionals on data clouds. *SIAM Journal on Mathematics of Data Science*, 1(1):93–120, 2019. doi: 10.1137/18M1188999.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021. doi: 10.1038/s41592-020-01050-x. URL https://www.nature.com/articles/s41592-020-01050-x.

Charles J. Geyer. On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 22(4): 1993–2010, 1994. doi: 10.1214/aos/1176325768.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8238499/.

Jesse He, Tristan Brugère, and Gal Mishne. Product manifold learning with independent coordinate selection. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, volume 221 of *Proceedings of Machine Learning Research*, pp. 267–277, 2023.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/b5b03f06271f8917685d14cea7c6c50a-Paper.pdf.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. GraphMAE2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM Web Conference 2023*, pp. 737–746, 2023. doi: 10.1145/3543507.3583379.

Geert-Jan Huizing, Laura Cantini, and Gabriel Peyré. Unsupervised ground metric learning using Wasserstein singular vectors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9429–9443, 2022.

Jiayan Jiang, Bo Wang, and Zhuowen Tu. Unsupervised metric learning by self-smoothing operator. In *2011 International Conference on Computer Vision*, pp. 794–801, 2011. doi: 10.1109/ICCV.2011.6126318.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 920–929, 2016.

Boris Landa and Xiuyuan Cheng. Robust inference of manifold density and geometry by doubly stochastic scaling. *SIAM Journal on Mathematics of Data Science*, 5(3):589–614, 2023. doi: 10.1137/22M1516968.

Roy R. Lederman and Ronen Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018. doi: 10.1016/j.acha.2015.09.002. URL https://www.sciencedirect.com/science/article/pii/S1063520315001190.

Ann B. Lee and Rafael Izbicki. A spectral series approach to high-dimensional nonparametric regression. *Electronic Journal of Statistics*, 10(1):423–463, 2016. doi: 10.1214/16-EJS1112.

Changhee Lee, Fergus Imrie, and Mihaela van der Schaar. Self-supervision enhanced feature selection with correlated gates. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oDFvtxzPOx.

John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009. doi: 10.1016/j.neucom.2008.12.017.

Shira Lifshitz, Ofir Lindenbaum, Gal Mishne, Ron Meir, and Hadas Benisty. Unsupervised feature selection through group discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23496–23504, 2026. doi: 10.1609/aaai.v40i28.39521.

Ofir Lindenbaum, Uri Shaham, Erez Peterfreund, Jonathan Svirsky, Nicolas Casey, and Yuval Kluger. Differentiable unsupervised feature selection based on a gated Laplacian. In *Advances in Neural Information Processing Systems 34*, pp. 1530–1542, 2021.

Yixin Liu, Yu Zheng, Daokun Zhang, Hongxu Chen, Hao Peng, and Shirui Pan. Towards unsupervised deep graph structure learning. In *Proceedings of the ACM Web Conference 2022*, pp. 1392–1403, 2022. doi: 10.1145/3485447.3512186.

Marco Loog. Semi-supervised linear discriminant analysis through moment-constraint parameter estimation. *Pattern Recognition Letters*, 37:24–31, 2014. doi: 10.1016/j.patrec.2013.03.004.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Nicholas F. Marshall and Ronald R. Coifman. Manifold learning with bi-stochastic kernels. *IMA Journal of Applied Mathematics*, 84(3):455–482, 2019. doi: 10.1093/imamat/hxy065.

André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, 2016.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles (eds.), *Algorithmic Learning Theory*, volume 9925 of *Lecture Notes in Computer Science*, pp. 3–17, Cham, 2016. Springer. doi: 10.1007/978-3-319-46379-7_1. URL https://doi.org/10.1007/978-3-319-46379-7_1.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Xueying Niu, Chaowei Zhang, Xiaojie Zhao, Lihua Hu, and Jifu Zhang. A multi-view ensemble clustering approach using joint affinity matrix. *Expert Systems with Applications*, 216:119484, 2023. doi: 10.1016/j.eswa.2022.119484.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323.

Daniel Segal, Ofir Lindenbaum, and Ariel Jaffe. Spectral self-supervised feature selection. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=t0EJiOd9Lg.

Jun Shao and C. F. J. Wu. A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197, 1989. doi: 10.1214/aos/1176347263.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. doi: 10.2140/pjm.1967.21.343.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017. doi: 10.1038/nmeth.4380.

Ronen Talmon and Hau-Tieng Wu. Latent common manifold learning with alternating diffusion: Analysis and applications. *Applied and Computational Harmonic Analysis*, 47(3):848–892, 2019. doi: 10.1016/j.acha.2017.12.006. URL https://www.sciencedirect.com/science/article/pii/S1063520318300034.

Joseph G. Taylor, Viktoriia Sharmanska, Kristian Kersting, David Weir, and Novi Quadrianto. Learning using unselected features (LUFe). In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2060–2066, 2016.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Hugues Van Assel, Titouan Vayer, Rémi Flamary, and Nicolas Courty. SNEkhorn: Dimension reduction with symmetric entropic affinities. In *Advances in Neural Information Processing Systems 36*, pp. 44470–44487, 2023. doi: 10.52202/075280-1925.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9780511802256.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks—ICANN 2001*, volume 2130 of *Lecture Notes in Computer Science*, pp. 485–491. Springer, 2001. doi: 10.1007/3-540-44668-0_68.

Bo Wang and Zhuowen Tu. Affinity learning via self-diffusion for image segmentation and clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2312–2319, 2012. doi: 10.1109/CVPR.2012.6247942.

Sheng Wang, Jianfeng Lu, Xingjian Gu, Haishun Du, and Jingyu Yang. Semi-supervised linear discriminant analysis for dimension reduction and classification. *Pattern Recognition*, 57:179–189, 2016. doi: 10.1016/j.patcog.2016.02.019.

Shulei Wang. Augmentation invariant manifold learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):978–1000, 2025. doi: 10.1093/jrsssb/qkaf003.

Kilian Q. Weinberger and Gerald Tesauro. Metric learning for kernel regression. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 612–619, 2007.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. doi: 10.1016/S0169-7439(01)00155-1.

Jiayi Xu, Qirui Huang, Yang Zhao, Jihong Pei, and Xuan Yang. Adaptive local neighborhood window Laplacian eigenmaps. In *Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition*, pp. 1–7, 2024. doi: 10.1145/3704323.3704335.

Junchen Yang, Ofir Lindenbaum, Yuval Kluger, and Ariel Jaffe. Multi-modal differentiable unsupervised feature selection. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2400–2410, 2023.

Shuo Yang, Sujay Sanghavi, Holakou Rahmanian, Jan Bakus, and S. V. N. Vishwanathan. Toward understanding privileged features distillation in learning-to-rank. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26658–26670, 2022. doi: 10.52202/068431-1933.

Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2018. doi: 10.1109/TIP.2017.2765836.

Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13:1–26, 2012.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004. doi: 10.1137/S1064827502419154.

Mingbo Zhao, Zhao Zhang, Tommy W. S. Chow, and Bing Li. A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. *Neural Networks*, 55:83–97, 2014. doi: 10.1016/j.neunet.2014.03.005.

Lili Zhou, Min Lu, and Hemant Ishwaran. Variable priority for unsupervised variable selection. *Pattern Recognition*, 172:112727, 2026. doi: 10.1016/j.patcog.2025.112727.

Jinfeng Zhuang, Jialei Wang, Steven C. H. Hoi, and Xiangyang Lan. Unsupervised multiple kernel learning. In *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pp. 129–144, 2011.