When Do Unseen Data Concentrate at Class-Specific Points? Kernel Data Piling of High-Dimensional Small-Sample-Size Data

Anonymous authors
Paper under double-blind review

Abstract

We consider classification problems in which observations are represented by high-dimensional feature vectors while only a small, fixed number of labeled training samples are available from each class. For this purpose, we work under a high-dimension-low-sample-size (HDLSS) asymptotic regime, where the feature dimension tends to infinity while the training sample sizes remain fixed. We study when independent test observations, after being mapped by a kernel classifier fitted to these labeled samples, concentrate at class-specific locations. We identify conditions under which this class-specific test concentration occurs and characterize when the resulting class locations are distinct for several commonly used kernels. We further show that components with very large within-class variation can prevent test-score concentration throughout the empirical span generated by the training sample. To address this failure, we propose a kernel correction that restores separated class-specific concentration under suitable conditions. Simulation studies provide empirical support for the theoretical results developed throughout the paper.

1 Introduction

Modern classification problems often combine many features with few labeled examples. In few-shot learning, a pretrained encoder maps observations to high-dimensional representations, while a new task supplies only a handful of labeled examples per class. A lightweight map fitted to these frozen features can nevertheless yield strong predictive performance (Tian et al., 2020; Luo et al., 2023; Bär et al., 2024; Huang et al., 2024). A useful geometric perspective is to examine whether its scores for independent test observations concentrate around well-separated class-specific points, allowing prediction by comparison with appropriately estimated class centers. For nonlinear maps, this geometry depends on both the class distributions and how the map transforms them. Understanding this interaction helps determine when concentration observed on the training data reflects class structure beyond those observations.

Geometry of this kind is known to arise in trained deep classifiers and to matter for prediction. Papyan et al. (2020) documented *neural collapse*: late in training, the last-layer features of the training observations from each class concentrate at their class mean, while the class means approach a maximally separated symmetric configuration that the final linear classifier matches. Subsequent work studied the emergence of this geometry in simplified feature models and training dynamics (Zhu et al., 2021; Han et al., 2022), as well as its development across network layers (Rangamani et al., 2023). Galanti et al. (2022) linked collapse geometry to few-shot transfer through nearest-center error bounds, while Goldblum et al. (2020); Wang et al. (2023); Li et al. (2024) used related neural collapse measures to assess pretrained-model transferability. More recently, Luthra et al. (2026) showed that full collapse is unnecessary: small within-class variation along class-separating directions can suffice even when substantial variation remains in orthogonal directions. These findings motivate examining the out-of-sample implications of classwise training collapse, not only as a feature of representation learning but also as a geometry induced by a map fitted on a fixed representation.

Maximal data piling (MDP) provides a concrete approach to this downstream perspective in the high-dimension, low-sample-size (HDLSS) setting, where dimension grows while the training sam-

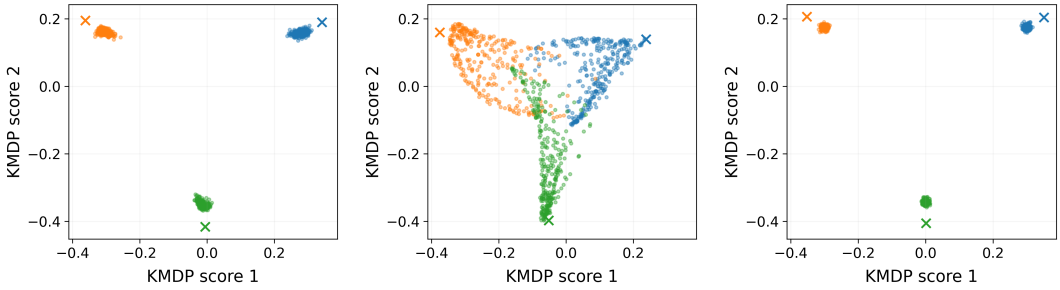
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

ple sizes remain fixed (Hall et al., 2005; Ahn et al., 2007). In binary linear classification, MDP maximizes between-class separation subject to identical within-class training projections (Ahn & Marron, 2010). Building on this geometry, Chang et al. (2021); Kim et al. (2024) established *second data piling* for independent test observations under suitable spiked-covariance conditions through projected negative-ridge constructions in binary and multiclass settings.

Training data piling is also related to broader questions of interpolation and generalization in high-dimensional classification. A prominent example arises in support vector machines (SVMs): when training observations are projected onto the SVM normal direction, support vectors share the two class-specific margin values. In HDLSS settings, Marron et al. (2007) identified such margin piling as a potential source of poor generalization. More recent work studies regimes of *support vector proliferation*, where every training observation lies on a hard-margin SVM margin (Hsu et al., 2021; Ardeshir et al., 2021). Under suitable conditions, the SVM coincides with the minimum-norm label interpolant and can nevertheless generalize well (Muthukumar et al., 2021). Related results have extended this connection to kernel SVMs and more general high-dimensional classification settings (Kaushik et al., 2024; Hashimoto et al., 2025).

The interpolation analyses above let the sample size grow with the dimension, whereas a few-shot task fixes the number of labeled examples. We therefore work in the HDLSS regime, letting the dimension grow with the classwise training sizes held fixed. Within it, we pose the piling problem over an entire reproducing kernel Hilbert space: among all score maps that send each class's training observations to a single point, which separates those points the most? We show that the maximizing subspace is unique, lies in the span of the kernel evaluations at the training sample, and admits a closed form, giving the *kernel maximal data piling* (KMDP) map, the kernel analogue of multiclass MDP (Kim et al., 2024) (Section 2). Our first asymptotic result gives conditions under which independent test scores also pile at class-specific locations, together with a criterion, explicit for each kernel, for those locations to be distinct (Section 3). The criterion separates kernels: the Gaussian kernel produces distinct piles when the classes differ only in residual variance, while the linear and polynomial kernels do not.

We further identify a regime in which components with large within-class variation prevent the test data piling. We then propose a spike-subspace correction that consistently estimates and removes the corresponding high-variation subspace, and show that this correction restores test piling under suitable conditions (Section 4). Figure 1 illustrates these findings in simulated three-class, five-shot problems: exact train data piling occurs in all three panels, while test scores concentrate at class-specific locations in (a), remain dispersed in the presence of large variation in (b), and concentrate again after correction in (c).



(a) Test piling (b) No test piling (c) Test piling after correction

Figure 1: Train and test scores in a simulated three-class, five-shot problem. Training observations pile exactly in all three panels. Test scores pile when the dominant within-class variation is small relative to p (a), do not when it is of order p (b), and pile again once the directions carrying it are estimated and removed from the features (c). Section 4.2 describes this correction. Crosses denote train piles, dots denote test scores, and colors indicate classes.

2 PILING OF TRAINING DATA

Let $\mathcal{H}$ be the RKHS of a positive-semidefinite kernel $k(\cdot,\cdot)$ on an input space $\mathcal{X}$. The finite-sample construction allows a general nonempty $\mathcal{X}$, whereas the subsequent asymptotic analysis takes $\mathcal{X} = \mathbb{R}^p$. An observation $X \in \mathcal{X}$ may contain measured covariates or transformed features from a pretrained encoder, which is held fixed for a few-shot task.

For $K \geq 2$ classes, we observe labeled training observations $X_{k,1}, \ldots, X_{k,n_k}$, with $n_k \geq 2$ for $k = 1, \ldots, K$. Write $N = \sum_{k=1}^{K} n_k$ and index the pooled sample as $(X_i, y_i)_{i=1}^{N}$. Later asymptotic analysis assumes $p \to \infty$ with n_k fixed. The notation $X_{k,0}$ denotes a class-k test observation independent of the training sample.

2.1 KERNEL MAXIMAL DATA PILING

Consider a *score map* $g = (g_1, \ldots, g_{K-1})^\top : \mathcal{X} \to \mathbb{R}^{K-1}$ with $g_1, \ldots, g_{K-1} \in \mathcal{H}$. Its outputs are real-valued discriminant coordinates, not class labels or probabilities. Let $\bar{g}_k = n_k^{-1} \sum_{i=1}^{n_k} g(X_{k,i})$ and $\bar{g} = N^{-1} \sum_{k=1}^{K} n_k \bar{g}_k$ denote the classwise and overall mean scores, respectively.

We seek a maximizer of the following *KMDP problem*:

$$\begin{aligned} \tilde{g} \in & \underset{g_1,\ldots,g_{K-1}\in\mathcal{H}}{\arg\max} \sum_{k=1}^{K} n_k \|\bar{g}_k - \bar{g}\|^2 \\ \text{subject to} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n_k} \|g(X_{k,i}) - \bar{g}_k\|^2 = 0, \quad \langle g_j, g_\ell \rangle_{\mathcal{H}} = \delta_{j\ell}, \quad 1 \leq j, \ell \leq K-1, \end{aligned} \tag{1}$$

where $\delta_{j\ell}$ denotes the Kronecker delta. The constraints enforce exact training-data piling, so all training observations within class k have score $\bar{g}_k$, while the orthonormality constraint normalizes the coordinate functions g_k and prevents the objective from being increased arbitrarily by rescaling.

Since

$$\sum_{k=1}^{K} n_k \|\bar{g}_k - \bar{g}\|^2 = \frac{1}{2N} \sum_{k=1}^{K}\sum_{\ell=1}^{K} n_k n_\ell \|\bar{g}_k - \bar{g}_\ell\|^2,$$

the objective maximizes the weighted sum of squared distances between the piles. The span of the maximizing coordinate functions is called the *kernel maximal data piling (KMDP) subspace*, $\mathcal{W}_{\mathrm{KMDP}}$.

For the kernel representation, define $G \in \mathbb{R}^{N\times N}$, $\kappa(x) \in \mathbb{R}^N$, and $E \in \mathbb{R}^{N\times K}$ by $G_{ij} := k(X_i, X_j)$, $\kappa(x)_i := k(x, X_i)$, and $E_{ik} := \mathbf{1}\{y_i = k\}$. Also let $\mathcal{S}_X := \mathrm{span}\{k(X_i,\cdot) : 1 \leq i \leq N\}$, and for $\alpha \in \mathbb{R}^N$, write $f_\alpha := \sum_{i=1}^{N} \alpha_i k(X_i,\cdot)$. The following lemma identifies the optimizer and shows that its coordinate functions lie in $\mathcal{S}_X$.

Lemma 1 (The kernel MDP subspace and its representation)**.** *Suppose $G \succ 0$. Let $T \in \mathbb{R}^{K\times(K-1)}$ satisfy $T^\top T = I_{K-1}$ and $T^\top \mathbf{1}_K = 0$. Let $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{K-1}] \in \mathbb{R}^{N\times(K-1)}$ and $h \in \mathbb{R}^{K-1}$ solve*

$$\begin{pmatrix} G & \mathbf{1}_N \\ \mathbf{1}_N^\top & 0 \end{pmatrix} \begin{pmatrix} Q \\ h^\top \end{pmatrix} = \begin{pmatrix} ET \\ 0 \end{pmatrix}. \tag{2}$$

Then Q has rank $K-1$, and (1) has a unique maximizing subspace $\mathcal{W}_{\mathrm{KMDP}} = \mathrm{span}\{f_{\boldsymbol{q}_1}, \ldots, f_{\boldsymbol{q}_{K-1}}\} \subset \mathcal{S}_X$. Every maximizing tuple is an RKHS-orthonormal basis of this subspace. In particular,

$$A = Q(Q^\top G Q)^{-1/2}, \qquad g(x) = A^\top \kappa(x) \tag{3}$$

solves (1) and satisfies $A^\top G A = I_{K-1}$.

To describe the training-score geometry induced by the KMDP map, note that the ith row of GA is $g(X_i)^\top$. From (2), $GA = (ET - \mathbf{1}_N h^\top)(Q^\top G Q)^{-1/2}$. Rows of ET are identical within each class and distinct across classes, and these properties are preserved after subtracting the common vector $h^\top$ and multiplying by the invertible normalizing matrix. Hence the fitted map produces exactly K distinct training piles. The second block equation of (2), $\mathbf{1}_N^\top Q = 0$, excludes the constant-training-evaluation direction that would shift all piles equally without affecting their separation.

The maximizing subspace is unique, although its orthonormal basis is not. Choosing another basis only rotates or reflects all score vectors, so pairwise distances are preserved. We use (3) to fix coordinates throughout the paper. This exact training-piling property, however, does not by itself imply the concentration of independent test scores.

3 PILING OF INDEPENDENT TEST DATA

Having established exact piling of the training observations, we now study the behavior of the KMDP score map on unseen test observations. Chang et al. (2021) called the phenomenon in which independent test observations from the same class share a common projected value *second data piling*, and Kim et al. (2024) extended this notion to multiclass classification. Here we study the analogous phenomenon for KMDP scores as $p \to \infty$ with fixed training sample sizes.

Let $\mathcal{F}_p$ be the σ-field of all fitting information, including the training sample. For the fitted score map g, define the conditional class-k test-score mean and covariance by $m_k := \mathbb{E}[g(X_{k,0}) \mid \mathcal{F}_p]$ and $C_k := \mathrm{Cov}(g(X_{k,0}) \mid \mathcal{F}_p)$. These quantities describe the conditional center and within-class variation of an independent class-k test score. Our second-piling results show that C_k vanishes while m_k converges to a class-specific limiting location. Throughout, $\searrow$ and $\nearrow$ denote monotone convergence from above and below, respectively, while $\xrightarrow{P}$ and $\xrightarrow{D}$ denote convergence in probability and in distribution.

3.1 A SPIKED HDLSS FEATURE MODEL

We model each observation from class k as the following spiked covariance model

$$X_{k,i} = \mu_k + p^{\beta/2} U \xi_{k,i} + R_k^{1/2} \varepsilon_{k,i}, \quad 0 \le \beta \le 1, \quad k \in \{1, \dots, K\}. \tag{4}$$

Models of this type are widely used in high-dimensional studies (Jung et al., 2018; Yata & Aoshima, 2020; Fan et al., 2021; Gedon et al., 2024; Yoon & Jung, 2025). Here $\mu_k \in \mathbb{R}^p$ is the deterministic class mean, and $U \in \mathbb{R}^{p \times r}$ has deterministic orthonormal columns representing the common spike directions. The random vectors $\xi_{k,i} \in \mathbb{R}^r$ and $\varepsilon_{k,i} \in \mathbb{R}^p$ are centered with covariances $\Theta_k \succ 0$ and I_p, respectively. The resulting covariance is $\mathrm{Cov}(X_{k,i}) = p^\beta U \Theta_k U^\top + R_k$, where $R_k \in \mathbb{R}^{p \times p}$ is the residual covariance. The spike contribution to the dimension-scaled input geometry is asymptotically negligible for $\beta < 1$ but contributes at a non-vanishing scale when $\beta = 1$. We suppress the dependence of μ_k, U, and R_k on p throughout.

Following common HDLSS asymptotic formulations, we impose the following regularity conditions on the dimension-scaled mean geometry and the diffuse residual variation.

Assumption 1 (Feature geometry and moments)**.** As $p \to \infty$, K, r, and $n_1, \dots, n_K$ remain fixed. For each class k, $\xi_{k,i}$ has a fixed distribution with finite fourth moments, and $\varepsilon_{k,i}$ has independent coordinates with uniformly bounded fourth moments. The observations are independent across (k, i), with $\xi_{k,i}$ independent of $\varepsilon_{k,i}$ within each observation. For every $k, l \in \{1, \dots, K\}$,

$$\mu_k^\top \mu_l / p \to \gamma_{kl} \in \mathbb{R}, \qquad U^\top \mu_k / \sqrt{p} \to \boldsymbol{a}_k \in \mathbb{R}^r, \tag{5}$$

$$\mathrm{tr}(R_k)/p \to \sigma_k^2 \in (0, \infty), \qquad \mathrm{tr}(R_k^2)/p^2 \to 0. \tag{6}$$

The limits in (5) are direct multiclass analogs of the mean-scaling and mean–spike alignment conditions used in double data piling theory (Chang et al., 2021; Kim et al., 2024). The first limit specifies a stable dimension-scaled geometry of the class means, while the second limit controls their asymptotic alignment with the finite-rank spike subspace. Unlike the separation assumptions imposed there, we do not impose mean separation at the feature-model level; the kernel-specific criteria stated later in the next section determine class separation.

The first limit in (6) retains a nondegenerate average residual variance, while the second requires the residual variation to be diffuse rather than concentrated in a few directions (Ahn et al., 2007; Jung & Marron, 2009). The model allows class-dependent Θ_k and R_k, which are weaker than Chang et al. (2021); Kim et al. (2024). Together with the fourth-moment and independence conditions, these assumptions yield concentration of residual norms and independent residual inner products without requiring Gaussianity or the ρ-mixing conditions used in earlier HDLSS and double-data-piling analyses (Jung & Marron, 2009; Chang et al., 2021; Kim et al., 2024).

3.2 KERNEL CONDITIONS FOR TEST PILING

We impose the following conditions on the kernel function.

Assumption 2 (Kernel conditions for test piling)**.** The kernel function has the form

$$k(x,z) = \psi\left(\frac{\|x\|^2}{p}, \frac{x^\top z}{p}, \frac{\|z\|^2}{p}\right).$$

The function $\psi(u,t,v)$ is continuous, symmetric in its first and third arguments, and defines a positive-semidefinite kernel. In addition, there exist a finite constant L and a fixed $d_\psi \geq 0$ such that

$$\psi(u,u,u) > \psi(u,t,u) \quad (0 \leq t < u), \qquad |\psi(u,t,v)| \leq L\{1 + (uv)^{d_\psi/2}\} \quad (|t| \leq \sqrt{uv}). \tag{7}$$

If $d_\psi > 2$, we additionally assume that, for some $\epsilon > 0$,

$$\sup_{k,p} \mathbb{E}(\|X_{k,0}\|/\sqrt{p})^{2d_\psi+\epsilon} < \infty. \tag{8}$$

We consider kernels whose values depend on the dimension-scaled squared norms and inner product of their two inputs. The first condition in (7) ensures that the self-similarity is strictly larger than the limiting similarity between two distinct observations from the same class. This keeps the limiting train Gram matrix positive definite. The second condition controls the growth of the kernel profile through the exponent d_ψ and is needed to obtain the moment control needed for test-score concentration. Gaussian, linear, and polynomial kernels satisfy these conditions, with their corresponding growth exponents d_ψ summarized in Table 1.

For kernels with $d_\psi > 2$, we need an additional moment condition (8) to the base fourth-moment conditions in Assumption 1. This condition gives a moment strictly above order d_ψ for the dimension-scaled squared norm $\|X_{k,0}\|^2/p$, which is needed for the required uniform integrability. Sub-Gaussian tails imply these moments but are not necessary.

3.3 TEST DATA PILING UNDER WEAK SPIKES

We first consider $\beta < 1$, for which the finite-rank variation is asymptotically negligible in the classwise kernel geometry. In this regime, the KMDP classifier also induces classwise piling of independent test observations. The following theorem characterizes the resulting test-pile locations and their separation.

Theorem 1 (Test data piling under weak spikes)**.** *Suppose Assumptions 1 and 2 hold with $\beta < 1$. Define $\ell, \eta \in \mathbb{R}^K$, $B \in \mathbb{R}^{K\times K}$, and $J \in \mathbb{R}^{(K-1)\times(K-1)}$ by*

$$\ell_k = \gamma_{kk} + \sigma_k^2, \qquad B_{kl} = \psi(\ell_k, \gamma_{kl}, \ell_l), \tag{9}$$
$$\eta_k = \psi(\ell_k, \ell_k, \ell_k) - B_{kk}, \qquad J = T^\top \{B + \mathrm{diag}(\eta_1/n_1, \ldots, \eta_K/n_K)\} T.$$

Then J is positive definite, and the class-k train data pile $\bar{g}_k$ converges to

$$t_k^{\mathrm{tr}} := J^{-1/2} T^\top \{B + \mathrm{diag}(\eta_1/n_1, \ldots, \eta_K/n_K)\} \boldsymbol{e}_k, \tag{10}$$

where $\boldsymbol{e}_k \in \mathbb{R}^K$ denotes the kth standard basis vector.

By contrast, for an independent class-k test observation, the KMDP map in (3) satisfies

$$m_k \xrightarrow{P} t_k := J^{-1/2} T^\top B \boldsymbol{e}_k, \qquad C_k \xrightarrow{P} 0. \tag{11}$$

For $k \neq l$, the limiting test data piles are distinct if and only if

$$B_{kk} + B_{ll} - 2B_{kl} > 0.$$

Together with exact train data piling, the theorem establishes double data piling under weak spikes. Whether the resulting classwise test data piles are distinct depends on the kernel through the limiting similarity matrix B; two classes k and l are separated if and only if $B_{kk} + B_{ll} - 2B_{kl} > 0$. The displacement in (10), however, shows that the train and test data piles need not coincide.

The limiting expressions also reveal the effect of training sample size. For a balanced design $n_k = n$, the following corollary quantifies the training–test displacement and how pairwise pile separations vary with n.

Corollary 1 (Effect of training sample size). *Suppose the conditions of Theorem 1 hold and $n_k = n$ for every class. Write $t_k(n)$ and $t_k^{\mathrm{tr}}(n)$ to make their dependence on n explicit. Then, for every class k, $\|t_k^{\mathrm{tr}}(n) - t_k(n)\|^2 \le \eta_k/n$. Moreover, for $k \neq l$, as n increases,*

$$\|t_k^{\mathrm{tr}}(n) - t_l^{\mathrm{tr}}(n)\|^2 = B_{kk} + B_{ll} - 2B_{kl} + \frac{\eta_k + \eta_l}{n} \searrow B_{kk} + B_{ll} - 2B_{kl},$$

whereas

$$\|t_k(n) - t_l(n)\|^2 \nearrow B_{kk} + B_{ll} - 2B_{kl}.$$

As more labeled observations become available, the training–test displacement vanishes, while test-pile separation is nondecreasing and training-pile separation decreases toward their common limit. Thus, with few observations, the separation among the interpolated training piles can overstate the separation observed in independent test data. The limit in n is understood sequentially, after taking the high-dimensional limit for each fixed n.

Table 1 summarizes kernel-specific separation criteria, together with examples of profiles satisfying Assumption 2, their growth exponents d_ψ, and limiting similarities B_{kl}. For the linear and polynomial kernels, test piles are distinct if and only if $\Delta_{kl} := \gamma_{kk} + \gamma_{ll} - 2\gamma_{kl} > 0$, corresponding to a positive limiting squared mean distance. In contrast, the Gaussian kernel also yields distinct piles when $\Delta_{kl} = 0$ but $\sigma_k \neq \sigma_l$. In this unequal variance case, the linear and polynomial test piles coincide despite within-class concentration, illustrating that concentration alone does not guarantee separation.

Kernel profile $\psi(u, t, v)$	d_ψ	Derived class similarity B_{kl}	Separated-pile criterion
$\exp[-(u+v-2t)/(2\tau)]$	0	$\exp[-(\ell_k + \ell_l - 2\gamma_{kl})/(2\tau)]$	$\Delta_{kl} > 0$ or $\sigma_k \neq \sigma_l$
t	1	γ_{kl}	$\Delta_{kl} > 0$
$(\nu + t)^m$	m	$(\nu + \gamma_{kl})^m$	$\Delta_{kl} > 0$

Table 1: Kernel profiles, their smallest admissible growth exponents d_ψ in (7), the corresponding limiting class similarities B_{kl}, and the resulting criteria for distinct test data piles. Here $\tau, \nu > 0$, m is a positive integer, $\ell_k = \gamma_{kk} + \sigma_k^2$, and $\Delta_{kl} = \gamma_{kk} + \gamma_{ll} - 2\gamma_{kl}$.

4 Critical spikes: obstruction and correction

The weak-spike results rely on finite-rank fluctuations becoming negligible in the dimension-scaled kernel geometry. We now ask whether test data piling remains attainable at $\beta = 1$, when the spike variation persists. In linear double data piling theory, ordinary MDP projections need not exhibit test data piling in this regime. Under suitable conditions, projected negative-ridge constructions nevertheless yield second-piling directions within the centered span of the training observations (Chang et al., 2021; Kim et al., 2024). This linear-score setting considered in DDP corresponds to the linear-kernel special case of our framework.

This raises the corresponding question for general kernels: can a suitable function in $\mathcal{S}_X$ eliminate the surviving test variation? This empirical span is also relevant to standard kernel learning. By the generalized representer theorem, any minimizer of an empirical-risk objective with a penalty strictly increasing in the RKHS norm lies in $\mathcal{S}_X$ (Schölkopf et al., 2001; Dinuzzo & Schölkopf, 2012). This covers a broad class of supervised kernel learning methods, including kernel ridge regression, kernel logistic regression, and kernel SVMs.

4.1 A critical-spike obstruction in the Gaussian span

We examine this question for the Gaussian kernel $k(x, z) = \exp\{-\|x - z\|^2/(2p\tau)\}$ with fixed $\tau > 0$. As shown in Table 1, this kernel can produce distinct test piles through residual variance differences even when the limiting mean difference vanishes, unlike the other kernels considered there. We investigate whether this nonlinear separation mechanism can still yield test data piling at $\beta = 1$ through a suitable choice of coefficients in $\mathcal{S}_X$.

For class k, let $V_{k,p} := \mathrm{Cov}\{\kappa(X_{k,0}) \mid \mathcal{F}_p\}$. Since $f_\alpha(x) = \alpha^\top \kappa(x)$ has squared RKHS norm $\alpha^\top G\alpha$, the minimum conditional test variance over unit-RKHS-norm functions in $\mathcal{S}_X$ is

$$\lambda^\star_{k,p} := \min_{\alpha^\top G\alpha=1} \alpha^\top V_{k,p}\alpha = \lambda_{\min}\left(G^{-1/2}V_{k,p}G^{-1/2}\right),$$

whenever $G \succ 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue. The following theorem shows that this minimum cannot vanish asymptotically.

Theorem 2 (An empirical Gaussian-span variance floor)**.** *Suppose Assumption 1 holds with $\beta = 1$ and $r \geq 1$. Assume that the distribution of $\xi_{k,i}$ is absolutely continuous with a density positive on a nonempty open subset of $\mathbb{R}^r$ for every class k.*

For the Gaussian kernel above, as $p \to \infty$ with N and $\tau > 0$ fixed, $\Pr(G \succ 0) \to 1$, and for every class k, $\lambda^\star_{k,p} \xrightarrow{D} \Lambda_k$ where Λ_k is an almost surely positive random variable.

Moreover, on the event $G \succ 0$, let $A = [\alpha_1, \ldots, \alpha_{K-1}] \in \mathbb{R}^{N\times(K-1)}$ be any $\mathcal{F}_p$-measurable coefficient matrix satisfying $A^\top GA = I_{K-1}$. Then,

$$\mathrm{Cov}\{A^\top\kappa(X_{k,0}) \mid \mathcal{F}_p\} = A^\top V_{k,p}A \succeq \lambda^\star_{k,p} I_{K-1}.$$

Hence, no normalized score map in $\mathcal{S}_X$, including KMDP, has vanishing within-class test covariance or induces test piling at the critical spike scale.

The theorem concerns the entire normalized empirical Gaussian span, not merely the choice of the KMDP coefficient. Thus, unlike in the linear double-data-piling construction, changing coefficients within this span cannot suppress the persistent spike variation. To recover test piling in this setting, we next consider changing the input representation before constructing the kernel.

4.2 Restoring test piling by feature correction

The obstruction in the previous subsection arises because the dominant spike variation remains visible to the kernel at the critical scale. A natural remedy, therefore, is to remove this variation before constructing the kernel. Recall that $\mathcal{F}_p$ contains all data used to fit the score map, including those used to estimate the projection. The fresh class-k observation $X_{k,0}$ used to define m_k and C_k remains independent of $\mathcal{F}_p$. The following theorem shows that consistent estimation of the spike projection restores test piling at $\beta = 1$.

Theorem 3 (Test data piling after spike correction)**.** *Suppose Assumptions 1 and 2 hold with $\beta = 1$. Let $\hat{P}_p$ be an $\mathcal{F}_p$-measurable rank-r orthogonal projector such that*

$$\|\hat{P}_p - UU^\top\|_{\mathrm{op}} \xrightarrow{P} 0 \quad \textit{as} \quad p \to \infty. \tag{12}$$

Define $\tilde{k}(x,z) = k\big((I-\hat{P}_p)x, (I-\hat{P}_p)z\big)$ and use this kernel throughout the KMDP construction. Define $\tilde{\ell}_k$, $\tilde{B}$, $\tilde{\eta}_k$, $\tilde{J}$, and $\tilde{t}_k$ by (9)–(11), substituting $\gamma_{kl} - \boldsymbol{a}_k^\top\boldsymbol{a}_l$ for γ_{kl}. Then the corrected train Gram matrix is positive definite with probability tending to one, and all conclusions of Theorem 1 hold for the corrected KMDP map with these substitutions. In particular, $m_k \xrightarrow{P} \tilde{t}_k$ and $\mathrm{tr}\, C_k \xrightarrow{P} 0$ as $p \to \infty$ and the train-pile limit is given by (10) with the same substitutions. For $k \neq l$, the limiting test data piles are distinct if and only if $\tilde{B}_{kk} + \tilde{B}_{ll} - 2\tilde{B}_{kl} > 0$.

The theorem shows that consistent removal of the spike subspace reduces the critical-spike problem to the same limiting geometry as the weak-spike regime. Both the random spike variation and the spike-aligned components of the class means are removed, which leads to the replacement $\gamma_{kl} \mapsto \gamma_{kl} - \boldsymbol{a}_k^\top\boldsymbol{a}_l$. Test data piling is therefore restored whenever the corrected class similarities remain distinct. Under the balanced design $n_k = n$, Corollary 1 also applies to the corrected geometry after replacing the relevant quantities by their corrected counterparts.

Estimating the correction from auxiliary data. We next describe a concrete way to obtain the consistent spike projection required by Theorem 3. A few training samples are not sufficient for consistent estimation of the spike subspace (Jung et al., 2012). A larger auxiliary sample from a related population can instead provide information about the shared dominant variation. For this

auxiliary sample, no labels are required, but the leading covariance subspace must identify the target spike subspace.

Let $W_1, \dots, W_{L_p}$ be i.i.d. auxiliary feature vectors, independent of the training and test observations. A sufficient condition, as $p \to \infty$, is

$$\begin{aligned} \mathrm{Cov}(W_1) &= pU\Theta_S U^\top + R_S, \qquad \Theta_S \succ 0 \text{ fixed}, \quad R_S \succeq 0, \\ L_p &\to \infty, \qquad \|R_S\|_{\mathrm{op}}/p \to 0, \qquad \mathbb{E}\|W_1 - \mathbb{E}W_1\|^4 = O(p^2). \end{aligned} \tag{13}$$

Let $\hat{P}_p$ be the projector onto the leading r eigenvectors of the sample-centered auxiliary covariance matrix. By the eigenspace perturbation bound of Yu et al. (2015), condition (13) is a sufficient condition of $\|\hat{P}_p - UU^\top\|_{\mathrm{op}} = O_{\mathbb{P}}\Big(L_p^{-1/2} + \|R_S\|_{\mathrm{op}}/p\Big)$, which verifies (12). When no suitable auxiliary sample is available, the unlabeled features of the test batch itself may instead be used to estimate the projection. Appendix A presents this transductive construction and establishes piling of the same test batch under projection consistency.

5 Experiments

We conduct simulations to examine finite-dimensional test data piling under different spike strengths, its recovery through correction, kernel-dependent separation, and training-sample-size effects. All observations follow (4) with $K = 3$, $r = 2$, $U = [\boldsymbol{e}_1, \boldsymbol{e}_2]$, and $R_k = \sigma_k^2 I_p$. The vectors $\xi_{k,i}$ and $\varepsilon_{k,i}$ are mutually independent and centered Gaussian, with covariances $\mathrm{diag}(4, 2)$ and I_p, respectively. Class means, residual variance levels, and spike strengths are specified for each experiment. Unless otherwise stated, each class has 5 training observations, p ranges from 128 to 4096, and each configuration uses 200 independent replications. All maps use the RKHS normalization in (3); Gaussian kernels use the fixed bandwidth parameter $\tau = 5$.

5.1 Validation of Theorems 1 and 3

We first examine whether test scores concentrate under weak spikes and whether correction recovers this behavior at the critical scale. To quantify test-score concentration relative to class separation, we use the class-distance-normalized variance (CDNV) of conditional score:

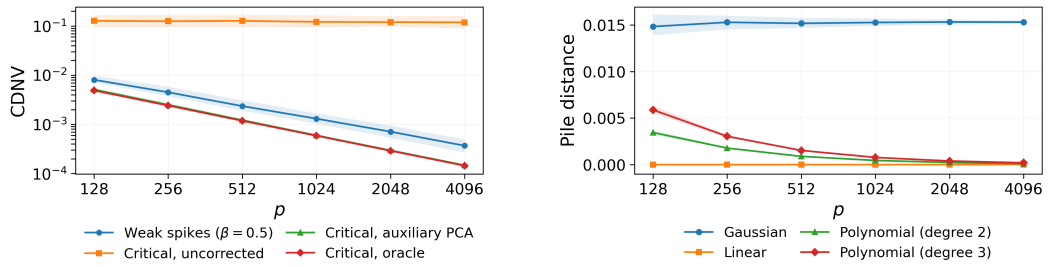
$$\mathcal{V}_{kl} = \frac{\mathrm{tr}\, C_k + \mathrm{tr}\, C_l}{2\|m_k - m_l\|^2}, \qquad m_k \neq m_l.$$

This is the CDNV defined in Galanti et al. (2022), applied to the fitted score distributions rather than the raw input features. A small value indicates that within-class variation is small relative to the squared distance between class centers. We estimate each $\mathcal{V}_{kl}$ from the empirical classwise means and covariance matrices of independent test scores and report the worst-pair value $\max_{k<l} \hat{\mathcal{V}}_{kl}$.

Figure 2(a) compares weak spikes ($\beta = 0.5$), critical spikes ($\beta = 1$), and critical spikes after auxiliary or oracle correction. The class means are set to $\mu_1 = \sqrt{p}\, e_3$, $\mu_2 = \sqrt{p}(-\frac{1}{2}e_3 + \frac{\sqrt{3}}{2}e_4)$, and $\mu_3 = \sqrt{p}(-\frac{1}{2}e_3 - \frac{\sqrt{3}}{2}e_4)$, with $\sigma_k = 1$ for all classes, in accordance with Assumption 1. The auxiliary correction estimates the spike projector from $\lceil 5\sqrt{p} \rceil$ independent observations with covariance $pU\,\mathrm{diag}(4, 2)U^\top + I_p$, whereas the oracle correction uses the true projector $UU^\top$. Each replication uses an independent test sample of size $\lceil 10\sqrt{p} \rceil$. The curves report the median of the maximum pairwise empirical CDNV across replications, with interquartile bands.

The median CDNV decreases with the feature dimension under weak spikes but remains approximately constant under uncorrected critical spikes. Both auxiliary and oracle correction substantially reduce the CDNV, with auxiliary correction closely matching the oracle benchmark. These trends support the predicted test-score concentration under weak spikes and its recovery through feature correction under critical spikes.

Figure 2(b) examines kernel-specific test-pile separation under equal class means, with $(\sigma_1, \sigma_2, \sigma_3) = (1, 1.25, 1.5)$ and $\beta = 1$. Within each replication, we use the same auxiliary projector for the Gaussian, linear, and polynomial kernels with $\nu = 1$. Since coincident limiting test-score means can make the CDNV denominator vanish, we instead report the analytically computed minimum pairwise distance between test piles.

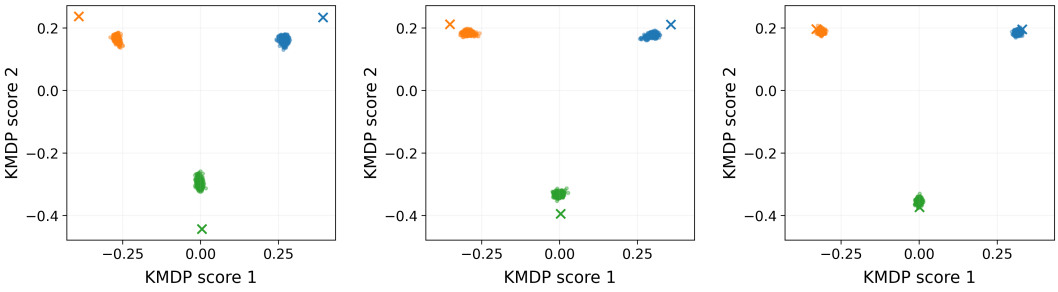


(a) Maximum test score CDNV

(b) Minimum distance between test piles

Figure 2: Test-score concentration and kernel-dependent separation. Curves and shaded bands show medians and interquartile ranges over 200 replications.

For the Gaussian kernel, this distance approaches a positive limit as the feature dimension increases. In contrast, it is identically zero for the linear kernel and decreases toward zero for two polynomial kernels. These trends support the predicted separation of limiting test piles through residual variance differences for the Gaussian kernel, but not for the linear or polynomial kernels, illustrating that within-class concentration alone does not guarantee between-class separation.



(a) $n = 2$.

(b) $n = 5$.

(c) $n = 20$.

Figure 3: Training (crosses) and test (dots) scores at $p = 4096$ as the training size increases. All panels use the same 300 test observations and coordinate scale.

5.2 Validation of Corollary 1

We examine the effect of training sample size in the mean-separated setting of Figure 2a, using $\beta = 0.5$ and the Gaussian kernel. Figure 3 fixes $p = 4096$ and compares $n = 2, 5, 20$ labeled training observations per class. As the training sample size increases, the training piles move closer together while the test piles become more widely separated. At the same time, the corresponding training and test piles within each class approach one another, consistent with the convergence to common class-specific locations predicted by Corollary 1.

Additional experiments in Appendix B examine sensitivity to the correction rank, the size of unlabeled data used for correction, and the Gaussian-kernel bandwidth.

6 Conclusion

Using HDLSS theory, we identify sufficient conditions for constructing a score map from a fixed number of labeled observations such that independent test scores concentrate at distinct class-specific locations. For frozen features satisfying these conditions, a simple downstream map can therefore induce favorable geometry in small-sample settings, such as few-shot learning. We also show that dominant variation can obstruct test piling while suitable feature correction can restore it. By explaining when test piling occurs and why it fails, our analysis motivates broader use of HDLSS asymptotics to understand geometric and statistical phenomena in small-sample learning.

AI USE STATEMENT

We used ChatGPT to polish the writing, proof arguments, and correct grammar. We also used it to debug the experiment source code. We also used it for writing the markdown descriptions for the source code. We do not use the LLM output directly. We have reviewed all AI-assisted work. All AI-assisted writing, proof review, and code debugging were incorporated only after the authors manually checked and revised the corresponding text, mathematical arguments, and code line by line.

REFERENCES

Jeongyoun Ahn and J. S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010.

Jeongyoun Ahn, J. S. Marron, Keith M. Muller, and Yueh-Yun Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766, 2007.

Navid Ardeshir, Clayton Sanford, and Daniel Hsu. Support vector machines and linear regression coincide with very high-dimensional features. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, pp. 4907–4918, Red Hook, NY, USA, 2021. Curran Associates Inc.

Andreas Bär, Neil Houlsby, Mostafa Dehghani, and Manoj Kumar. Frozen feature augmentation for few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16046–16057, 2024.

Woonyoung Chang, Jeongyoun Ahn, and Sungkyu Jung. Double data piling leads to perfect classification. *Electronic Journal of Statistics*, 15(2):6382–6428, 2021.

Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for Hilbert spaces: a necessary and sufficient condition. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 25, pp. 189–196, Red Hook, NY, USA, 2012. Curran Associates, Inc.

Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, and Ziwei Zhu. Robust high-dimensional factor models with applications to statistical machine learning. *Statistical Science*, 36(2):303–327, 2021.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Daniel Gedon, Antonio H. Ribeiro, and Thomas B. Schön. No double descent in principal component regression: A high-dimensional analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 15271–15293. PMLR, 2024.

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3607–3616. PMLR, 2020.

Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (3):427–444, 2005.

X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: proximity to and dynamics on the central path. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Ichiro Hashimoto, Stanislav Volgushev, and Piotr Zwiernik. Universality of benign overfitting in binary linear classification. *arXiv preprint arXiv:2501.10538*, 2025.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 91–99. PMLR, 2021.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Sungkyu Jung and J. S. Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

Sungkyu Jung, Arusharka Sen, and J. S. Marron. Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, 109:190–203, 2012.

Sungkyu Jung, Myung Hee Lee, and Jeongyoun Ahn. On the number of principal components in high dimensions. *Biometrika*, 105(2):389–402, 2018.

Chiraag Kaushik, Andrew D. McRae, Mark Davenport, and Vidya Muthukumar. New equivalences between interpolation and SVMs: Kernels and structured features. *SIAM Journal on Mathematics of Data Science*, 6(3):761–787, 2024.

Taehyun Kim, Woonyoung Chang, Jeongyoun Ahn, and Sungkyu Jung. Double data piling: a high-dimensional solution for asymptotically perfect multi-category classification. *Journal of the Korean Statistical Society*, 53(3):704–737, 2024.

Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Understanding and improving transfer learning of deep models via neural collapse. *Transactions on Machine Learning Research*, 2024.

Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 23103–23123. PMLR, 2023.

Achleshwar Luthra, Yash Salunkhe, and Tomer Galanti. Directional neural collapse explains few-shot transfer in self-supervised learning. In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR, 2026.

J. S. Marron, Michael J. Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28729–28745. PMLR, 2023.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *Proceedings of the 16th European Conference on Computer Vision*, pp. 266–282, Berlin, Heidelberg, 2020. Springer.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5549–5558, 2023.

Kazuyoshi Yata and Makoto Aoshima. Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47(3): 899–921, 2020.

Dongsun Yoon and Sungkyu Jung. Adaptive reference-guided estimation of principal component subspace in high dimensions. *Stat*, 14(3):e70081, 2025.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, pp. 29820–29834, Red Hook, NY, USA, 2021. Curran Associates Inc.

A TRANSDUCTIVE SPIKE-SUBSPACE CORRECTION

When an independent auxiliary sample is unavailable, the unlabeled features of the test batch can be used to estimate the spike projection. Let $(Z_j, Y_j)_{j=1}^{M_p}$ be i.i.d. test observations, independent of the training sample, where $Z_j \in \mathbb{R}^p$ is the feature vector and $Y_j \in \{1, \dots, K\}$ is its true class label. Assume fixed class probabilities $\Pr(Y_j = k) = \pi_k > 0$ and the class-conditional feature distributions in (4). Only the features Z_j, not their labels, are used for fitting.

With r specified, estimate $\hat{P}_p$ from the leading r eigenvectors of the sample-centered test covariance matrix. The sufficient conditions in (13), with $W_1 = Z_1$ and $L_p = M_p$, guarantee (12). More generally, under the feature-model moment conditions and $M_p \to \infty$, consistency follows if the leading r-dimensional eigenspace of the marginal test covariance converges to $\mathrm{span}(U)$ and its eigengap divided by p remains bounded away from zero. Between-class mean variation is therefore allowed, provided that it does not asymptotically change this leading eigenspace or close the gap.

For this setting, $\mathcal{F}_p$ is generated by the labeled training sample and the unlabeled test features used to estimate the correction. Theorem 3 applies to a fresh test observation independent of $\mathcal{F}_p$, but does not directly establish piling of the batch used in projection estimation. Since its scores are fixed conditional on $\mathcal{F}_p$, we describe their concentration through the empirical within-class score geometry.

Theorem 4 (Test data piling under transductive correction)**.** *Suppose Assumptions 1 and 2 hold with $\beta = 1$, and consider the test batch above with $M_p \to \infty$. Let $\hat{P}_p$ be an $\mathcal{F}_p$-measurable rank-r orthogonal projector satisfying (12). Construct the corrected kernel and KMDP map as in Theorem 3.*

Let $M_{k,p} = \sum_{j=1}^{M_p} \mathbf{1}\{Y_j = k\}$. For every class k, $\Pr(M_{k,p} > 0) \to 1$, and on this event,

$$\frac{1}{M_{k,p}} \sum_{j:Y_j=k} \|g(Z_j) - \tilde{t}_k\|^2 \xrightarrow{\mathbb{P}} 0 \tag{14}$$

as $p \to \infty$ with fixed training sample sizes. If $\tilde{B}_{kk} + \tilde{B}_{ll} - 2\tilde{B}_{kl} > 0$ holds for every class pair, the limiting test piles are pairwise distinct.

The theorem establishes classwise piling for the same test batch used to stimate the projection, without sample splitting. Equation (14) implies that the empirical class means converge to $\tilde{t}_k$ and the within-class score variation vanishes. The true test labels are used only to describe these evaluation quantities, not to construct the correction. The conclusion concerns average squared distance within each class, rather than uniform convergence of all scores in a growing batch.

Proof of Theorem 4. All limits are as $p \to \infty$, with the training sample sizes fixed. The test pairs are i.i.d. with fixed class probabilities $\pi_k > 0$ and are independent of the training sample, as specified above. Write k_p and $\tilde{k}_p$ for the original and corrected kernels at dimension p, and put $P = UU^\top$. For arbitrary $x, z \in \mathbb{R}^p$, orthogonality of the projectors gives

$$\frac{\left|\langle (I - \hat{P}_p)x, (I - \hat{P}_p)z\rangle - \langle (I - P)x, (I - P)z\rangle\right|}{p} \le \|\hat{P}_p - P\|_{\mathrm{op}} \frac{\|x\|\,\|z\|}{p}.$$

This deterministic bound does not require independence between the estimated projector and the observations to which it is applied.

Let $\tilde{G}$ and $\tilde{\kappa}$ denote the corrected training Gram matrix and kernel evaluation vector. Since the training sample is fixed in size and $\|X_i\|/\sqrt{p} = O_{\mathbb{P}}(1)$, the oracle comparison in the proof of Theorem 3 remains valid and gives

$$\tilde{G} \xrightarrow{\mathbb{P}} \tilde{G}_0 := E\tilde{B}E^\top + \mathrm{diag}_{i \le N}(\tilde{\eta}_{y_i}) \succ 0.$$

Put $D_n = \mathrm{diag}(n_1, \dots, n_K)$. The same block-system inversion and normalization then yield $A \xrightarrow{\mathbb{P}} A_0 := ED_n^{-1}T\tilde{J}^{-1/2}$. Neither conclusion requires the projector to be independent of the training sample or the test batch. On the exceptional event $\tilde{G} \not\succ 0$, whose probability tends to zero, define $A = 0$ and $g = 0$.

Introduce an auxiliary index I_p, uniform on $\{1, \dots, M_p\}$ and independent of all data, solely for this proof. The selected pair (Z_{I_p}, Y_{I_p}) has the prescribed mixture law and is independent of the training sample, although it need not be independent of $\hat{P}_p$ or g. Conditional on $Y_{I_p} = k$, the oracle-projected observation $(I - P)Z_{I_p}$ has the normalized norm and test-to-training inner-product limits established in the proof of Theorem 3. The displayed perturbation bound and (12) transfer these limits to the estimated projection, since $\|Z_{I_p}\|/\sqrt{p} = O_{\mathbb{P}}(1)$. Conditioning on $Y_{I_p} = k$ preserves these probability limits because π_k is fixed and positive. Continuity of the kernel profile therefore implies

$$\|\tilde{\kappa}(Z_{I_p}) - E\tilde{B}\boldsymbol{e}_{Y_{I_p}}\| \xrightarrow{\mathbb{P}} 0, \qquad \|g(Z_{I_p}) - \tilde{t}_{Y_{I_p}}\|^2 \xrightarrow{\mathbb{P}} 0,$$

where the second conclusion uses $A \to A_0$ and $A_0^\top E\tilde{B}\boldsymbol{e}_k = \tilde{t}_k$.

We next strengthen the latter convergence to convergence of expectations. For each realized projector, the corrected kernel is positive semidefinite. On $\tilde{G} \succ 0$, the coordinates of g are orthonormal in its RKHS by Lemma 1. Bessel's inequality, (7), and $\|(I - \hat{P}_p)x\| \le \|x\|$ give the pointwise bound

$$\|g(x)\|^2 \le \tilde{k}_p(x, x) \le L\left\{1 + \left(\frac{\|x\|^2}{p}\right)^{d_\psi}\right\}.$$

This also holds for the zero-score extension, and in particular for observations used to fit the projector. For $d_\psi > 2$, (8) makes the right-hand side uniformly integrable under every class distribution. For $0 \le d_\psi \le 2$, write $v_{k,0} = R_k^{1/2}\varepsilon_{k,0}$ and use

$$\frac{\|X_{k,0}\|^4}{p^2} \le 27\left\{\frac{\|\mu_k\|^4}{p^2} + \|\xi_{k,0}\|^4 + \frac{\|v_{k,0}\|^4}{p^2}\right\}.$$

The first term is bounded and deterministic, and the second has a fixed distribution with finite expectation. The residual assumptions give $\|v_{k,0}\|^2/p \xrightarrow{L^2} \sigma_k^2$, and hence $\|v_{k,0}\|^4/p^2 \xrightarrow{L^1} \sigma_k^4$. Thus $\{\|X_{k,0}\|^4/p^2\}_p$ is uniformly integrable, which also controls $(\|X_{k,0}\|^2/p)^{d_\psi}$ for $d_\psi \le 2$. Since the law of Z_{I_p} is a fixed finite mixture of these class laws and the $\tilde{t}_k$ are fixed, the squared score errors are uniformly integrable. Consequently, $\mathbb{E}\|g(Z_{I_p}) - \tilde{t}_{Y_{I_p}}\|^2 \to 0$.

By uniformity and independence of I_p, averaging over this index gives

$$\mathbb{E}\left[\frac{1}{M_p}\sum_{j=1}^{M_p}\|g(Z_j) - \tilde{t}_{Y_j}\|^2\right] = \mathbb{E}\|g(Z_{I_p}) - \tilde{t}_{Y_{I_p}}\|^2 \longrightarrow 0.$$

Hence the full-batch average converges to zero in probability by Markov's inequality. Also, the binomial variance bound gives $M_{k,p}/M_p \xrightarrow{\mathbb{P}} \pi_k$, and $\Pr(M_{k,p} = 0) = (1-\pi_k)^{M_p} \to 0$. On $M_{k,p} > 0$, the class-k average in (14) is at most $M_p/M_{k,p}$ times the full-batch average. The multiplier converges to $1/\pi_k$ in probability on these events, proving (14). Since K is fixed, the conclusions hold simultaneously for all classes. Finally, the deterministic limiting locations are those of Theorem 3, whose separation criterion implies that they are pairwise distinct under the stated condition. □

B ADDITIONAL EXPERIMENTS

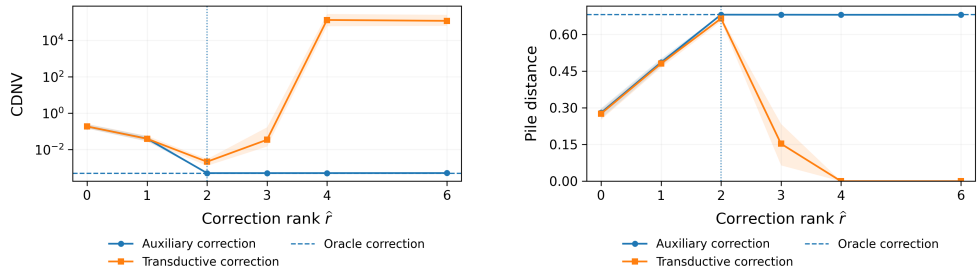
We examine sensitivity to the correction rank, the amount and source of unlabeled data, and the Gaussian-kernel bandwidth. All studies use $K = 3$, $n_k = 5$, $p = 1024$, and $r = 2$ with $\beta = 1$. Independently for each class-k observation, $k \in \{1, 2, 3\}$, we generate

$$X_k = \sqrt{p}\, V b_k + \sqrt{p}\, U\Theta^{1/2}\xi + \sigma_k\varepsilon, \qquad \xi \sim N_2(0, I_2), \quad \varepsilon \sim N_p(\mathbf{0}_p, I_p),$$

with ξ and ε independent, $U = [\boldsymbol{e}_1, \boldsymbol{e}_2]$, $V = [\boldsymbol{e}_3, \boldsymbol{e}_4]$, and $\Theta = \mathrm{diag}(4, 2)$. In the *mean-difference* setting, the $b_k \in \mathbb{R}^2$ form an equilateral triangle of unit circumradius centered at the origin and $\sigma_k = 1$. In the *variance-difference* setting, $b_k = 0$ and $(\sigma_1, \sigma_2, \sigma_3) = (1, 1.25, 1.5)$. Unlabeled target batches are i.i.d. draws from the equal-prior mixture of these class distributions. Auxiliary observations are i.i.d. $N_p(\mathbf{0}_p, pU\Theta U^\top + I_p)$, independently of all target observations. The kernel is $k(x, z) = \exp\{-\|x - z\|^2/(2p\tau)\}$, with bandwidth parameter $\tau = 5$ unless varied.

For a specified correction rank, let $\hat{P}$ be the orthogonal projector onto that number of leading eigenvectors of the sample-centered covariance of the unlabeled correction features. Both training and test features are transformed by $I_p - \hat{P}$, and KMDP is refitted whenever the correction rank or bandwidth changes. No correction uses $\hat{P} = 0$, whereas oracle correction uses $\hat{P} = UU^\top$. Test labels are used only for post-fit evaluation.

Each study uses 200 independent replications. Within each setting and replication, methods share the training set, which is held fixed across parameter values. Independent evaluation uses the same 1000 test observations per class across those values; transductive evaluation uses the correction-fitting test batch, as specified below. We report the worst-case empirical pairwise CDNV of the test score and the worst-case pairwise test pile distance. Lines and shading in each figure show medians and interquartile ranges of these per-replication metrics, respectively.



(a) Maximum test score CDNV

(b) Minimum distance between test piles

Figure 4: Correction-rank sensitivity in the mean-difference setting. (a) reports maximum pairwise empirical CDNV; (b) minimum pairwise distance between empirical test-score centers across $\hat{r}$. The vertical line marks $r = 2$.

B.1 Effect of correction rank

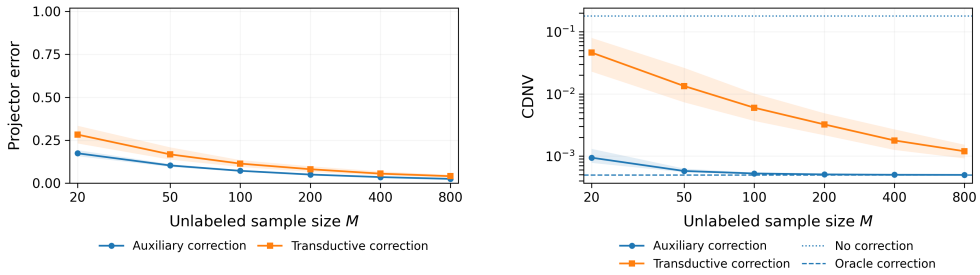
In the mean-difference setting, let $\hat{r}$ denote the prescribed number of directions removed by correction. We vary $\hat{r} \in \{0, 1, 2, 3, 4, 6\}$, with zero denoting no correction and the true rank $r = 2$ unchanged. Auxiliary correction uses 384 independent auxiliary observations and the fixed independent evaluation set described above. Transductive correction uses one fixed i.i.d. target test batch of 384 total observations for both fitting and post-fit evaluation. Each method reuses its own evaluation observations across ranks. At $\hat{r} = 0$, the fitted maps coincide, but empirical metrics may differ because the evaluation sets differ.

Both corrections improve markedly at $\hat{r} = r$ relative to no correction (Figure 4). Their behavior under rank overestimation differs: auxiliary correction remains stable, whereas transductive correction loses separation and develops much larger CDNV. The target mixture contains between-class variation absent from the zero-mean auxiliary population. Its next two population eigendirections after the spikes span the mean-signal plane V; the recorded retained-mean fractions support the interpretation that overspecified transductive correction removes this signal. Thus tolerance to rank overestimation depends on the source of the correction data and on preserving class signal, not only on removing spike variation.

B.2 Effect of source data size

In the mean-difference setting, let M denote the total number of unlabeled observations used to estimate the correction. We vary $M \in \{20, 50, 100, 200, 400, 800\}$ while fixing $\hat{r} = 2$. Auxiliary correction uses M independent auxiliary observations and a fixed independent evaluation set. Transductive correction fits on M target test features and evaluates those same observations, with labels used only afterward. Test batches are sampled from the target mixture without fixing class counts or resampling based on the resulting class composition.

Both routes estimate the spike subspace more accurately as M increases, accompanied by smaller test CDNV (Figure 5). Auxiliary correction approaches the oracle reference sooner in this experiment; transductive correction improves steadily but retains larger CDNV over the displayed range. The target mixture includes between-class variation, although its leading two population eigendirections still span U. Note that these curves do not provide a fair comparison between two correction methods: only the correction sample size varies for auxiliary correction, whereas fitting and evaluation sizes both vary for transductive correction.



(a) Projector estimation error

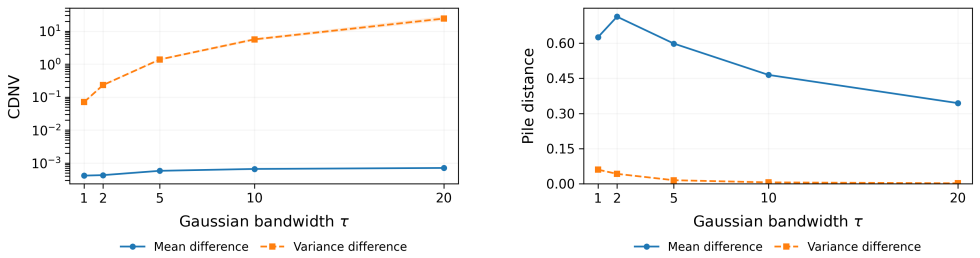
(b) Maximum test score CDNV

Figure 5: Effect of unlabeled sample size. (a) reports the estimation error of projection; (b) reports maximum pairwise empirical CDNV across M. The horizontal axes are logarithmic in M.

B.3 Effect of the bandwidth parameter.

We fix auxiliary correction at $\hat{r} = 2$ using 384 auxiliary observations and vary the Gaussian-kernel bandwidth parameter $\tau \in \{1, 2, 5, 10, 20\}$. We use the same auxiliary sample and fitted projector for both settings within each replication; training and test samples are generated separately for each setting and held fixed across bandwidths. Each panel of Figure 6 overlays the mean-difference and variance-difference settings. Mean-difference CDNV remains small throughout the sweep; its pile distance first increases and then decreases. In contrast, variance-difference CDNV increases strongly while its pile distance decreases.

In both settings, larger bandwidths may attenuate class-discriminating distance contrasts, weakening test-pile separation relative to within-class variation. Since the CDNV is scale-invariant, with both within-class variance and squared between-class distance scaling by the same factor, its increase cannot be explained by a common rescaling of the scores.



(a) Maximum test score CDNV

(b) Minimum distance between test piles

Figure 6: Gaussian-bandwidth sensitivity under auxiliary correction. Both panels contain the mean-difference and variance-difference settings. (a) reports maximum pairwise empirical CDNV; (b) reports minimum pairwise distance between empirical test-score centers across τ.

C TECHNICAL DETAILS

C.1 PROOF OF LEMMA 1

Proof. Let $\mathcal{N}_W := \{f \in \mathcal{H} : f(X_{k,i}) = f(X_{k,j}) \text{ for all } k = 1, \dots, K,\ 1 \le i, j \le n_k\}$ be the subspace of scalar functions with classwise constant support values. By the reproducing property, every $f \in \mathcal{S}_X^{\perp}$ vanishes on the support, so $\mathcal{S}_X^{\perp} \subset \mathcal{N}_W$, then, $\mathcal{N}_W = (\mathcal{N}_W \cap \mathcal{S}_X) \oplus (\mathcal{N}_W \cap \mathcal{S}_X^{\perp}) = (\mathcal{N}_W \cap \mathcal{S}_X) \oplus \mathcal{S}_X^{\perp}$. For $f_a = \sum_{i=1}^{N} a_i k(X_i, \cdot)$, its vector of training evaluations is Ga. Since $G \succ 0$, the map $f_a \mapsto Ga$ is a linear isomorphism from $\mathcal{S}_X$ onto $\mathbb{R}^N$.

Classwise constant evaluation vectors are exactly those in $\mathrm{col}(E)$, so $\mathcal{N}_W \cap \mathcal{S}_X = \{f_a : Ga \in \mathrm{col}(E)\}$ has dimension $\mathrm{rank}(E) = K$.

Define $f_0 := f_{G^{-1}\mathbf{1}_N}$, so $f_0(X_i) = 1$ for every training observation. Within $\mathcal{N}_W \cap \mathcal{S}_X$, its RKHS-orthogonal complement is $\mathcal{W}_0 = \{f_a : Ga \in \mathrm{col}(E),\ \mathbf{1}_N^{\top} a = 0\}$, because $\langle f_a, f_0 \rangle_{\mathcal{H}} = a^{\top} G(G^{-1}\mathbf{1}_N) = \mathbf{1}_N^{\top} a$.

Thus, $\dim \mathcal{W}_0 = K - 1$ and $\mathcal{N}_W = \mathcal{W}_0 \oplus \mathrm{span}\{f_0\} \oplus \mathcal{S}_X^{\perp}$.

Here f_0 is constant on the training set; it need not be constant on the entire input space.

For scalar f, let $\bar{f}_k$ and $\bar{f}$ be its class and pooled averages on the training set, and define $b(f) := \sum_{k=1}^{K} n_k(\bar{f}_k - \bar{f})^2$.

On $\mathcal{N}_W$, this quadratic form depends only on the $\mathcal{W}_0$ component: the f_0 component shifts all train values equally, and the $\mathcal{S}_X^{\perp}$ component vanishes there. It is positive definite on $\mathcal{W}_0$. Indeed, $b(f_a) = 0$ there implies $Ga = t\mathbf{1}_N$ for some t, whence $f_a = t f_0$; orthogonality to f_0 forces $t = 0$.

Choose an RKHS-orthonormal eigenbasis $u_1, \dots, u_{K-1}$ for this positive quadratic form on $\mathcal{W}_0$, with eigenvalues $\lambda_j > 0$. For any feasible tuple $(g_1, \dots, g_{K-1})$ in (1), Bessel's inequality gives

$$\sum_{k=1}^{K} n_k \|\bar{g}_k - \bar{g}\|^2 = \sum_{\ell=1}^{K-1} \lambda_\ell \sum_{j=1}^{K-1} |\langle g_j, u_\ell \rangle_{\mathcal{H}}|^2 \le \sum_{\ell=1}^{K-1} \lambda_\ell.$$

Equality is attained by any orthonormal basis of $\mathcal{W}_0$. Conversely, equality requires every u_ℓ to belong to $\mathrm{span}\{g_1, \dots, g_{K-1}\}$, since all λ_ℓ are positive. Both spaces have dimension $K - 1$, so this span must equal $\mathcal{W}_0$. This proves existence and uniqueness of the maximizing subspace in the entire RKHS, and shows that every optimizer lies in $\mathcal{S}_X$.

It remains to identify the explicit basis. The block system (2) is invertible because $G \succ 0$ and $\mathbf{1}_N^{\top} G^{-1} \mathbf{1}_N > 0$. It gives

$$GQ = ET - \mathbf{1}_N h^{\top}, \qquad \mathbf{1}_N^{\top} Q = 0.$$

Hence every $f_{\boldsymbol{q}_j}$ belongs to $\mathcal{W}_0$. If $Qv = 0$, then $ETv = \mathbf{1}_N h^{\top} v$, and, because every class is represented, Tv is constant across classes. As $Tv \perp \mathbf{1}_K$ and $T^{\top}T = I_{K-1}$, this implies $v = 0$. Thus $\mathrm{rank}(Q) = K - 1$ and the functions $f_{\boldsymbol{q}_j}$ span $\mathcal{W}_0$. Finally, $Q^{\top} G Q \succ 0$ and

$$A^{\top} G A = (Q^{\top} G Q)^{-1/2} (Q^{\top} G Q)(Q^{\top} G Q)^{-1/2} = I_{K-1}.$$

Thus, the coordinate functions defined by A form an RKHS-orthonormal basis of $\mathcal{W}_0$, so $g(x) = A^{\top}\kappa(x)$ solves (1). □

C.2 PROOF OF THEOREM 1

Proof. All limits are as $p \to \infty$, with $K, r, n_1, \dots, n_K$ and T fixed. We first establish the normalized feature limits and then identify the limiting KMDP coefficients and conditional test moments.

Write $v_{k,i} = R_k^{1/2} \varepsilon_{k,i}$. By (5)–(6), $\|\mu_k\|^2 = O(p)$, $\mathrm{tr}(R_k) = O(p)$, and $\|R_k\|_{\mathrm{op}}/p \to 0$. Expanding the residual quadratic form and using independent coordinates with uniformly bounded fourth moments gives, for a constant C independent of p,

$$\mathbb{E}\big(\|v_{k,i}\|^2 - \mathrm{tr}(R_k)\big)^2 \le C\,\mathrm{tr}(R_k^2) = o(p^2).$$

For distinct observations, $\mathbb{E}(v_{k,i}^\top v_{l,j}/p)^2 = \operatorname{tr}(R_k R_l)/p^2 \to 0$ by the trace Cauchy–Schwarz inequality. Also, $\mathbb{E}(\mu_k^\top v_{l,j}/p)^2 \le \|\mu_k\|^2 \|R_l\|_{\mathrm{op}}/p^2 \to 0$. Moreover, $\mathbb{E}\|(\mu_k + v_{k,i})/\sqrt{p}\|^4 = O(1)$, whereas

$$\mathbb{E}\left\|p^{(\beta-1)/2} U \xi_{k,i}\right\|^4 = p^{2\beta-2}\mathbb{E}\|\xi_{k,i}\|^4 \longrightarrow 0.$$

Cauchy–Schwarz therefore makes every normalized product involving at least one spike term negligible in L^2. Expanding the norms and inner products yields

$$\frac{\|X_{k,i}\|^2}{p} \xrightarrow{L^2} \ell_k, \qquad \frac{X_{k,i}^\top X_{l,j}}{p} \xrightarrow{L^2} \gamma_{kl} \quad \text{for } (k,i) \neq (l,j). \tag{15}$$

These statements include the independent test observations.

Continuity of ψ now gives $G \xrightarrow{\mathbb{P}} G_0$ and $\kappa(X_{k,0}) \xrightarrow{\mathbb{P}} EB\boldsymbol{e}_k$, where

$$G_0 = EBE^\top + \operatorname{diag}(\eta_{y_1}, \dots, \eta_{y_N}).$$

Since $0 \le \gamma_{kk} < \ell_k$, the first condition in (7) implies $\eta_k > 0$. To verify $B \succeq 0$, consider, solely for this argument, m independent observations per class, with m fixed. The matrix of kernel similarities averaged over all pairs in each class block is positive semidefinite, since it is the Gram matrix of the class-average kernel sections. Its limit as $p \to \infty$ is $B + m^{-1}\operatorname{diag}(\eta_1, \dots, \eta_K)$, which is therefore positive semidefinite. Letting $m \to \infty$ in this deterministic matrix expression proves $B \succeq 0$. Consequently, $G_0 \succeq (\min_k \eta_k) I_N \succ 0$ and $J \succ 0$. In particular, $\Pr(G \succ 0) \to 1$. On the exceptional event $G \not\succ 0$, define $A = 0$ and $g = 0$; this convention does not affect the asserted probability limits.

Put $D_n = \operatorname{diag}(n_1, \dots, n_K)$ and $H = B + \operatorname{diag}(\eta_1/n_1, \dots, \eta_K/n_K)$, so $J = T^\top H T$. Since $E^\top E = D_n$ and $\mathbf{1}_N = E\mathbf{1}_K$,

$$G_0 E D_n^{-1} T = EHT, \qquad \mathbf{1}_N^\top E D_n^{-1} T = 0.$$

Let $Q_0 = E D_n^{-1} T J^{-1}$. The identity $T^\top(HTJ^{-1} - T) = 0$ shows that each column of $HTJ^{-1} - T$ is constant across classes. Hence there is a vector h_0 such that $G_0 Q_0 + \mathbf{1}_N h_0^\top = ET$, while $\mathbf{1}_N^\top Q_0 = 0$. The limiting block system is invertible because $G_0 \succ 0$ and $\mathbf{1}_N^\top G_0^{-1} \mathbf{1}_N > 0$. Thus continuity of its inverse and of positive-definite matrix square roots, together with $Q_0^\top G_0 Q_0 = J^{-1}$, gives

$$Q \xrightarrow{\mathbb{P}} Q_0, \qquad A \xrightarrow{\mathbb{P}} A_0 := E D_n^{-1} T J^{-1/2}. \tag{16}$$

The first convergence is understood on the events where the block system is defined, with any fixed extension otherwise.

The ith row of GA is $g(X_i)^\top$. Since $GA \xrightarrow{\mathbb{P}} G_0 A_0 = EHTJ^{-1/2}$, averaging the rows belonging to class k yields $\bar{g}_k \xrightarrow{\mathbb{P}} J^{-1/2} T^\top H \boldsymbol{e}_k = t_k^{\mathrm{tr}}$, as asserted in (10). For an independent test observation,

$$g(X_{k,0}) = A^\top \kappa(X_{k,0}) \xrightarrow{\mathbb{P}} A_0^\top EB\boldsymbol{e}_k = J^{-1/2} T^\top B \boldsymbol{e}_k = t_k. \tag{17}$$

To pass to the conditional moments, more than probability convergence is needed. On $G \succ 0$, the coordinate functions of g are RKHS-orthonormal by Lemma 1. Bessel's inequality and (7) therefore give the pointwise bound

$$\|g(x)\|^2 \le k_p(x,x) \le L\left\{1 + \left(\frac{\|x\|^2}{p}\right)^{d_\psi}\right\}. \tag{18}$$

The same bound holds for the zero-score extension. For $0 \le d_\psi \le 2$, (15) implies uniform integrability of $\{(\|X_{k,0}\|^2/p)^{d_\psi}\}_p$: indeed, $(\|X_{k,0}\|^2/p)^2 \to \ell_k^2$ in L^1, and $x^{d_\psi} \le 1 + x^2$ for $x \ge 0$. For $d_\psi > 2$, (8) gives a uniformly bounded moment of order $1 + \epsilon/(2d_\psi) > 1$ for this family. Thus (18) makes $\{\|g(X_{k,0}) - t_k\|^2\}_p$ uniformly integrable, and (17) implies $\mathbb{E}\|g(X_{k,0}) - t_k\|^2 \to 0$. By the definitions of m_k and C_k,

$$\|m_k - t_k\|^2 + \operatorname{tr} C_k = \mathbb{E}\left[\|g(X_{k,0}) - t_k\|^2 \mid \mathcal{F}_p\right].$$

The nonnegative right-hand side has expectation tending to zero, so Markov's inequality gives $m_k \xrightarrow{\mathbb{P}} t_k$ and $\operatorname{tr} C_k \xrightarrow{\mathbb{P}} 0$. Since $C_k \succeq 0$, this also gives $\|C_k\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0$, proving (11).

Finally, fix $k \neq l$ and put $q = \boldsymbol{e}_k - \boldsymbol{e}_l$. Because J is invertible, $t_k = t_l$ if and only if $T^\top B q = 0$. In that case $Bq \in \operatorname{span}\{\mathbf{1}_K\}$, and $q \perp \mathbf{1}_K$ gives $q^\top B q = 0$. Conversely, $q^\top B q = 0$ and $B \succeq 0$ imply $Bq = 0$, hence $t_k = t_l$. As $q^\top B q = B_{kk} + B_{ll} - 2B_{kl} \ge 0$, the stated necessary and sufficient condition for distinct test piles follows. □

C.3 Proof of Corollary 1

Proof. Throughout the comparison across n, the limiting quantities $B, \eta_1, \ldots, \eta_K$, and the contrast matrix T are fixed. Let

$$D_\eta = \mathrm{diag}(\eta_1, \ldots, \eta_K), \qquad M_B = T^\top B T, \qquad M_\eta = T^\top D_\eta T.$$

Under the balanced design $n_k = n$, $J_n = M_B + M_\eta/n$.

Since $B \succeq 0$, all $\eta_k > 0$, and T has orthonormal columns, $M_B \succeq 0$ and $M_\eta \succ 0$. Hence

$$J_n \succeq \frac{M_\eta}{n} \succ 0, \qquad J_n^{-1} \preceq n M_\eta^{-1}.$$

By (10),

$$t_k^{\mathrm{tr}}(n) - t_k(n) = \frac{\eta_k}{n} J_n^{-1/2} T^\top \boldsymbol{e}_k.$$

Put $V = D_\eta^{1/2} T$. Then $M_\eta = V^\top V$ and $T^\top \boldsymbol{e}_k = \eta_k^{-1/2} V^\top \boldsymbol{e}_k$, so

$$\begin{aligned} \|t_k^{\mathrm{tr}}(n) - t_k(n)\|^2 &= \frac{\eta_k^2}{n^2} (T^\top \boldsymbol{e}_k)^\top J_n^{-1} (T^\top \boldsymbol{e}_k) \le \frac{\eta_k^2}{n} (T^\top \boldsymbol{e}_k)^\top M_\eta^{-1} (T^\top \boldsymbol{e}_k) \\ &= \frac{\eta_k}{n} \boldsymbol{e}_k^\top V (V^\top V)^{-1} V^\top \boldsymbol{e}_k \le \frac{\eta_k}{n}, \end{aligned}$$

because $V(V^\top V)^{-1} V^\top$ is an orthogonal projector. This proves the displacement bound and implies $t_k^{\mathrm{tr}}(n) - t_k(n) \to 0$ as $n \to \infty$.

Next fix $k \neq \ell$ and set $q =: \boldsymbol{e}_k - \boldsymbol{e}_\ell$ and $w = T^\top q$.

Since $q \perp \mathbf{1}_K$ and $TT^\top = I_K - \mathbf{1}_K \mathbf{1}_K^\top / K$, we have $q = Tw$. Using (11) and (10),

$$t_k^{\mathrm{tr}}(n) - t_\ell^{\mathrm{tr}}(n) = J_n^{-1/2} T^\top \left(B + \frac{D_\eta}{n} \right) T w = J_n^{1/2} w.$$

Consequently,

$$\|t_k^{\mathrm{tr}}(n) - t_\ell^{\mathrm{tr}}(n)\|^2 = w^\top J_n w = q^\top B q + \frac{1}{n} q^\top D_\eta q = B_{kk} + B_{\ell\ell} - 2B_{k\ell} + \frac{\eta_k + \eta_\ell}{n}.$$

Since $\eta_k + \eta_\ell > 0$, this quantity decreases strictly to $B_{kk} + B_{\ell\ell} - 2B_{k\ell}$.

For the test piles,

$$t_k(n) - t_\ell(n) = J_n^{-1/2} M_B w,$$

and therefore

$$\|t_k(n) - t_\ell(n)\|^2 = w^\top M_B J_n^{-1} M_B w. \tag{19}$$

If $n_1 < n_2$, then $J_{n_1} \succeq J_{n_2}$ and hence $J_{n_1}^{-1} \preceq J_{n_2}^{-1}$. Thus (19) is nondecreasing in n.

To identify its limit, let

$$R = M_\eta^{-1/2} M_B M_\eta^{-1/2} \succeq 0.$$

Then

$$M_B J_n^{-1} M_B = M_\eta^{1/2} R \left(R + \frac{1}{n} I_{K-1} \right)^{-1} R M_\eta^{1/2}.$$

For every eigenvalue $\lambda \geq 0$ of R, $\lambda^2/(\lambda + 1/n) \to \lambda$. The spectral decomposition therefore gives

$$R \left(R + \frac{1}{n} I_{K-1} \right)^{-1} R \longrightarrow R, \qquad M_B J_n^{-1} M_B \longrightarrow M_B.$$

It follows that

$$\|t_k(n) - t_\ell(n)\|^2 \longrightarrow w^\top M_B w = q^\top B q = B_{kk} + B_{\ell\ell} - 2B_{k\ell}.$$

Together with monotonicity, this proves the claimed nondecreasing convergence. □

C.4 Proof of Theorem 2

Proof. All limits are as $p \to \infty$, with the training sample sizes and $\tau > 0$ fixed. No coupling of latent factors across dimensions is assumed. Index the training observations as $(X_i, y_i)_{i=1}^N$, and write ξ_i for the factor associated with X_i.

The matrix $(\gamma_{kl} - \boldsymbol{a}_k^\top \boldsymbol{a}_l)_{k,l=1}^K$ is positive semidefinite, being the limit of the Gram matrix $(\mu_k^\top (I - UU^\top)\mu_l/p)_{k,l=1}^K$. Choose deterministic $b_1, \dots, b_K \in \mathbb{R}^K$ such that $b_k^\top b_l = \gamma_{kl} - \boldsymbol{a}_k^\top \boldsymbol{a}_l$. Put $v_{k,i} = R_k^{1/2} \varepsilon_{k,i}$. The residual assumptions imply, for distinct observations,

$$\mathbb{E}\big(\|v_{k,i}\|^2 - \operatorname{tr} R_k\big)^2 \le C \operatorname{tr}(R_k^2) = o(p^2), \qquad \mathbb{E}(v_{k,i}^\top v_{l,j})^2 = \operatorname{tr}(R_k R_l) = o(p^2).$$

The first estimate follows from independent coordinates with bounded fourth moments, and the second uses trace Cauchy–Schwarz. Furthermore, $\|\mu_k\|^2 = O(p)$ and $\|R_l\|_{\mathrm{op}}/p \to 0$ give $\mathbb{E}|\mu_k^\top v_{l,j}/p|^2 \to 0$ and $\mathbb{E}\|U^\top v_{l,j}/\sqrt{p}\|^2 \to 0$. Expanding the squared distances and using the tightness of the factors therefore yields

$$\frac{\|X_{k,i} - X_{l,j}\|^2}{p} = \|\boldsymbol{a}_k + \xi_{k,i} - \boldsymbol{a}_l - \xi_{l,j}\|^2 + \|b_k - b_l\|^2 + \sigma_k^2 + \sigma_l^2 + o_{\mathbb{P}}(1) \tag{20}$$

whenever $(k, i) \neq (l, j)$, including pairs involving a fresh test observation.

Define the fitting-measurable coordinates $z_{i,p} = U^\top X_i/\sqrt{p}$ and $z_p = (z_{1,p}, \dots, z_{N,p})$. Then $z_{i,p} = \boldsymbol{a}_{y_i} + \xi_i + o_{\mathbb{P}}(1)$. Since the joint distribution of the independent training factors does not depend on p, we have $z_p \xrightarrow{d} Z = (Z_1, \dots, Z_N)$, where $Z_i = \boldsymbol{a}_{y_i} + \Xi_i$ and the independent Ξ_i have the corresponding class-specific factor distributions. For $z = (z_1, \dots, z_N) \in (\mathbb{R}^r)^N$, put

$$H_{ij}(z) = \exp\left\{-\frac{\|z_i - z_j\|^2 + \|b_{y_i} - b_{y_j}\|^2}{2\tau}\right\}, \qquad D = \operatorname{diag}_{i \le N}\left(e^{-\sigma_{y_i}^2/(2\tau)}\right),$$

and define $G_0(z) = DH(z)D + I_N - D^2$. Equation (20) and $G_{ii} = 1$ imply $\|G - G_0(z_p)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0$. The matrix $H(z)$ is a Gaussian Gram matrix on the points (z_i, b_{y_i}), so $H(z) \succeq 0$ and

$$G_0(z) \succeq I_N - D^2 \succeq \min_{1 \le k \le K}\{1 - e^{-\sigma_k^2/\tau}\} I_N \succ 0.$$

The bound is uniform in z, and hence $\Pr(G \succ 0) \to 1$.

For class k, let ν_k be its factor distribution and define $\varphi_k(u; z) = (\varphi_{k,i}(u; z))_{i=1}^N$ by

$$\varphi_{k,i}(u; z) = \exp\left\{-\frac{\|\boldsymbol{a}_k + u - z_i\|^2 + \|b_k - b_{y_i}\|^2 + \sigma_k^2 + \sigma_{y_i}^2}{2\tau}\right\}. \tag{21}$$

Put $V_{k,0}(z) = \operatorname{Cov}_{u \sim \nu_k}\{\varphi_k(u; z)\}$, where z is held fixed in the covariance. Represent the fresh test observation using a factor and residual independent of $\mathcal{F}_p$. By (20), $\|\kappa(X_{k,0}) - \varphi_k(\xi_{k,0}; z_p)\| \xrightarrow{\mathbb{P}} 0$. Both vectors have coordinates in $[0, 1]$, so their difference also converges to zero in L^1. Thus $\mathbb{E}[\|\kappa(X_{k,0}) - \varphi_k(\xi_{k,0}; z_p)\| \mid \mathcal{F}_p] \xrightarrow{\mathbb{P}} 0$ by Markov's inequality. Since z_p is $\mathcal{F}_p$-measurable and $\xi_{k,0}$ is independent of $\mathcal{F}_p$, the conditional covariance of $\varphi_k(\xi_{k,0}; z_p)$ equals $V_{k,0}(z_p)$. The same conditional error controls both first and second moments:

$$\|V_{k,p} - V_{k,0}(z_p)\|_{\mathrm{op}} \le 4\sqrt{N}\, \mathbb{E}[\|\kappa(X_{k,0}) - \varphi_k(\xi_{k,0}; z_p)\| \mid \mathcal{F}_p] \xrightarrow{\mathbb{P}} 0.$$

Bounded convergence makes $V_{k,0}(z)$ continuous in z. Together with continuity of $G_0(z)$, continuous mapping and Slutsky's theorem therefore give the joint convergence

$$(G, V_{k,p}) \xrightarrow{d} (G_0(Z), V_{k,0}(Z)). \tag{22}$$

It remains to show $V_{k,0}(Z) \succ 0$ almost surely. The centers $Z_1, \dots, Z_N$ are pairwise distinct almost surely, by independence and absolute continuity of the training factors. Fix a configuration z with distinct centers. If $\alpha^\top V_{k,0}(z)\alpha = 0$, then $F(u) = \alpha^\top \varphi_k(u; z)$ is constant ν_k-almost surely. The density is positive on a nonempty open set, so continuity makes F constant throughout that set. The

identity theorem for real-analytic functions extends this constancy to $\mathbb{R}^r$. Each Gaussian term tends to zero as $\|u\| \to \infty$, so the constant must be zero. The functions in (21) are positive multiples of Gaussian translates with distinct centers $z_i - \boldsymbol{a}_k$. Their linear independence follows by evaluating any vanishing linear combination at these centers: the resulting Gaussian Gram matrix is strictly positive definite, forcing every coefficient to vanish. Hence $\alpha = 0$, and $V_{k,0}(z) \succ 0$. This holds almost surely simultaneously for all classes, since K is fixed.

By (22) and continuity of the inverse square root and smallest eigenvalue on positive-definite matrices,

$$\lambda^\star_{k,p} \xrightarrow{d} \Lambda_k := \lambda_{\min}\Big(G_0(Z)^{-1/2} V_{k,0}(Z) G_0(Z)^{-1/2}\Big), \qquad \Pr(\Lambda_k > 0) = 1.$$

Here $\lambda^\star_{k,p}$ may be set to zero on $G \not\succ 0$, whose probability tends to zero. The random limit depends on Z; no common realization of the factors across dimensions is required.

Finally, on $G \succ 0$, the defining minimum gives $V_{k,p} \succeq \lambda^\star_{k,p} G$. For any $\mathcal{F}_p$-measurable A with $A^\top G A = I_{K-1}$, conditioning fixes A, and therefore

$$\mathrm{Cov}\{A^\top \kappa(X_{k,0}) \mid \mathcal{F}_p\} = A^\top V_{k,p} A \succeq \lambda^\star_{k,p} A^\top G A = \lambda^\star_{k,p} I_{K-1}.$$

If these conditional covariances converged to zero in probability, then $\lambda^\star_{k,p}$ would also converge to zero in probability, contradicting its almost surely positive distributional limit. To rule out test piling itself, note that RKHS orthonormality and Bessel's inequality give $\|A^\top \kappa(x)\|^2 \le k_p(x,x) = 1$. If two conditionally independent class-k test scores concentrated at a common location, their difference would converge to zero in probability. Its squared norm is bounded by 4, so its expectation, which equals twice the expected trace of the conditional score covariance, would tend to zero. Markov's inequality would then force that conditional trace to converge to zero in probability, contradicting the variance bound. Assigning zero scores on the exceptional events $G \not\succ 0$ does not change these conclusions. This proves the obstruction for every normalized empirical-span map, including KMDP. □

C.5 Proof of Theorem 3

Proof. All limits are as $p \to \infty$, with the classwise training sample sizes fixed. We write k_p and $\tilde{k}_p$ to display the dimension dependence of the original and corrected kernels. The test observation $X_{k,0}$ is independent of $\mathcal{F}_p$, as specified in the sampling setup. Let $P := UU^\top$, $v_{k,i} := R_k^{1/2}\varepsilon_{k,i}$, $X^\circ_{k,i} := (I-P)X_{k,i}$, and $\tilde{X}_{k,i} := (I - \hat{P}_p)X_{k,i}$. The oracle projection removes the spike exactly, so $X^\circ_{k,i} = (I-P)(\mu_k + v_{k,i})$.

The residual moment estimates in the proof of Theorem 1 give $\|v_{k,i}\|^2/p \xrightarrow{L^2} \sigma_k^2$, $\mu_k^\top v_{l,j}/p \xrightarrow{\mathbb{P}} 0$, and $v_{k,i}^\top v_{l,j}/p \xrightarrow{\mathbb{P}} 0$ for distinct observations. Moreover, (6) implies $\|R_k\|_{\mathrm{op}}/p \to 0$, and hence $\mathbb{E}\|U^\top v_{k,i}/\sqrt{p}\|^2 \le r\|R_k\|_{\mathrm{op}}/p \to 0$. Combining these facts with $\mu_k^\top(I-P)\mu_l/p \to \gamma_{kl} - \boldsymbol{a}_k^\top \boldsymbol{a}_l$ yields

$$\frac{\|X^\circ_{k,i}\|^2}{p} \xrightarrow{\mathbb{P}} \tilde{\ell}_k, \qquad \frac{(X^\circ_{k,i})^\top X^\circ_{l,j}}{p} \xrightarrow{\mathbb{P}} \gamma_{kl} - \boldsymbol{a}_k^\top \boldsymbol{a}_l \quad ((k,i) \neq (l,j)). \tag{23}$$

These limits include fresh test observations.

The feature model also gives $\|X_{k,i}\|/\sqrt{p} = O_{\mathbb{P}}(1)$. Since both P and $\hat{P}_p$ are orthogonal projectors,

$$\left|\frac{\tilde{X}_{k,i}^\top \tilde{X}_{l,j} - (X^\circ_{k,i})^\top X^\circ_{l,j}}{p}\right| \le \|\hat{P}_p - P\|_{\mathrm{op}} \frac{\|X_{k,i}\|\,\|X_{l,j}\|}{p} = o_{\mathbb{P}}(1).$$

This bound holds also for $(k,i) = (l,j)$ and requires no independence between $\hat{P}_p$ and the observations. Thus (23) remains valid with X° replaced by $\tilde{X}$.

Let $\tilde{G}$ and $\tilde{\kappa}$ denote the corrected training Gram matrix and test-to-training kernel vector, and put $\tilde{G}_0 = E\tilde{B}E^\top + \mathrm{diag}_{i \le N}(\tilde{\eta}_{y_i})$. By continuity of ψ and fixed N,

$$\tilde{G} \xrightarrow{\mathbb{P}} \tilde{G}_0, \qquad \tilde{\kappa}(X_{k,0}) \xrightarrow{\mathbb{P}} E\tilde{B}\boldsymbol{e}_k.$$

Because $\gamma_{kk} - \boldsymbol{a}_k^\top \boldsymbol{a}_k \geq 0$ and $\sigma_k^2 > 0$, (7) gives $\tilde{\eta}_k > 0$. The class-average kernel-Gram argument from the proof of Theorem 1, applied to oracle-projected observations, gives $\tilde{B} + m^{-1}\operatorname{diag}(\tilde{\eta}_1, \ldots, \tilde{\eta}_K) \succeq 0$ for every fixed positive integer m. Letting $m \to \infty$ in this deterministic expression proves $\tilde{B} \succeq 0$. Consequently, $\tilde{G}_0 \succ 0$ and $\tilde{J} \succ 0$, and $\Pr(\tilde{G} \succ 0) \to 1$. On the exceptional event $\tilde{G} \not\succ 0$, set $A = 0$ and $g = 0$.

Let $D_n = \operatorname{diag}(n_1, \ldots, n_K)$ and let A be computed from the corrected Gram matrix as in (3). The block-system inversion and normalization in the proof of Theorem 1 apply to the same limiting matrix structure, with the tilded quantities, and yield $A \xrightarrow{\mathbb{P}} E D_n^{-1} T \tilde{J}^{-1/2}$. The rows of $\tilde{G}A$ therefore give

$$\bar{g}_k \xrightarrow{\mathbb{P}} \tilde{J}^{-1/2} T^\top \left\{ \tilde{B} + \operatorname{diag}(\tilde{\eta}_1/n_1, \ldots, \tilde{\eta}_K/n_K) \right\} \boldsymbol{e}_k.$$

This is the training-pile limit in (10) with the stated substitutions. Similarly, $g(X_{k,0}) = A^\top \tilde{\kappa}(X_{k,0}) \xrightarrow{\mathbb{P}} \tilde{t}_k$.

It remains to justify convergence of the conditional moments. For each realization of $\hat{P}_p$, composition with $I - \hat{P}_p$ preserves positive semidefiniteness of the kernel. On $\tilde{G} \succ 0$, Lemma 1 makes the coordinates of g orthonormal in the RKHS of $\tilde{k}_p$. Bessel's inequality, (7), and $\|(I - \hat{P}_p)x\| \leq \|x\|$ give

$$\|g(x)\|^2 \leq \tilde{k}_p(x, x) \leq L \left\{ 1 + \left(\frac{\|x\|^2}{p} \right)^{d_\psi} \right\}. \tag{24}$$

The same bound holds for the zero-score extension. For $d_\psi > 2$, (8) implies uniform integrability of the right-hand side. For $0 \leq d_\psi \leq 2$, note that at $\beta = 1$,

$$\frac{\|X_{k,0}\|^4}{p^2} \leq 27 \left\{ \frac{\|\mu_k\|^4}{p^2} + \|\xi_{k,0}\|^4 + \frac{\|v_{k,0}\|^4}{p^2} \right\}.$$

The first term is bounded and deterministic; the second has a fixed distribution with finite expectation; and the third converges to σ_k^4 in L^1, since $\|v_{k,0}\|^2/p \xrightarrow{L^2} \sigma_k^2$. Thus $\{\|X_{k,0}\|^4/p^2\}_p$ is uniformly integrable, which also controls $(\|X_{k,0}\|^2/p)^{d_\psi}$ for $d_\psi \leq 2$. It follows from (24) that $\{\|g(X_{k,0}) - \tilde{t}_k\|^2\}_p$ is uniformly integrable. Together with probability convergence, this gives $\mathbb{E}\|g(X_{k,0}) - \tilde{t}_k\|^2 \to 0$. By the definitions of m_k and C_k,

$$\|m_k - \tilde{t}_k\|^2 + \operatorname{tr} C_k = \mathbb{E}\left[\|g(X_{k,0}) - \tilde{t}_k\|^2 \mid \mathcal{F}_p \right].$$

The nonnegative right-hand side has expectation tending to zero. Markov's inequality therefore proves $m_k \xrightarrow{\mathbb{P}} \tilde{t}_k$ and $\operatorname{tr} C_k \xrightarrow{\mathbb{P}} 0$.

Finally, put $q = \boldsymbol{e}_k - \boldsymbol{e}_l$ for $k \neq l$. Invertibility of $\tilde{J}$ gives $\tilde{t}_k = \tilde{t}_l$ if and only if $T^\top \tilde{B} q = 0$. The latter implies $\tilde{B}q \in \operatorname{span}\{\mathbf{1}_K\}$, so $q^\top \tilde{B} q = 0$ because $q \perp \mathbf{1}_K$. Conversely, $q^\top \tilde{B} q = 0$ and $\tilde{B} \succeq 0$ imply $\tilde{B}q = 0$, hence $\tilde{t}_k = \tilde{t}_l$. Since $q^\top \tilde{B} q = \tilde{B}_{kk} + \tilde{B}_{ll} - 2\tilde{B}_{kl} \geq 0$, the stated separation criterion follows. □