TRANSDUCTIVE LEARNING OVER UNBOUNDED LOSSES: AN UNBOUNDED TRANSDUCTIVE LOCAL COMPLEXITY APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop Unbounded Transductive Local Complexity (UTLC) for transductive learning with unbounded losses, such as squared-loss regression without norm constraints. We first develop a novel variance rescaling method to extend the existing Sharper Transductive Local Complexity (STLC) (Yang, 2026) concentration to unbounded loss functions. Under standard localization and Bernstein conditions, a single complexity fixed point controls empirical risk minimization. UTLC is applied to two transductive learning problems with squared-loss. For Transductive Kernel Learning (TKL) over an unbounded set in a Reproducing Kernel Hilbert Space (RKHS), UTLC gives spectral excess-risk rates for unbounded squared loss, including controlled positive ridge chosen from training labels. UTLC with bounded loss on a fixed RKHS ball recovers the rate of Yang (2026, Theorem 5.2) only as a special case under the same assumptions for Yang (2026, Theorem 5.2). UTLC also renders a sharp rate for noisy unbounded TKL with i.i.d. sub-Gaussian noise in labels. With independent inputs and a two-sided polynomial population Eigenvalue Decay Rate (EDR), UTLC attains the minimax optimal rate on bounded RKHS signal sets, matching a Gaussian-noise lower bound under the same polynomial population EDR. The learner optimizes over an unbounded RKHS set, and its squared loss is unbounded. To the best of our knowledge, this is the first transductive local-complexity analysis to establish minimax optimal bounds for noisy TKL with unbounded squared loss.

1 INTRODUCTION

Transductive learning reveals the labels of a uniformly sampled training subset of a fixed full sample and evaluates predictions on the complementary test subset. We write m and u for the training and test sizes, and $n = m + u$ is the full sample size. Classical analyses use VC dimension, global complexity, or stability (Vapnik, 1982; 1998; El-Yaniv & Pechyony, 2009; Cortes & Mohri, 2006; Cortes et al., 2009). Local-complexity methods sharpen these bounds by measuring the class near a risk minimizer (Bartlett et al., 2005; Tolstikhin et al., 2014). More recently, sharper Transductive Local Complexity (STLC) (Yang, 2026) uses a sharper local transductive complexity (TC) method to remove an earlier logarithmic confidence factor in (Yang, 2025b), avoid the undesirable restrictive training/test size regime $m \gg u^2$ or $u \gg m^2$ in (Yang, 2025a), and remove the undesirable factors n/u, n/m in (Tolstikhin et al., 2014), achieving sharper bounds than the previous methods.

Local-complexity methods yield sharp transductive excess-risk bounds (Tolstikhin et al., 2014; Yang, 2025a;b; 2026), but these results assume uniformly bounded loss classes. This restriction excludes important regression settings. For squared loss, no finite uniform bound exists when predictions can grow arbitrarily at a sample point, as can occur in kernel learning without a norm constraint. Gaussian observation noise also prevents a deterministic loss bound uniform over noise realizations. These settings motivate extending transductive local-complexity analysis beyond bounded losses while retaining its sharp risk rates. We develop Unbounded Transductive Local Complexity (UTLC) to address this challenge. Our contributions are as follows.

Contribution. We introduce Unbounded Transductive Local Complexity (UTLC) through a novel variance-rescaling method. Variance rescaling places functions with different variances in one class

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.

with a finite envelope, so bounded STLC concentration inequalities (Yang, 2026) apply. Undoing the rescaling leads to concentration inequalities on the original class. Theorem 4.1 extends concentration under variance control and the standard localized TC condition, to be defined in (4.1), which is widely used in local-complexity-based transductive learning (Tolstikhin et al., 2014; Yang, 2025a;b; 2026). With the Bernstein geometry of Assumption 1, Theorem 4.2 gives an excess-risk bound for empirical risk minimization (ERM) on the training set for unbounded losses. Applying Theorem 4.2 with bounded losses give a special case recovering the rate of Yang (2026, Theorem 4.4) under its Bernstein and localization assumptions. Appendix C.3 gives a squared-loss example satisfying Assumption 1.

The first application of UTLC is Transductive Kernel Learning (TKL) over an unbounded RKHS set, where Theorem 4.3 renders spectral rates as the upper bound for the excess risk of the empirical loss minimizer of ridge least squares on the training set. As a special case of Theorem 4.3, bounded loss on a fixed RKHS ball gives (D.49), recovering the rate of Yang (2026, Theorem 5.2) under its same assumptions. Theorem 4.3 also introduces a novel proof strategy of independent interest which involves orthogonal projection of loss functions onto the span of different empirical loss minimizers on different subsets of the full sample, therefore preserves these prediction differences while removing unused kernel directions from the complexity bound.

We further apply UTLC to noisy unbounded TKL with i.i.d. sub-Gaussian label noise. We bound the prediction error, which is the mean squared error (MSE) on the test set between the empirical loss minimizer on the training set and the true regression function. With a polynomial population Eigenvalue Decay Rate (EDR) of order $j^{-2\beta}$ for some $\beta > 1/2$ and all $j \geq 1$, Theorem 4.4 bounds prediction error by $\mathcal{O}(m^{-2\beta/(2\beta+1)})$ for $u \geq m$. Under the same polynomial population EDR, Theorem D.5 proves a matching Gaussian-noise lower bound on bounded RKHS signal sets. The empirical loss minimizer's optimization domain remains unbounded. To the best of our knowledge, this is the first transductive local-complexity analysis to establish matching minimax bounds for noisy TKL with potentially unbounded squared loss. While the inductive minimax rate is classical (Caponnetto & De Vito, 2007), our Theorem 4.4 establishes it for the complementary test set under a transductive learning setup. Section 3.2 summarizes our main results, which are detailed in Sections 4. The appendix gives the detailed proofs.

2 Notation

Table 1 summarizes the notation for the fixed-sample analysis. Throughout, $m, u \geq 2$ are integers and $n = m + u$. As in a standard transductive learning setup, we consider the indexed full sample $((\mathbf{x}_i, y_i))_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y} \subseteq \mathbb{R}$. The learner observes the unlabeled indexed feature sample $\mathbf{X}_n := (\mathbf{x}_i)_{i=1}^n$ before the split. Sampling is over indices, so repeated feature values are allowed. Write $\mathbb{N} := \{1, 2, \ldots\}$ and $[n] := \{1, \ldots, n\}$. A loss function h on $\mathcal{X} \times \mathcal{Y}$, or on $\mathcal{X}$, is identified with its restriction to the indexed full sample, with $h(i) = h(\mathbf{x}_i, y_i)$ or $h(i) = h(\mathbf{x}_i)$. For nonempty $A \subseteq [n]$ and $h : [n] \to \mathbb{R}$, write $P_A h := |A|^{-1} \sum_{i \in A} h(i)$ for the subset average and $P_n h := P_{[n]} h$ for the full-sample mean. The full-sample variance is $\mathrm{Var}\,[h] := P_n(h - P_n h)^2$. We note that throughout this paper, $\mathrm{Var}\,[h]$ is finite for every loss function h, but there may not be a finite uniform bound for $\mathrm{Var}\,[\cdot]$ over the entire function class.

Bold lowercase and uppercase letters denote vectors and matrices, respectively, and $\mathbf{0}$ denotes a zero vector or zero matrix of the required dimension. The vector $\mathbf{1}$ has every coordinate equal to

Symbol	Description
$n = m + u$	Full sample size
$N_{u,m}$	$\min\{u, m\}$
$(\mathbf{x}_i, y_i)$	Input and label at index i
$\mathbf{X}_n$	Full sample
$\mathcal{M}, \mathcal{U}$	Training and test index sets
$P_A h$	Average of h over A
$P_n, P_\mathcal{M}, P_\mathcal{U}$	Full, training, and test averages
$\ell_f(i)$	Loss of predictor f at index i
$g(\mathcal{H})$	Test–train supremum over $\mathcal{H}$
$B_\infty(\mathcal{H})$	Envelope
$\mathcal{H}^{[2]}$	Class of centered squares
$\mathfrak{R}_p^\eta(\mathcal{H})$	One-sided transductive complexity
$V(h)$	Localization functional
$\psi,\ r^*$	Sub-root majorant and its fixed point
$\mathcal{E}_\mathcal{U}(\widehat{f}_m)$	Excess risk of the empirical loss minimizer $\widehat{f}_m$

Table 1: Main notation.

one, and $\mathbf{I}_p$ is the identity matrix of integer order
$p \geq 1$. Products, powers, and absolute values of functions are taken pointwise and those of vectors coordinatewise. In particular, $(h - P_n h)(i) = h(i) - P_n h$, and $P_n h^2$ means $P_n(h^2)$. For finite vectors, $\|\mathbf{v}\|_2$ is the Euclidean norm, and $\|\mathbf{v}\|_\infty$ is the maximum absolute coordinate. We write $a \vee b \coloneqq \max\{a, b\}$, $a_+ \coloneqq \max\{a, 0\}$, and $\mathbb{1}_{\{E\}}$ for the indicator of an event E.

The random test index set $\mathcal{U}$ is uniform among the u-subsets of $[n]$, and the training index set $\mathcal{M} \coloneqq [n] \setminus \mathcal{U}$ has size m. Recall that $N_{u,m} = \min\{u, m\} \geq 2$. For a nonempty function class $\mathcal{H} \subset \mathbb{R}^{[n]}$ of real-valued functions on $[n]$, we define the test–train process

$$g(\mathcal{H}) \coloneqq \sup_{h \in \mathcal{H}} (P_{\mathcal{U}} h - P_{\mathcal{M}} h). \tag{2.1}$$

This is the empirical process studied in Yang (2025a;b; 2026) for bounded loss functions. For one function, write $g(h) \coloneqq g(\{h\}) = P_{\mathcal{U}} h - P_{\mathcal{M}} h$. Conditional on the full sample, only the split is random in (2.1). Throughout this paper the random split means the split that generates $\mathcal{U}, \mathcal{M}$. The notation $g(\mathcal{H})$ suppresses this dependence. The split-balance factor $c_{\text{split}} \coloneqq n/\max\{u, m\} \in (1, 2]$ gives $g(h) = c_{\text{split}}(P_{\mathcal{U}} - P_n)h$ when $u \leq m$ and $g(h) = c_{\text{split}}(P_n - P_{\mathcal{M}})h$ otherwise.

For $1 \leq p < n$, let A_p be uniform among the p-subsets of $[n]$. Following STLC notation (Yang, 2026), for $h : [n] \to \mathbb{R}$ we define the two one-sided subset deviations by $R_p^+ h \coloneqq (P_{A_p} - P_n)h$ and $R_p^- h \coloneqq -R_p^+ h$. For a nonempty class $\mathcal{H} \subset \mathbb{R}^{[n]}$ and $\eta \in \{+, -\}$, the Transductive Complexity (TC) (Yang, 2025a;b) is defined as follows.

Definition 2.1 (Transductive Complexity (Yang, 2025a, Definition 2.2)). Transductive Complexity (TC) of a function class $\mathcal{H}$ is defined as $\mathfrak{R}_p^\eta(\mathcal{H}) \coloneqq \mathbb{E}_{A_p}\left[\sup_{h \in \mathcal{H}} R_p^\eta h\right]$.

For centered-square applications, we set $\mathcal{H}^{[2]} \coloneqq \left\{(h - P_n h)^2 : h \in \mathcal{H}\right\}$. The second-order quantity $\mathfrak{R}_p^+(\mathcal{H}^{[2]})$ measures deviations of centered squares. TC differs from the ternary-sign complexity of El-Yaniv & Pechyony (2009) and the signed permutational symmetrization of Tolstikhin et al. (2015), and it measures the one-sided deviation of the actual subset from the full-sample mean. The convention $0 \in \mathcal{H}$ makes every one-sided supremum nonnegative. For a nonempty class $\mathcal{H}$, we define its envelope by

$$B_\infty(\mathcal{H}) \coloneqq \max_{i \leq n} \sup_{h \in \mathcal{H}} |h(i) - P_n h|. \tag{2.2}$$

An envelope bound for $\mathcal{H}$ is a deterministic $b \in [0, \infty]$ satisfying $B_\infty(\mathcal{H}) \leq b$.

Appendix A defines the inductive Rademacher complexity (RC) used to bound TC. For nonnegative quantities a, b, we write $a \lesssim b$ if $a \leq Cb$ for some constant $C > 0$, and $a \asymp b$ or $a = \Theta(b)$ if $cb \leq a \leq Cb$ for constants $0 < c \leq C < \infty$. These comparisons hold uniformly over the indicated range. The constants are independent of varying sample sizes, indices, and confidence parameters. Their allowed dependence on fixed parameters is stated locally, and subscripts specify a particular dependence.

3 Problem Setup and Summary of Main Results

3.1 Problem Setup

Let $\mathcal{F}$ be a nonempty prediction class. For each $f \in \mathcal{F}$, let $\ell_f(\mathbf{x}, y) = \ell(f(\mathbf{x}), y)$ be its loss function. Write $\ell_f(i) = \ell_f(\mathbf{x}_i, y_i) \in \mathbb{R}$ for $i \in [n]$. The UTLC inequality in Theorem 4.1 and the ERM bound in Theorem 4.2 allow signed loss functions. The transductive excess risk of an empirical loss minimizer on the training set is $\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) = P_{\mathcal{U}}\ell_{\widehat{f}_m} - \inf_{f \in \mathcal{F}} P_{\mathcal{U}}\ell_f = P_{\mathcal{U}}(\ell_{\widehat{f}_m} - \ell_{\widehat{f}_u})$, with a training minimizer $\widehat{f}_m \in \arg\min_{f \in \mathcal{F}} P_{\mathcal{M}}\ell_f$ and a test oracle $\widehat{f}_u \in \arg\min_{f \in \mathcal{F}} P_{\mathcal{U}}\ell_f$. Although every sample loss $\ell_f(i)$ is finite, the loss class is unbounded on the full sample when $\sup_{f \in \mathcal{F}} |\ell_f(i)| = +\infty$ for some $i \in [n]$. Theorems 4.1 and 4.2 permit unbounded losses with potentially infinite envelope under their geometric assumptions and the localized TC condition (4.1).

3.2 Summary of Main Results

UTLC extends transductive local-complexity analysis to unbounded losses, including classes with infinite envelopes. Its novel variance-rescaling method yields concentration and ERM excess-risk

Quantity to be bounded	**UTLC bounds (ours)**	**Yang (2026) bounds**	**Recovery of bounded-loss rate (under the same assumptions as Yang (2026))**
Test-train process	Theorem 4.1, $\lesssim r^* + p^{-1}$	Fixed point $+p^{-1}$	Yang (2026, Theorem 4.2), when applying Theorem 4.1 to bounded loss
ERM excess risk	Theorem 4.2, $(\theta = 1)$ $\lesssim r_{\mathrm{exc}} + p^{-1}$	Two fixed points $+p^{-1}$	Yang (2026, Theorem 4.4), when applying Theorem 4.2 $(\theta = 1)$ to bounded loss
Excess risk of U-TKL,	Theorem 4.3, Spectral bound (4.5)	Spectral bound (4.5)	Yang (2026, Theorem 5.2), when applying Theorem 4.3 to a RKHS ball.
Prediction error of noisy U-TKL,	Theorem 4.4, $\lesssim m^{-\alpha}$, minimax optimal	Not Applicable	No corresponding result in the literature under the same noisy U-TKL setup. Our rate holds on an unbounded RKHS set.

Table 2: Upper bound (rate) comparison, and each row is compared to the current state-of-the-art under the corresponding setup. U-TKL stands for Unbounded TKL. In the first three rows, our UTLC recovers the bounded-loss results in Yang (2026) under the same assumptions as Yang (2026). Here $p = N_{u,m}$ and $\alpha = 2\beta/(2\beta+1)$ for some $\beta > 1/2$. The UTLC and ERM fixed points are r^* and r_{exc}. We remark that the UTLC results in all rows hold for unbounded losses under their corresponding conditions.

bounds under the standard localization and geometric conditions (Theorems 4.1 and 4.2). Bounded losses remain a special case recovering the corresponding rates of Yang (2026) under the same assumptions of Yang (2026). UTLC is applied to derive bounds for the excess risk of unbounded TKL and the prediction error of noisy unbounded TKL. For unbounded TKL, Theorem 4.3 gives spectral excess-risk bounds while allowing optimization over an unbounded RKHS set. For noisy unbounded TKL, Theorem 4.4 gives a sharp rate for the prediction error with a data-based ridge choice, where both optimization of the empirical loss minimizer and the true regression function are over an unbounded RKHS set. Together with the matching lower bound in Theorem D.5, Theorem 4.4 establishes minimax optimal prediction error rates with high probability on fixed RKHS signal balls under two-sided population EDR and positive Gaussian noise. These results show that sharp transductive guarantees, including minimax optimal prediction under label noise, are compatible with unbounded squared loss and an unbounded optimization domain. Table 2 summarizes our main results. It is noted that all existing local-complexity-based transductive bounds on bounded loss (Tolstikhin et al., 2014; Yang, 2025a;b; 2026) are not applicable for both unbounded TKL and noisy unbounded TKL.

4 Main Results

4.1 Unbounded Transductive Local Complexity (UTLC)

We encode localized complexity by a sub-root function. A function $\psi : [0,\infty) \to [0,\infty)$ is a sub-root function (Bartlett et al., 2005) if it is nondecreasing and $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing on $(0,\infty)$. A nonzero sub-root function has a unique positive fixed point $r^* > 0$, characterized by $\psi(r^*) = r^*$. We designate $r^* = 0$ for the zero function. For $r \geq r^*$, the sub-root property gives $\psi(r) \leq \sqrt{r r^*} \leq r$, which permits absorption of the complexity term above the fixed point.

We use a surrogate localization functional $V : \mathcal{H} \to [0,\infty)$ satisfying $\mathrm{Var}[h] \leq V(h) < +\infty$ for every $h \in \mathcal{H}$. It defines the localized class $\mathcal{H}(r) := \{h \in \mathcal{H} : V(h) \leq r\} \cup \{0\}$. We require a sub-root majorant ψ satisfying the following localized TC condition,

$$\psi(r) \geq \max\left\{ \max_{\eta\in\{+,-\}} \mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{H}(r)), \mathfrak{R}^{+}_{N_{u,m}}\left(\mathcal{H}(r)^{[2]}\right) \right\}, \qquad r \geq 0. \tag{4.1}$$

Condition (4.1) controls localized mean deviations in both orientations and fluctuations of centered squares. Enlarging V shrinks $\mathcal{H}(r)$ and can restore a common sub-root majorant ψ. Lemma A.4 in Appendix A illustrates both the obstruction and this remedy for an infinite-envelope class. Our

concentration input is the centered form of Yang (2026, Theorem 4.1), stated as Theorem B.1 in Appendix B. We set $c_0 = \sqrt{219}$.

Theorem 4.1 (UTLC Inequality)**.** Let $0 \in \mathcal{H} \subset \mathbb{R}^{[n]}$ and let $V : \mathcal{H} \to [0, \infty)$ satisfy $\mathrm{Var}[h] \le V(h) < \infty$ for every $h \in \mathcal{H}$. Define $\mathcal{H}(r)$ as above and suppose a sub-root function ψ satisfies (4.1), with fixed point $r^* \ge 0$. Set $c_1 := 64(4 + c_0)^2$. For every $A, x \ge 1$, with probability at least $1 - 2\exp(-x)$ over the random split, for all $h \in \mathcal{H}$,

$$|g(h)| \le \frac{V(h)}{A} + c_1 A \left(r^* + \frac{x}{N_{u,m}} \right) + c_1 \min \left\{ A(n-1)\frac{x^2}{N_{u,m}^2}, B_\infty(\mathcal{H})\frac{x}{N_{u,m}} \right\}. \tag{4.2}$$

For either fixed $\varepsilon \in \{+1, -1\}$, the same right-hand side bounds $\varepsilon g(h)$ simultaneously over $\mathcal{H}$ with probability at least $1 - \exp(-x)$ over the random split.

Lemma A.5 gives $r^* < \infty$, so that the RHS of (4.2) is always finite even with infinite envelope, $B_\infty(\mathcal{H}) = +\infty$, under the conditions of Theorem 4.1.

Recovery of the bounded STLC inequality (Yang, 2026, Theorem 4.2). If $|h(i)| \le H_0 < \infty$ for a positive H_0 and all $h \in \mathcal{H}$, $i \in [n]$, then $B_\infty(\mathcal{H}) \le 2H_0$, and the envelope branch of (4.2) gives $g(h) \le V(h)/A + c_1 A r^* + c_1(A + 2H_0)x/N_{u,m}$ uniformly with probability at least $1 - \exp(-x)$ over the random split. This recovers the fixed-point form of Yang (2026, Theorem 4.2) as a bounded-class special case. Lemma B.3 converts its localization hypothesis to (4.1), preserving fixed-point order for fixed H_0. Unbounded classes are covered under $\mathrm{Var}[h] \le V(h) < \infty$ and (4.1). A finite envelope replaces $2H_0$ by $B_\infty(\mathcal{H})$ above. An infinite envelope adds $c_1 A(n-1)x^2/N_{u,m}^2$, of order Ax^2/n for balanced splits.

Proof roadmap for Theorem 4.1 is presented as follows. Appendix B supplies the full argument.

(1) Rescale by variance. Lemma B.2 forms one class $\mathcal{A}_r$ from $4^{-k}h$ on each shell $4^{k-1}r < V(h) \le 4^k r$, $k \ge 1$, and leaves functions with $V(h) \le r$ unchanged. The key reduction is $B_\infty(\mathcal{A}_r) \le \min\left\{ B_\infty(\mathcal{H}), \sqrt{(n-1)r} \right\}$, which is finite even for an infinite original envelope.

(2) Concentrate and undo the rescaling. Condition (4.1) and sub-root growth bound its first-order and centered-square TCs by $2\psi(r)$. We choose a deterministic radius from r^* and the envelope alternatives, and apply Theorem B.1 once. Undoing the rescaling obtains (4.2).

4.2 Excess Risk of Empirical Risk Minimizer

Fast excess-risk bounds use low variance near a full-sample loss minimizer, called the full-sample oracle. The centered Bernstein condition below permits unbounded excesses.

Assumption 1 (Transductive ERM Geometry)**.** (1) Each $\ell_f(i)$ is finite for $f \in \mathcal{F}$ and $i \in [n]$, and there exists $f_n^\star \in \arg\min_{f \in \mathcal{F}} P_n \ell_f$.

(2) There are $B > 0$ and $0 < \theta \le 1$ such that $\mathrm{Var}[g_f] \le B\,(P_n g_f)^\theta$ for every $f \in \mathcal{F}$, where $g_f := \ell_f - \ell_{f_n^\star}$.

Assumption 1 is used broadly in the literature as the Bernstein geometry of prior fast-rate local-complexity analyses (Bartlett et al., 2005; Tolstikhin et al., 2014; Yang, 2025b; 2026). It makes fluctuation variance small when full-sample excess risk is small. At $\theta = 1$, centering weakens the cited transductive second-moment condition and allows an infinite envelope. Lemma C.4 gives three sufficient conditions for Assumption 1, provided its finiteness and minimizer-existence requirements hold. Example C.5 verifies Assumption 1 for a squared-loss class with an infinite envelope and an oracle-excess fixed point at most $1/N_{u,m}$. Both results are presented in Appendix C.3. Lemma D.1 also shows that Assumption 1 holds for the unbounded squared-loss for TKL.

The function g_f is the oracle excess, whereas $g(g_f)$ is its test–train deviation. Set $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}\}$, $V_\star(g_f) = B(P_n g_f)^\theta$, and $\mathcal{G}_\star(r) = \{g_f : V_\star(g_f) \le r\}$. The oracle excess $g_{f_n^\star}$ is zero, so it belongs to every localization. Assume a finite-valued sub-root function $\psi_\star$ satisfies (4.1) for $\mathcal{G}_\star$ and $V_\star$. Its designated fixed point r_{exc} is finite by Lemma C.1 in Appendix C.

Theorem 4.2 (Unbounded-Loss Transductive ERM)**.** Suppose Assumption 1 holds and $\psi_\star$ satisfies the localized TC condition (4.1) for $\mathcal{G}_\star$ and $V_\star$. Suppose also that the empirical minimum is attained

on every subset of size m or u, and fix any corresponding ERM selections $\widehat{f}_m$, $\widehat{f}_u$ as in Section 3. Fix $x \geq 1$, and put

$$s = x + \log 2, \qquad A_0 = 2(1 \vee B), \qquad C_B = 3c_1 A_0,$$
$$\delta_s = r_{\mathrm{exc}} + \frac{s}{N_{u,m}}, \qquad \tau_s = (n-1)\frac{s^2}{N_{u,m}^2}, \qquad \Phi_\theta(v) = v \vee v^{1/(2-\theta)} \quad (v \geq 0).$$

Here c_1 is the constant in Theorem 4.1. Then with probability at least $1 - \exp(-x)$ over the random split,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) \leq C_B \min\left\{\Phi_\theta(\delta_s + \tau_s), \Phi_\theta(\delta_s) + B_\infty(\mathcal{G}_\star)\frac{s}{N_{u,m}}\right\}. \tag{4.3}$$

The constant C_B is independent of θ. At $\theta = 1$, $\Phi_1(v) = v$, so the right-hand side is exactly $C_B[\delta_s + \min\{\tau_s, B_\infty(\mathcal{G}_\star)s/N_{u,m}\}]$. When $B_\infty(\mathcal{G}_\star) = \infty$, the first candidate $\Phi_\theta(\delta_s + \tau_s)$ remains finite.

For $0 \leq v \leq 1$, $\Phi_\theta(v) = v^{1/(2-\theta)}$, so (4.3) displays the Bernstein exponent explicitly. At $\theta = 1$, bounded losses recover the rate of Yang (2026, Theorem 4.4) as a special case under its Bernstein, localization, and minimizer-existence assumptions (Lemma B.3). Under Assumption 1, attained minima, and (4.1), unbounded losses also give order $r_{\mathrm{exc}} + s/N_{u,m}$ for uniformly bounded $B, B_\infty(\mathcal{G}_\star)$. An infinite envelope adds $(n-1)s^2/N_{u,m}^2$ in (4.3). Appendix C proves Theorem 4.2. The proof follows the roadmap below.

(1) Control both empirical minimizers. Apply the one-sided forms of Theorem 4.1 to $\mathcal{G}_\star$ in opposite directions at confidence parameter $s = x + \log 2$. Empirical optimality and $P_n = (m/n)P_{\mathcal{M}} + (u/n)P_{\mathcal{U}}$ bound the full-sample excesses $P_n g_{\widehat{f}_m}$ and $P_n g_{\widehat{f}_u}$ on the intersection of these events.

(2) Compare through the full-sample oracle. Write the loss difference between the learner and test oracle as $\ell_{\widehat{f}_m} - \ell_{\widehat{f}_u} = g_{\widehat{f}_m} - g_{\widehat{f}_u}$. Combine the positive deviation bound for $g_{\widehat{f}_m}$ and the negative one for $g_{\widehat{f}_u}$ with training optimality. The bounds from step (1) and Lemma C.3 then give (4.3) for every $0 < \theta \leq 1$, with the single fixed point r_{exc} and no additional failure probability.

4.3 Unbounded transductive kernel learning (TKL)

In this section, we apply UTLC to squared-loss TKL over an unbounded set in a RKHS, termed the unbounded TKL. In this section we bound ordinary excess risk for fixed labels. Noisy unbounded TKL with sub-Gaussian label noise is studied in the next subsection.

Let K be a positive semidefinite kernel with the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$, whose inner product and norm are $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ and $\|\cdot\|_{\mathcal{H}_K}$. Its Gram matrix is $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. For nonempty $A \subseteq [n]$, write $\mathbf{K}_A = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j\in A}$ for its subset Gram matrix. The labels $\mathbf{y} = (y_i)_{i=1}^n \in \mathbb{R}^n$ are fixed until the noisy extension in Section 4.4. Write $y(i) = y_i$ for the label function on $[n]$. We optimize over the full, finite-dimensional and closed span $\mathcal{F} = \mathcal{H}_{\mathbf{X}_n} = \mathrm{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. If $\mathbf{K} \neq \mathbf{0}$, $\mathcal{F}$ has no finite RKHS radius or squared-loss bound.

Fix $\lambda \geq 0$ and define the ridge-regularized least-squares objective $J_A(f) := P_A(f-y)^2 + \lambda\|f\|_{\mathcal{H}_K}^2$ for nonempty $A \subseteq [n]$. Let $\widehat{f}_{A,\lambda} \in \arg\min_{f\in\mathcal{F}} J_A(f)$ minimize J_A over $\mathcal{F}$, uniquely for $\lambda > 0$, with minimum-RKHS-norm selection at $\lambda = 0$. At zero ridge, RKHS gradient descent initialized at zero, also called Landweber iteration (Yao et al., 2007), converges in RKHS norm to $\widehat{f}_{A,0}$ for a sufficiently small constant step size. Appendix D.1 proves this for the update (D.6). We refer to these $\widehat{f}_{A,\lambda}$ as subset solutions in the sequel. We write $\widehat{f}_{m,\lambda} = \widehat{f}_{\mathcal{M},\lambda}$ and $\widehat{f}_{u,\lambda} = \widehat{f}_{\mathcal{U},\lambda}$. We consider squared loss, $\ell_f(i) = (f(\mathbf{x}_i) - y_i)^2$, and $\mathcal{E}_{\mathcal{U}}(f) = P_{\mathcal{U}}(\ell_f - \ell_{\widehat{f}_{u,0}})$. The ridge penalty enters only the fitting and intermediate analysis.

A positive ridge discourages large RKHS norms and gives a unique subset solution without imposing a norm constraint. We first establish a spectral bound at zero ridge, then choose a positive ridge using training labels and observed inputs. By controlling the resulting change in predictions, we obtain the corresponding bound for the positive-ridge subset solution.

For this section, a relevant subset is $A \subseteq [n]$ with $|A| \in \{m, u, n\}$. The convex hull of their zero-ridge subset solutions, $\mathcal{F}_0^{\mathrm{sel}}$ defined in (D.3), contains the zero-ridge training and test subset solutions

for every split and is fixed before sampling. Its geometry supplies the localization used in the proof. Let Π_0 be the RKHS orthogonal projection onto the span of $\widehat{f}_{A,0} - \widehat{f}_{B,0}$ over relevant A, B, as defined in (D.1) of the appendix at $\lambda = 0$, and set $(\mathbf{K}_0)_{ij} = \langle \Pi_0 K(\cdot, \mathbf{x}_i), \Pi_0 K(\cdot, \mathbf{x}_j) \rangle_{\mathcal{H}_K}$. We set $K_{\max} = \max_{i \le n} K(\mathbf{x}_i, \mathbf{x}_i)$ and define the diameter D_0 and residual maximum L_0 as

$$D_0 = \max_{\substack{A,B \subseteq [n] \\ |A|,|B| \in \{m,u,n\}}} \left\| \widehat{f}_{A,0} - \widehat{f}_{B,0} \right\|_{\mathcal{H}_K}, \qquad L_0 = \max_{\substack{A \subseteq [n],\ |A| \in \{m,u,n\} \\ i \in [n],\ (\mathbf{K}_0)_{ii} > 0}} \left(\widehat{f}_{A,0}(\mathbf{x}_i) - y_i \right)^2.$$

An empty set in the definition of L_0 renders a zero residual maximum.

We use the orthogonal projection Π_0 because excess-risk comparisons depend only on differences between subset solutions. For every difference d of two members of $\mathcal{F}_0^{\mathrm{sel}}$, we have $d(\mathbf{x}_i) = \langle d, \Pi_0 K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_K}$. Projection therefore preserves these prediction differences while removing unused kernel directions from the complexity bound. Let $\widehat{\lambda}_j^{(0)} = \lambda_j(\mathbf{K}_0/n)$, $j \in [n]$, be the nonincreasing eigenvalues, counted with multiplicity. We write $\lambda_j(\mathbf{K}/n)$ for the ordered original eigenvalues of $\mathbf{K}/n$.

For symmetric matrices, $\mathbf{A} \preceq \mathbf{B}$ means $\boldsymbol{a}^\top (\mathbf{B} - \mathbf{A}) \boldsymbol{a} \ge 0$ for every vector $\boldsymbol{a}$. Lemma D.1 proves $\mathbf{0} \preceq \mathbf{K}_0 \preceq \mathbf{K}$ and identifies $(\mathbf{K}_0)_{ii} = 0$ with agreement of predictions of all relevant subset solutions at $\mathbf{x}_i$. Projection therefore removes unused directions from the spectral bound, while the residual maximum L_0 excludes coordinates whose common loss cancels from every comparison.

The following theorem gives the excess risk of the empirical loss minimizer $\widehat{f}_{m,\lambda_{\mathcal{M}}}$ obtained by minimization of the ridge-regularized least-squares objective $J_{\mathcal{M}}$ at a selected ridge $\lambda_{\mathcal{M}}$.

Theorem 4.3 (Spectral Rates For TKL With Training-Label Ridge Selection)**.** In the unbounded TKL setup above, fix a deterministic prediction tolerance $\varepsilon > 0$. Recall that $\mathcal{M}$ is the index set of the training set. If $\mathbf{K}_{\mathcal{M}} \neq \mathbf{0}$, let $\kappa_{\mathcal{M}}$ be the smallest positive eigenvalue of $\mathbf{K}_{\mathcal{M}}/m$ and choose

$$\lambda_{\mathcal{M}} = \kappa_{\mathcal{M}} \frac{\varepsilon}{\sqrt{K_{\max}} \left\| \widehat{f}_{m,0} \right\|_{\mathcal{H}_K} + \varepsilon}. \tag{4.4}$$

If $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$, use any fixed positive value as $\lambda_{\mathcal{M}}$. This ridge is strictly positive and uses only training labels and the observed inputs. For a universal $C > 0$ and every $x \ge 1$, there is an event $\Omega_0(x)$ of probability at least $1 - \exp(-x)$ over the random split on which

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \le 2C \min_{0 \le Q \le \mathrm{rank}(\mathbf{K}_0)} \left\{ \frac{L_0(Q + x)}{N_{u,m}} + D_0 \sqrt{\frac{L_0}{N_{u,m}} \sum_{j > Q} \lambda_j(\mathbf{K}/n)} \right\} + 2\varepsilon^2. \tag{4.5}$$

Furthermore, if $\widehat{\lambda}_j^{(0)} \le C_K j^{-2\beta}$ for $j \in [n]$, where $C_K > 0$ and $\beta > 1/2$, then on the same event $\Omega_0(x)$,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \le 2C_\beta \min \left\{ \frac{L_0 \mathrm{rank}(\mathbf{K}_0)}{N_{u,m}}, (C_K D_0^2)^{1/(2\beta+1)} \left(\frac{L_0}{N_{u,m}} \right)^{2\beta/(2\beta+1)} \right\} + \frac{2C_\beta L_0 x}{N_{u,m}} + 2\varepsilon^2, \tag{4.6}$$

where C_β depends only on β. A bound $\lambda_j(\mathbf{K}/n) \le C_K j^{-2\beta}$ on the original spectrum also suffices. The bounds (4.5) and (4.6) also hold on $\Omega_0(x)$ for the zero-ridge learner $\widehat{f}_{m,0}$, with the term $2\varepsilon^2$ omitted. For each fixed x, all these conclusions hold on $\Omega_0(x)$ simultaneously for every $\varepsilon > 0$, with the ridge recomputed from (4.4) for each tolerance $\varepsilon > 0$ when $\mathbf{K}_{\mathcal{M}} \neq \mathbf{0}$.

Bounded-loss special case at zero ridge. Appendix D.7 proves (D.49), recovering Yang (2026, Theorem 5.2) as a bounded-loss special case. Both empirical minimizers there are taken without ridge over the fixed ball $\mathcal{F}_{\mu_0} = \{ f \in \mathcal{F} : \|f\|_{\mathcal{H}_K} \le \mu_0 \}$, where $\mu_0 = \Theta(1) > 0$. For fixed $M > 0$, $L \ge 0$, and $B' > 0$, the loss satisfies $0 \le \ell_f(i) \le M$ and is L-Lipschitz on each permitted prediction interval. Its full-sample minimizer on the ball, denoted here by $f_n^\star$, satisfies $P_n(f - f_n^\star)^2 \le B' P_n(\ell_f - \ell_{f_n^\star})$ for all $f \in \mathcal{F}_\mu$. The bound balances the spectral cutoff against the square root of the remaining eigenvalue tail, with confidence term $x/N_{u,m}$ and constants depending only on M, L, B', μ. For

squared loss, $L = 2\sqrt{M}$ and $B' = 1$ suffice. If the ball contains all relevant zero-ridge subset solutions, then $D_0 \le 2\mu$ and $L_0 \le M$, and the learner and test oracle are preserved. For unbounded squared loss, (4.5) retains this spectral order when the diameter D_0 and residual maximum L_0 stay bounded, plus $2\varepsilon^2$ for positive ridge.

Proof roadmap for Theorem 4.3. Appendix D.2 gives the complete proof.

(1) Verify the ERM assumptions. The deterministic hull $\mathcal{F}_0^{\mathrm{sel}}$ has finite losses and contains the training, test, and full-sample subset solutions, preserving all required minima. For $L_0 > 0$, put $f_n^\star = \widehat{f}_{[n],0}$ and $\widetilde{g}_f = (\ell_f - \ell_{f_n^\star})/L_0$. The normal equation and residual control give $P_n\widetilde{g}_f = P_n(f - f_n^\star)^2/L_0$ and $\mathrm{Var}\,[\widetilde{g}_f] \le 4P_n\widetilde{g}_f$. Hence Assumption 1 holds for the normalized losses with $B = 4$ and $\theta = 1$. Lemma D.1 handles $L_0 = 0$ directly.

(2) Obtain the zero-ridge spectral bound. Lemma D.1 also proves (D.5) and an envelope bound of 2 for the normalized oracle excesses. Applying Theorem 4.2 and undoing normalization gives the zero-ridge case of Theorem D.3 on $\Omega_0(x)$. Corollary D.4 evaluates its fixed point using D_0, L_0, and $\widehat{\lambda}_j^{(0)}$. Since $\widehat{\lambda}_j^{(0)} \le \lambda_j(\mathbf{K}/n)$, replacing the projected eigenvalue tail by the original tail gives the spectral expression in (4.5). An EDR bound on the original spectrum therefore also suffices.

(3) Control the prediction change. If $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$, both training subset solutions are zero. Otherwise, (D.19), the reproducing property, and the ridge choice (4.4) bound every prediction change by ε.

(4) Transfer the bound to positive ridge. The test least-squares identity (D.15) gives

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \le 2\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,0}) + 2\varepsilon^2.$$

Inserting the zero-ridge spectral bound proves (4.5). Using its sharper projected-spectrum EDR bound proves (4.6). The comparison is deterministic, so both (4.5) and (4.6) hold on $\Omega_0(x)$ without a new probability bound at the selected ridge.

If we choose $\varepsilon = N_{u,m}^{-1}$, then the additional error from positive ridge, $2\varepsilon^2$ in (4.5) and (4.6), is $2N_{u,m}^{-2}$. For a fixed confidence parameter $x \ge 1$, two useful cases follow from (4.6). If the residual maximum L_0 and projected rank $\mathrm{rank}(\mathbf{K}_0)$ remain bounded as $N_{u,m}$ grows, then $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \lesssim N_{u,m}^{-1}$. If instead the projected eigenvalues satisfy the EDR assumption with fixed $\beta > 1/2$ and C_K, D_0, L_0 remain bounded, then $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \lesssim N_{u,m}^{-2\beta/(2\beta+1)}$. In both cases, the bound for $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}})$ holds with probability at least $1 - \exp(-x)$ over the random split at each sample size. These conditions make the excess-risk bound tend to zero without imposing an RKHS norm constraint on the learner.

Appendix D.5 expresses the diameter and residual maximum in terms of subset Gram matrices and explains how the bounds remain unchanged when the kernel and ridge are scaled together. Appendix D.6 illustrates how projection and cancellation of common losses sharpen the bounds.

4.4 Noisy Unbounded Transductive Kernel Learning

In this subsection, we apply UTLC to derive a sharp bound for the prediction error for noisy unbounded TKL, where the labels contain i.i.d. sub-Gaussian noise.

Let ρ be a probability measure on the Borel input space $\mathcal{X}$. Let K be measurable with separable RKHS $\mathcal{H}_K$ and $\sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x},\mathbf{x}) \le \kappa^2$ for some $\kappa > 0$. We write $K_{\mathbf{x}} = K(\cdot,\mathbf{x})$ and define the integral operator $T_\rho h := \int h(\mathbf{x})K_{\mathbf{x}}\,d\rho(\mathbf{x})$. Its ordered eigenvalues are ν_j, with zeros appended in the finite-rank case. Draw the full sample $\mathbf{X}_n = \{\mathbf{x}_i\}_{i=1}^n$ independently from ρ and, independently, choose the uniform split with $2 \le m \le u$ from $\mathbf{X}_n$. Let $\mathbf{K}$ be the Gram matrix of K over $\mathbf{X}_n$, and let $f^* \in \mathcal{H}_K$ be the true regression function. The i.i.d. label noises are independent of the inputs and split and have sub-Gaussian variance proxy $\sigma_0^2 \ge 0$:

$$y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \qquad \mathbb{E}\,[\varepsilon_i] = 0, \qquad \mathbb{E}\left[e^{t\varepsilon_i}\right] \le e^{\sigma_0^2 t^2/2} \quad (t \in \mathbb{R},\ i \in [n]). \tag{4.7}$$

We write $\Pr_{f^*}$ and $\mathbb{E}_{f^*}$ for probability and expectation over the inputs, random split, and label noise in (4.7). The prediction error is $\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f^*) = P_{\mathcal{U}}(\widehat{f} - f^*)^2$. The learner still minimizes $J_{\mathcal{M}}$ over $\mathcal{F} = \mathcal{H}_{\mathbf{X}_n}$ as in Section 4.3. Projecting f^* onto the span $\mathcal{H}_{\mathbf{X}_n}$ preserves all sample predictions and cannot increase its norm, as proved in Appendix D.3. The following theorem gives the sharp rate for the prediction error for noisy unbounded TKL.

Theorem 4.4 (Minimax Rate For Noisy Unbounded TKL)**.** In the above noisy unbounded TKL setup, suppose $\nu_j \le a_+ j^{-2\beta}$ for every $j \ge 1$, where $a_+ > 0$ and $\beta > 1/2$ are constants. Put $\alpha = 2\beta/(2\beta+1)$ and choose $\lambda_m = m^{-\alpha}$. Then there is a constant $C > 0$, depending only on β, a_+, κ, such that, for every fixed $f^* \in \mathcal{H}_K$ and every $u \ge m \ge 2$, (4.8) holds with probability at least $1 - 4\exp\left(-m^{1/(2\beta+1)}\right)$ over the inputs, random split, and label noise, and (4.9) holds in expectation:

$$\mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\lambda_m}, f^*) \le C\left(\|f^*\|_{\mathcal{H}_K}^2 + \sigma_0^2\right) m^{-\alpha}, \tag{4.8}$$

$$\mathbb{E}_{f^*}\left[\mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\lambda_m}, f^*)\right] \le C\left(\|f^*\|_{\mathcal{H}_K}^2 + \sigma_0^2\right) m^{-\alpha}. \tag{4.9}$$

For ordered eigenvalues $\widehat{\nu}_j = \lambda_j(\mathbf{K}/n)$, both bounds also hold with λ_m replaced by

$$\widehat{\lambda}_m = \frac{8\log m}{m} + \min_{Q\in\{0,\dots,n\}}\left[\frac{Q}{m} + \sqrt{\frac{1}{m}\sum_{j=Q+1}^{n}\widehat{\nu}_j}\right]. \tag{4.10}$$

Here (4.8) has probability at least $1 - 4m^{-8}$, and C may be enlarged.

Rule (4.10) uses only m and observed inputs, and both ridge choices, (4.10) and $\lambda_m = m^{-\alpha}$, give (4.8)–(4.9) for every fixed $f^* \in \mathcal{H}_K$, without the need for a finite upper bound for the diameter D_0 or residual maximum L_0. The constant C depends only on β, a_+, κ.

We remark that the empirical loss minimizer, $\widehat{f}_{m,\lambda_m}$, is obtained by optimization over the unbounded RKHS set $\mathcal{F} = \mathcal{H}_{\mathbf{X}_n}$. As a result, all existing local-complexity-based transductive learning analyses of bounded loss (Tolstikhin et al., 2014; Yang, 2025a;b; 2026) are not applicable for both unbounded TKL in Section 4.3 and noisy unbounded TKL in this subsection.

Proof roadmap for Theorem 4.4 is presented below. The proof applies UTLC Theorem 4.1, not Theorem 4.3.

(1) Verify the localized TC condition. For $h_v(i) = v(\mathbf{x}_i)^2/\kappa^2$, $\|v\|_{\mathcal{H}_K} \le 1$, use $V(h) = P_n h$. TC comparison and contraction (Theorems A.3 and A.2) and centering (Lemma A.1) give (4.1) with majorant (D.32).

(2) Apply UTLC uniformly. The one-sided Theorem 4.1 with $A = 2$ and homogeneity give (D.33) for every $d \in \mathcal{H}_K$. The cutoff (4.10) controls its fixed point and the training effective dimension.

(3) Control ridge error and match rates. The ridge equation (D.35) and noise concentration give (D.39). Population EDR and a fourth-moment argument give (4.8)–(4.9).

For $\mu_0 = \Theta(1) > 0$, we write $\mathcal{H}_K(\mu_0) = \left\{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \le \mu_0\right\}$. We now explain that the prediction error bound (4.8) of Theorem 4.4 is minimax optimal.

Minimax Optimality. Under the additional lower EDR $\nu_j \ge a_- j^{-2\beta}$ for some constant $a_- > 0$ and Gaussian noise $N(0, \sigma_0^2)$ with $\sigma_0 > 0$, Theorem D.5 in Appendix D.4 gives a matching high-probability minimax lower bound of $\Omega(m^{-\alpha})$ over $\mathcal{H}_K(\mu_0)$ for the prediction error. Together with the high-probability upper bound of $\mathcal{O}(m^{-\alpha})$ for the prediction error in (4.8), this establishes minimax optimality of the prediction-error rate $m^{-\alpha}$ in high probability with $u \ge m$, proved by (4.8) of Theorem 4.4.

5 Conclusion

We introduce UTLC, which extends transductive local-complexity analysis to unbounded losses with potentially infinite envelopes through a novel variance rescaling method. Under the standard localization and geometric conditions, UTLC yields ERM excess-risk bounds and spectral excess-risk bounds for unbounded TKL without imposing an RKHS norm constraint. For noisy unbounded TKL, UTLC with a data-based ridge choice achieves minimax optimal prediction error on fixed RKHS signal balls under a two-sided polynomial population EDR and positive Gaussian noise, while the learner's optimization domain remains unbounded. Bounded-loss STLC rates in the current state-of-the-art are recovered as special cases of UTLC under their original assumptions.

AI USE STATEMENT

In this work, we used generative AI tools for interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, and generating synthetic data sets, implementing methods, cleaning and reformatting dataset, assisting with translation, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for improving readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8. URL https://doi.org/10.1007/s10208-006-0196-8.

Corinna Cortes and Mehryar Mohri. On transductive regression. In *Advances in Neural Information Processing Systems*, NIPS'06, pp. 305–312, Cambridge, MA, USA, 2006. MIT Press.

Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability analysis and learning bounds for transductive regression algorithms, 2009. URL https://arxiv.org/abs/0904.0814.

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher complexity and its applications. *J. Artif. Int. Res.*, 35(1):193–234, jun 2009.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 9.1–9.24. PMLR, 2012. URL https://proceedings.mlr.press/v23/hsu12.html.

Ilya Tolstikhin, Nikita Zhivotovskiy, and Gilles Blanchard. Permutational Rademacher complexity: A new complexity measure for transductive learning. In *Algorithmic Learning Theory*, volume 9355 of *Lecture Notes in Computer Science*, pp. 209–223. Springer, 2015. doi: 10.1007/978-3-319-24486-0_14.

Ilya O. Tolstikhin, Gilles Blanchard, and Marius Kloft. Localized complexities for transductive learning. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Conference on Learning Theory*, volume 35 of *JMLR Workshop and Conference Proceedings*, pp. 857–884. JMLR.org, 2014.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1982. ISBN 0-387-90733-5. Translated from the Russian by Samuel Kotz.

Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. ISBN 0-471-03003-1. A Wiley-Interscience Publication.

Yingzhen Yang. A new concentration inequality for sampling without replacement and its application for transductive learning. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=NRVdvg7VMn.

Yingzhen Yang. Improved generalization bounds for transductive learning by transductive local complexity and its applications. arXiv preprint arXiv:2309.16858, 2025b. URL https://arxiv.org/abs/2309.16858v4.

Yingzhen Yang. Even sharper bounds for transductive learning and its applications. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. URL https://arxiv.org/pdf/2609.28459.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2. URL https://doi.org/10.1007/s00365-006-0663-2.

Guide to the appendix. The proofs follow the order of the main results, with the required lemmas established before each theorem proof. Sections A–C develop the general UTLC and ERM arguments. Section D applies them to fixed-label and noisy TKL. The following map covers the four main-text theorems and their supporting claims and comparisons.

(1) Section A supplies the common mathematical tools. Lemma A.1 controls centering in inductive RC, Theorem A.2 verifies contraction, and Theorem A.3 compares TC with inductive RC. Lemma A.4 explains the role of the localized TC condition (4.1). Lemma A.5 proves the fixed-point existence and finiteness used in Section 4.1.

(2) Section B proves Theorem 4.1. Section B.1 first verifies the cited finite-envelope concentration input, Theorem B.1. Section B.2 then proves the variance-rescaling Lemma B.2 and uses it to obtain both the one-sided and two-sided UTLC bounds. Section B.3 proves Lemma B.3, which justifies the bounded-loss STLC and ERM comparisons in the main text and Table 2.

(3) Section C proves Theorem 4.2. Section C.1 establishes fixed-point finiteness, the oracle decomposition, and nonlinear absorption in Lemmas C.1, C.2, and C.3. Section C.2 then proves (C.1) and the explicit specialization (4.3) for every $0 < \theta \leq 1$. Section C.3 verifies the three sufficient Bernstein geometries in Lemma C.4 and the infinite-envelope squared-loss construction in Example C.5, as cited after Assumption 1.

(4) Section D.1 establishes the deterministic foundation for Section 4.3. It defines the convex hull of subset solutions and proves Lemma D.1, including the representer formulas, minimum-norm selection, convergence of gradient descent from zero, projected-kernel order and rank properties, Bernstein geometry, and the localized TC bound (D.5). These results supply the hypotheses needed to apply the generic ERM theorem.

(5) Section D.2 proves Theorem 4.3 through four consecutive arguments. Section D.2.1 proves the deterministic risk and ridge comparisons in Lemma D.2. Section D.2.2 applies the preceding localization and the parameterized ERM bound (C.1) to prove Theorem D.3. Section D.2.3 evaluates its fixed point in Corollary D.4. Finally, Section D.2.4 transfers the zero-ridge spectral bound to the training-label ridge choice, proving (4.5) and (4.6), including the zero training Gram matrix case and the stated uniformity over prediction tolerances.

(6) Section D.3 proves Theorem 4.4. It derives empirical spectral-tail bounds from population EDR, verifies both localized TC bounds for normalized squared predictions, and applies UTLC uniformly over prediction differences. The ridge normal equation and a sub-Gaussian quadratic-form bound give (4.8). A fourth-moment argument proves (4.9). The section also verifies the critical population rate and a concrete Fourier-kernel example. Its final step proves that the data-based ridge (4.10) attains the same rate without the population EDR exponent.

(7) Section D.4 states and proves Theorem D.5, supporting the minimax comparison in Section 4.4. Under its population EDR and Gaussian-noise assumptions, the proof constructs separated signals, bounds their distinguishability from training labels, and transfers the separation to the test inputs. It then obtains the worst-case fixed-input lower bound for bounded RKHS signal sets.

(8) Section D.5 proves Lemma D.6 and the exact subset-matrix formulas for the diameter and residual maximum cited in Section 4.3. It also verifies invariance of the subset solutions and bounds under joint rescaling of the kernel and ridge.

(9) Section D.6 verifies the kernel comparisons cited in Section 4.3. The examples isolate the gains from projection and cancellation of common losses, and show why finite scales of the subset solutions alone do not ensure vanishing excess risk.

(10) Section D.7 proves the bounded-TKL specialization (D.49), including the squared-loss case and preservation of the learner and test oracle when the ball contains all relevant zero-ridge subset solutions. This supplies the bounded-loss comparison with Yang (2026, Theorem 5.2) in Section 4.3 and Table 2.

In the proofs below, circled numbers identify derivation steps and their accompanying explanations. The numbering starts at one in each proof and restarts at one after nine. Repeated numbers refer only to their local calculation.

A MATHEMATICAL TOOLS

We collect the tools used in Sections B–D. The averages, TC, and centered squares are defined in Section 2. For an integer $p \geq 1$, let $J = (J_1, \ldots, J_p)$ have independent uniform coordinates on $[n]$. Independently of J, let $\sigma = (\sigma_1, \ldots, \sigma_p)$ consist of independent Rademacher variables, so $\Pr\{\sigma_j = 1\} = \Pr\{\sigma_j = -1\} = 1/2$. For a nonempty class $\mathcal{H}$ of real-valued functions on $[n]$, define its inductive Rademacher complexity by $\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}) = \mathbb{E}_{J,\sigma}\left[\sup_{h\in\mathcal{H}} p^{-1}\sum_{j=1}^{p}\sigma_j h(J_j)\right]$. The supremum contains no absolute value. Theorem A.3 compares both TC orientations with this RC. Lemma A.1 controls the additional centering operation. Lemma A.4 explains why the common TC majorant condition (4.1) is a separate hypothesis.

Lemma A.1 (Centering And Inductive Rademacher Complexity)**.** Let $1 \leq p \leq n$ be an integer and let $\mathcal{H} \subset \mathbb{R}^{[n]}$ be nonempty. For the inductive RC defined above,

$$\mathfrak{R}_p^{(\mathrm{ind})}\left(\{h - P_n h : h \in \mathcal{H}\}\right) \leq 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}). \tag{A.1}$$

Proof of Lemma A.1. Both RCs in (A.1) use independent uniform indices $J_1, \ldots, J_p$ on $[n]$ and independent Rademacher variables $\sigma_1, \ldots, \sigma_p$, independent of J.

Proof outline.

(1) Ghost sample. Express the full-sample mean as an expectation over independent indices and move that expectation outside the supremum.
(2) Separate the samples. Bound the supremum of a difference by two suprema and use the symmetry of the Rademacher vector.

Step 1: Ghost sample. The centering term is deterministic, so we first represent it by an independent sample that can be compared with the observed sample. Fix $h_0 \in \mathcal{H}$. Every process supremum is bounded below by its value at h_0, which is bounded on the finite sampling space. Its expectation is therefore well defined, possibly equal to $+\infty$. Evaluating the raw RC supremum at h_0 and averaging the independent Rademacher variables gives $\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}) \geq \mathbb{E}_{J,\sigma}\left[p^{-1}\sum_j \sigma_j h_0(J_j)\right] = 0$. Its expectation therefore lies in $[0, \infty]$, although the supremum need not be nonnegative at each realization. If it is infinite, the claim holds in the extended-real sense. Suppose henceforth that it is finite. For any $i \in [n]$, the events $J_1 = \cdots = J_p = i$, $\sigma_1 = \cdots = \sigma_p = 1$ and $J_1 = \cdots = J_p = i$, $\sigma_1 = \cdots = \sigma_p = -1$ each have probability $(2n)^{-p}$ over (J, σ). The corresponding raw suprema are $\sup_{h\in\mathcal{H}} h(i)$ and $\sup_{h\in\mathcal{H}}(-h(i))$. Finiteness of the RC therefore implies $\sup_{h\in\mathcal{H}} |h(i)| < \infty$ for every coordinate. All subsequent suprema and expectations are consequently finite.

Let $J' = (J'_1, \ldots, J'_p)$ be a copy of J, independent of (J, σ). For fixed (J, σ) and every $h \in \mathcal{H}$, $\mathbb{E}_{J'}\left[\sum_j \sigma_j h(J'_j)\right] = (\sum_j \sigma_j) P_n h$. It follows that

$$\begin{aligned}\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_j \sigma_j (h(J_j) - P_n h) &\overset{①}{=} \sup_{h\in\mathcal{H}} \mathbb{E}_{J'}\left[\frac{1}{p}\sum_j \sigma_j (h(J_j) - h(J'_j))\right] \\ &\overset{②}{\leq} \mathbb{E}_{J'}\left[\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_j \sigma_j (h(J_j) - h(J'_j))\right].\end{aligned}$$

Step ① replaces $P_n h$ by the ghost expectation. Step ② bounds the supremum of expectations by the expectation of the supremum.

Averaging over (J, σ) bounds the left side of (A.1).

Step 2: Separate the samples. Separating the two sample suprema recovers two copies of the original complexity:

$$\begin{aligned}\mathfrak{R}_p^{(\mathrm{ind})}\left(\{h - P_n h : h \in \mathcal{H}\}\right) &\overset{③}{\leq} \mathbb{E}_{J,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_j \sigma_j h(J_j)\right] + \mathbb{E}_{J',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_j (-\sigma_j) h(J'_j)\right] \\ &\overset{④}{=} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}).\end{aligned}$$

Here ③ separates the two sample suprema. ④ uses the identical laws of (J, σ) and $(J', -\sigma)$.

This proves (A.1) without requiring symmetry of $\mathcal{H}$. □

Theorem A.2 (Contraction Property Of Inductive Rademacher Complexity (Yang, 2026, Theorem A.1))**.** Let $\mathcal{H}$ be a nonempty class of real-valued functions on $[n]$. Suppose that, for some $L \geq 0$ and every $i \in [n]$, the map $\phi_i \colon \mathbb{R} \to \mathbb{R}$ satisfies $\phi_i(0) = 0$ and $|\phi_i(a) - \phi_i(b)| \leq L\,|a - b|$ for all $a, b \in \mathbb{R}$. Define $\phi \circ \mathcal{H} \coloneqq \{i \mapsto \phi_i\,(h(i)) : h \in \mathcal{H}\}$. If $L > 0$, then $\mathfrak{R}_p^{(\mathrm{ind})}(\phi \circ \mathcal{H}) \leq L\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H})$ for every $p \in \{u, m\}$. When $L = 0$, the transformed class consists only of zero and its RC is zero. The usual scalar contraction inequality is recovered by taking $\phi_i = \phi$ for every $i \in [n]$.

Verification Of Theorem A.2. We verify the cited contraction inequality with the convention used here: the Rademacher supremum has no absolute value. Fix $p \in \{m, u\}$. The indices $J_1, \ldots, J_p$ are independent and uniform on $[n]$, and the independent Rademacher variables $\sigma_1, \ldots, \sigma_p$ are independent of them.

Proof outline.

(1) One-coordinate comparison. Condition on the indices and all but one Rademacher variable, and use the Lipschitz inequality between two class members.
(2) Iteration and averaging. Replace the coordinate maps one at a time, then average over the indices and restore the factor $1/p$.

Step 1: One-coordinate comparison. We average one Rademacher variable to compare the transformed and original coordinates. If $L = 0$, the conditions $\phi_i(0) = 0$ and zero Lipschitz constant give $\phi_i = 0$ for every i, proving the assertion. For $L > 0$, an infinite RC on the right makes the assertion automatic. Otherwise, the finite-RC argument in Step 1 of Lemma A.1 shows that $\mathcal{H}$ is bounded on every coordinate. All suprema below are therefore finite.

Condition on J and fix a coordinate $k \in \{1, \ldots, p\}$. Hold the other Rademacher variables fixed. Write A_h for the sum of their contributions for $h \in \mathcal{H}$. Each such contribution may use either its original coordinate map or multiplication by L. It is independent of σ_k. Averaging over the two values of σ_k gives

$$\begin{aligned}
\mathbb{E}_{\sigma_k}\left[\sup_{h\in\mathcal{H}}\left[A_h + \sigma_k\phi_{J_k}(h(J_k))\right]\right] &\overset{①}{=} \frac{1}{2}\sup_{h,h'\in\mathcal{H}}\left[A_h + A_{h'} + \phi_{J_k}(h(J_k)) - \phi_{J_k}(h'(J_k))\right] \\
&\overset{②}{\leq} \frac{1}{2}\sup_{h,h'\in\mathcal{H}}\left[A_h + A_{h'} + L\,|h(J_k) - h'(J_k)|\right] \\
&\overset{③}{=} \frac{1}{2}\sup_{h,h'\in\mathcal{H}}\left[A_h + A_{h'} + L(h(J_k) - h'(J_k))\right] \overset{④}{=} \mathbb{E}_{\sigma_k}\left[\sup_{h\in\mathcal{H}}\left[A_h + \sigma_k L h(J_k)\right]\right].
\end{aligned}$$

Steps ① and ④ average over $\sigma_k = \pm 1$ and separate or recombine the suprema. Approximating maximizers justify these identities without attainment. Step ② uses Lipschitz continuity. In ③, exchanging h, h' leaves $A_h + A_{h'}$ unchanged, so the two signed suprema coincide.

Step 2: Iteration and averaging. Average over the remaining Rademacher variables and repeat the comparison for $k = 1, \ldots, p$. Since A_h allows already replaced coordinates, this gives, for every J,

$$\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{k=1}^{p}\sigma_k\phi_{J_k}(h(J_k))\right] \overset{⑤}{\leq} L\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{k=1}^{p}\sigma_k h(J_k)\right].$$

Here ⑤ iterates the conditional one-coordinate comparison.

Dividing by p and averaging over J proves the claim. Repeated indices are allowed because the conditional comparison uses only Rademacher independence. □

The following theorem relates transductive complexity to its with-replacement counterpart.

Theorem A.3 ((Yang, 2025a, Theorem 2.1), (Yang, 2025b, Theorem 3.1))**.** Let $\mathcal{H}$ be a nonempty class of real-valued functions on $[n]$. Then $\max\{\mathfrak{R}_u^+(\mathcal{H}), \mathfrak{R}_u^-(\mathcal{H})\} \leq 2\mathfrak{R}_u^{(\mathrm{ind})}(\mathcal{H})$ and $\max\{\mathfrak{R}_m^+(\mathcal{H}), \mathfrak{R}_m^-(\mathcal{H})\} \leq 2\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{H})$.

Verification Of Theorem A.3. Fix $p \in \{m, u\}$ and let A_p be a uniform p-subset. Recall that

$$\mathfrak{R}_p^+(\mathcal{H}) = \mathbb{E}_{A_p}\left[\sup_{h\in\mathcal{H}}(P_{A_p} - P_n)h\right].$$

We verify the sampling comparison and symmetrization used in the cited results, keeping the same class $\mathcal{H}$ throughout.

Proof outline.

(1) Sampling comparison. Apply the finite-population convex comparison to a finite class, then pass to a dense subclass.
(2) Ghost-sample symmetrization. Introduce independent indices, randomize the differences by exchanging index pairs, and separate the two suprema. Reflection gives the negative orientation.

Step 1: Sampling comparison. To use an inductive complexity, we first replace sampling without replacement by sampling with replacement through convexity. If $\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}) = \infty$, there is nothing to prove. Otherwise $\mathcal{H}$ is bounded on each coordinate, by the argument in Step 1 of Lemma A.1. First take a finite subclass $\{h_1, \ldots, h_M\}$. Associate with each population index i the vector $\mathbf{c}_i = (h_l(i) - P_n h_l)_{l=1}^M \in \mathbb{R}^M$. For $\mathbf{v} = (v_1, \ldots, v_M) \in \mathbb{R}^M$, the function $\mathbf{v} \mapsto \max_{l\le M} v_l$ is continuous and convex. The vector form of Hoeffding's sampling comparison (Yang, 2025a, Theorem A.6) therefore gives

$$\mathbb{E}_{A_p}\left[\max_{l\le M}(P_{A_p} - P_n)h_l\right] \overset{①}{\le} \mathbb{E}_J\left[\max_{l\le M}\left[\frac{1}{p}\sum_{k=1}^p h_l(J_k) - P_n h_l\right]\right].$$

Step ① compares averages of the same finite population $(\mathbf{c}_i)_{i=1}^n$ sampled without and with replacement. Here $J_1, \ldots, J_p$ are independent uniform indices.

Equip $\mathcal{H}$ with the metric induced by the evaluation map $h \mapsto (h(i))_{i=1}^n \in \mathbb{R}^n$. Its image has a countable dense subset. Choose corresponding functions $(h_l)_{l\ge1}$ in $\mathcal{H}$ and apply the displayed comparison to the first M functions. For a finite dense subset, repeat its members in this enumeration. Each subset or signed-sum evaluation is continuous in these sample values. As a result, the increasing maxima tend to the corresponding supremum over $\mathcal{H}$ as $M \to \infty$. The sampling spaces are finite and the coordinate bounds are uniform over $\mathcal{H}$, so we may take these limits inside their expectations. We have proved

$$\mathfrak{R}_p^+(\mathcal{H}) \overset{②}{\le} \mathbb{E}_J\left[\sup_{h\in\mathcal{H}}\left[\frac{1}{p}\sum_{k=1}^p h(J_k) - P_n h\right]\right].$$

Here ② passes to the dense subclass limit on the finite sampling space.

Step 2: Ghost-sample symmetrization. An independent copy turns the mean deviation into symmetric differences, to which Rademacher variables can be attached. Let $J' = (J'_1, \ldots, J'_p)$ be an independent copy of J, and let $\sigma_1, \ldots, \sigma_p$ be independent Rademacher variables, independent of both samples. For each h, the mean of $p^{-1}\sum_k h(J'_k)$ is $P_n h$. Moving this expectation outside the supremum gives

$$\begin{aligned}\mathfrak{R}_p^+(\mathcal{H}) &\overset{③}{\le} \mathbb{E}_{J,J'}\left[\sup_{h\in\mathcal{H}}\frac{1}{p}\sum_{k=1}^p(h(J_k) - h(J'_k))\right] \overset{④}{=} \mathbb{E}_{J,J',\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{p}\sum_{k=1}^p\sigma_k(h(J_k) - h(J'_k))\right] \\ &\overset{⑤}{\le} \mathbb{E}_{J,\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{p}\sum_{k=1}^p\sigma_k h(J_k)\right] + \mathbb{E}_{J',\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{p}\sum_{k=1}^p(-\sigma_k)h(J'_k)\right] \overset{⑥}{=} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}).\end{aligned}$$

Step ③ moves the ghost expectation outside the supremum. For ④, exchange J_k, J'_k whenever $\sigma_k = -1$. Independence preserves the joint law of these pairs. Step ⑤ separates the suprema, and ⑥ uses the identical laws of (J, σ) and $(J', -\sigma)$.

For the other orientation, define the reflected class $-\mathcal{H} = \{-h : h \in \mathcal{H}\}$. Then $\mathfrak{R}_p^-(\mathcal{H}) = \mathfrak{R}_p^+(-\mathcal{H})$ and $\mathfrak{R}_p^{(\mathrm{ind})}(-\mathcal{H}) = \mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H})$ by Rademacher symmetry. Applying the positive comparison to $-\mathcal{H}$ proves the negative one. Neither argument inserts an absolute value into the RC supremum or assumes that $\mathcal{H}$ is symmetric. □

Lemma A.4 (Scalar Multiples And Sub-Root Majorants)**.** Let $\mathbf{z} \in \mathbb{R}^n$ satisfy $P_n\mathbf{z} = 0$ and $P_n\mathbf{z}^2 = 1$, let $1 \le p < n$ be an integer, and set $\mathcal{H} = \{t\mathbf{z} : t \ge 0\}$. For the quadratic surrogate $V_2(t\mathbf{z}) = t^2$, write $\mathcal{H}_2(r) = \{t\mathbf{z} : 0 \le t \le \sqrt{r}\}$ and define

$$\alpha_+ = \mathbb{E}_{A_p}\left[(P_{A_p}\mathbf{z})_+\right], \quad \alpha_- = \mathbb{E}_{A_p}\left[(-P_{A_p}\mathbf{z})_+\right], \quad \beta = \mathbb{E}_{A_p}\left[\left((P_{A_p} - P_n)\mathbf{z}^2\right)_+\right].$$

Here A_p is a uniform p-subset and $a_+ = \max\{a, 0\}$. Then, for $r \ge 0$ and $\eta \in \{+, -\}$,

$$\mathfrak{R}_p^{\eta}(\mathcal{H}_2(r)) = \alpha_\eta\sqrt{r}, \qquad \mathfrak{R}_p^{+}(\mathcal{H}_2(r)^{[2]}) = \beta r.$$

If $\mathbf{z}^2$ is nonconstant, then $\beta > 0$, so these two TC quantities admit no common finite sub-root majorant on all radii. If $\mathbf{z}^2$ is constant, then $\beta = 0$ and $(\max_\eta \alpha_\eta)\sqrt{r}$ is such a majorant. For the larger surrogate $V_4(t\mathbf{z}) = t^2 + t^4$, a common majorant always exists and may be taken as $\max\{\alpha_+, \alpha_-, \beta\}\sqrt{r}$.

Proof of Lemma A.4. The vector $\mathbf{z}$ is centered and has empirical variance one. As a result, $\mathrm{Var}[t\mathbf{z}] = t^2$, so both V_2 and V_4 are finite surrogate localization functionals on $\mathcal{H} = \{t\mathbf{z} : t \ge 0\}$. Write $\mathcal{H}_j(r) = \{h \in \mathcal{H} : V_j(h) \le r\}$ for $j = 2, 4$. In $\mathcal{H}_j(r)^{[2]}$, localization by $V_j \le r$ precedes centering and squaring. The constants $\alpha_+, \alpha_-, \beta$ are the expected positive parts of the subset deviations defined in the statement.

Proof outline.

(1) Exact quadratic localization. Maximize the linear and squared subset deviations over the permitted scalar interval.
(2) Positive square deviation. Compare two subsets whose squared averages differ and use the zero mean of a subset deviation.
(3) Larger surrogate. Use the two scalar constraints supplied by $V_4 \le r$ to bound both TC quantities by a multiple of $\sqrt{r}$.

Step 1: Exact quadratic localization. For each fixed subset, localization reduces to maximizing over $0 \le t \le \sqrt{r}$. The positive linear deviation is $tP_{A_p}\mathbf{z}$, whose supremum satisfies $\sup_{0\le t\le\sqrt{r}} tP_{A_p}\mathbf{z} \overset{①}{=} \sqrt{r}(P_{A_p}\mathbf{z})_+$. Here ① maximizes a linear function at the appropriate endpoint of $[0, \sqrt{r}]$. Reversing the sign replaces this positive part by $(-P_{A_p}\mathbf{z})_+$. Since $t\mathbf{z}$ is centered, its square is $t^2\mathbf{z}^2$. Its positive subset deviation satisfies $\sup_{0\le t\le\sqrt{r}} t^2(P_{A_p} - P_n)\mathbf{z}^2 \overset{②}{=} r((P_{A_p} - P_n)\mathbf{z}^2)_+$. Here ② maximizes over $t^2 \in [0, r]$. Taking expectations proves both identities, including $r = 0$. All constants are finite on the finite subset space.

Step 2: Positive square deviation. We show that a nonconstant squared ray has positive TC growing linearly in the radius. Suppose $z_i^2 \ne z_j^2$ for two indices. There exists a set $B \subseteq [n] \setminus \{i, j\}$ of size $p - 1$, since $p \le n - 1$. The squared averages over $B \cup \{i\}$ and $B \cup \{j\}$ differ by $(z_i^2 - z_j^2)/p \ne 0$. Hence $(P_{A_p} - P_n)\mathbf{z}^2$ is nonconstant. Its expectation is zero, because each index has inclusion probability p/n under the uniform choice of A_p. A nonconstant, mean-zero random variable on a finite probability space takes a positive value with positive probability. Therefore $\beta > 0$. For a finite sub-root function ψ and any fixed $r_0 > 0$, $\psi(r) \overset{③}{\le} \psi(r_0)\sqrt{r/r_0}$ for $r \ge r_0$. Here ③ uses the nonincreasing sub-root ratio at r_0 and r. This cannot dominate βr as $r \to \infty$. If $\mathbf{z}^2$ is constant, its subset deviation is zero on every subset. The first identity in Step 1 then proves the stated majorant.

Step 3: Larger surrogate. Adding a fourth-order term restricts the squared multiplier strongly enough to restore square-root growth for both complexities. For $r \ge 0$, every multiplier allowed by $t^2 + t^4 \le r$ satisfies $t \le \sqrt{r}$ and $t^2 \le \sqrt{r}$. Applying the endpoint bounds from Step 1 therefore gives

$$\mathfrak{R}_p^{\eta}(\mathcal{H}_4(r)) \overset{④}{\le} \alpha_\eta\sqrt{r}, \qquad \mathfrak{R}_p^{+}(\mathcal{H}_4(r)^{[2]}) \overset{⑤}{\le} \beta\sqrt{r}.$$

Steps ④–⑤ use the bounds on t and t^2, respectively. Their maximum proves the claim, including $r = 0$. □

The next lemma records the standard sub-root fixed-point property (Bartlett et al., 2005, Lemma 3.2) and an explicit finite upper bound. It applies directly to the majorant in Theorem 4.1.

Lemma A.5 (Finite Fixed Points Of Sub-Root Functions)**.** Let $\psi : [0,\infty) \to [0,\infty)$ be a finite-valued sub-root function. If ψ is not identically zero, it has a unique positive fixed point r^* satisfying $\psi(r^*) = r^*$. If $\psi \equiv 0$, set $r^* = 0$. For every $r_0 > 0$,

$$0 \le r^* = \inf\{r > 0 : \psi(r) \le r\} \le \max\left\{r_0, \frac{\psi(r_0)^2}{r_0}\right\} < \infty. \tag{A.2}$$

For every $r > 0$, $\psi(r) \le r$ if and only if $r \ge r^*$. Moreover, $\psi(r) \le \sqrt{rr^*} \le r$ whenever $r \ge r^*$.

Proof of Lemma A.5. Recall that ψ is finite and nondecreasing, and $\psi(r)/\sqrt{r}$ is nonincreasing on $(0,\infty)$. If $\psi \equiv 0$, every conclusion follows with $r^* = 0$. We henceforth consider the nonzero case.

Proof outline.

(1) Continuity. Combine the two monotonicity properties.
(2) Existence and uniqueness. Compare ψ with the identity and locate its unique crossing.
(3) Explicit fixed-point bound. Compare with a finite reference value.

Step 1: Continuity. For $0 < a < b$, the defining properties give

$$\psi(a) \le \psi(b) \overset{①}{\le} \sqrt{b/a}\,\psi(a).$$

Here ① uses the nonincreasing sub-root ratio. Letting either endpoint approach the other proves continuity on $(0,\infty)$ by squeezing. Continuity at zero is not needed.

Step 2: Existence and uniqueness. Choose $s > 0$ with $c = \psi(s)/\sqrt{s} > 0$. The sub-root property gives $\psi(r) \ge c\sqrt{r} > r$ for $0 < r < \min\{s, c^2\}$ and $\psi(r) \le c\sqrt{r} < r$ for $r > \max\{s, c^2\}$. Continuity therefore gives a positive fixed point r^*. Comparison of the sub-root ratio with its value $\sqrt{r^*}$ at that point gives

$$\psi(r) \overset{②}{\ge} \sqrt{rr^*} > r \quad (0 < r < r^*), \qquad \psi(r) \overset{③}{\le} \sqrt{rr^*} < r \quad (r > r^*).$$

Steps ②–③ use the ratio on the two sides of r^*. These inequalities prove uniqueness, the identity $\{r > 0 : \psi(r) \le r\} = [r^*, \infty)$, and the claimed bound for $r \ge r^*$. For $\psi \equiv 0$, the same set is $(0,\infty)$ and its infimum is zero.

Step 3: Explicit fixed-point bound. Fix $r_0 > 0$. If $r^* \le r_0$, the bound is immediate. Otherwise,

$$\sqrt{r^*} = \frac{\psi(r^*)}{\sqrt{r^*}} \overset{④}{\le} \frac{\psi(r_0)}{\sqrt{r_0}}, \qquad r^* \le \frac{\psi(r_0)^2}{r_0}.$$

Step ④ again uses the sub-root ratio. The two cases prove (A.2). □

The finiteness in (A.2) uses the finite value $\psi(r_0)$. It does not assert that a majorant satisfying the localized TC condition (4.1) exists for every class.

B Proof of Theorem 4.1

B.1 Finite-envelope concentration

For $0 \in \mathcal{H}$, put $v(\mathcal{H}) = (\sup_{h\in\mathcal{H}} \mathrm{Var}\,[h])^{1/2}$ and $c_{\mathrm{split}} = n/\max\{u, m\} \in (1, 2]$. Centering leaves the split process unchanged and replaces the raw second-moment bound by $v^2(\mathcal{H})$. The following cited finite-envelope input is not itself an unbounded-class extension.

Theorem B.1 (Centered Test-Train Concentration Under A Finite Envelope, The Centered Form Of The Cited STLC Theorem (Yang, 2026, Theorem 4.1))**.** Let $0 \in \mathcal{H} \subset \mathbb{R}^{[n]}$ be a nonempty function class with $B_\infty(\mathcal{H}) < \infty$. Set $c_0 := \sqrt{219}$ and choose $\eta = +$ if $u \le m$, and $\eta = -$ if $m < u$. Then, for every $x > 0$, with probability at least $1 - \exp(-x)$ over the random split,

$$g(\mathcal{H}) \le c_{\mathrm{split}}\mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{H}) + c_0 v(\mathcal{H})\sqrt{\frac{x}{N_{u,m}}} + \frac{c_0}{2}\mathfrak{R}^{+}_{N_{u,m}}(\mathcal{H}^{[2]}) + (16B_\infty(\mathcal{H}) + c_0/2)\,\frac{x}{N_{u,m}}. \tag{B.1}$$

Furthermore, for every split, $g(\mathcal{H}) \le c_{\mathrm{split}} B_\infty(\mathcal{H}) \le 2B_\infty(\mathcal{H})$.

Verification Of Theorem B.1. Recall $v^2(\mathcal{H}) = \sup_{h\in\mathcal{H}} \mathrm{Var}\,[h]$ and $c_{\mathrm{split}} = n/\max\{u,m\}$.

Proof outline.

(1) Centering. Apply the cited concentration theorem to centered functions.
(2) Split orientation. Identify the expected supremum through the smaller subset.
(3) Deterministic envelope bound. Bound the same subset average by its envelope.

Step 1: Centering. If $v(\mathcal{H}) = 0$, every $h \in \mathcal{H}$ is constant on $[n]$, so the split process and both TC terms vanish. Both conclusions follow directly. Otherwise, put $\mathcal{H}_c = \{h - P_n h : h \in \mathcal{H}\}$. Cancellation of constants gives

$$g(\mathcal{H}_c) = g(\mathcal{H}), \qquad \sup_{z\in\mathcal{H}_c} P_n z^2 = v^2(\mathcal{H}), \qquad \sup_{z\in\mathcal{H}_c} \max_i |z(i)| = B_\infty(\mathcal{H}).$$

The class contains zero, its ordinary square class is $\mathcal{H}^{[2]}$, and $0 < v(\mathcal{H}) \le B_\infty(\mathcal{H}) < \infty$. Consequently, Yang (2026, Theorem 4.1) applies with $H_0 = B_\infty(\mathcal{H})$ and second-moment bound $v^2(\mathcal{H})$. It gives, with probability at least $1 - e^{-x}$ over the random split,

$$g(\mathcal{H}) \overset{①}{\le} \mathbb{E}\,[g(\mathcal{H})] + c_0 v(\mathcal{H})\sqrt{\frac{x}{N_{u,m}}} + \frac{c_0}{2}\mathfrak{R}^{+}_{N_{u,m}}(\mathcal{H}^{[2]}) + (16B_\infty(\mathcal{H}) + c_0/2)\,\frac{x}{N_{u,m}}.$$

Step ① is the cited theorem applied to $\mathcal{H}_c$. No class supremum needs to be attained.

Step 2: Split orientation. To identify the expectation, use $P_n = (u/n)P_{\mathcal{U}} + (m/n)P_{\mathcal{M}}$:

$$g(h) \overset{②}{=} \frac{n}{m}(P_{\mathcal{U}} - P_n)h = \frac{n}{u}(P_n - P_{\mathcal{M}})h.$$

Step ② substitutes the full-sample mean. For $u \le m$, use the first expression, with coefficient c_{split} and smaller subset $\mathcal{U}$. For $m < u$, use the second expression and $\mathcal{M}$. Taking the supremum and expectation gives $\mathbb{E}\,[g(\mathcal{H})] = c_{\mathrm{split}}\mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{H})$, with the stated orientation. Substitution proves (B.1).

Step 3: Deterministic envelope bound. Set $z_h = h - P_n h$. The identities in Step 2 and $|z_h(i)| \le B_\infty(\mathcal{H})$ give, for every split,

$$|g(h)| \overset{③}{\le} c_{\mathrm{split}} \begin{cases} P_{\mathcal{U}}|z_h|, & u \le m, \\ P_{\mathcal{M}}|z_h|, & m < u \end{cases} \le c_{\mathrm{split}} B_\infty(\mathcal{H}).$$

Step ③ is the triangle inequality for the smaller-subset average. Taking the supremum and using $c_{\mathrm{split}} \le 2$ completes the proof. □

B.2 Variance rescaling and the UTLC proof

To apply this input to Theorem 4.1, recall that $V(h) \ge \mathrm{Var}\,[h]$ is finite and $\mathcal{H}(r) = \{h \in \mathcal{H} : V(h) \le r\} \cup \{0\}$. Condition (4.1) bounds both signs of TC of $\mathcal{H}(r)$ and the positive TC of $\mathcal{H}(r)^{[2]}$ by the sub-root function ψ. For $r > 0$ and $h \in \mathcal{H}$, let $w_r(h)$ be the smallest number of the form $4^k r$, where $k \in \mathbb{N} \cup \{0\}$, that is at least $V(h)$. Define $\mathcal{A}_r := \{0\} \cup \{rh/w_r(h) : h \in \mathcal{H}\}$. As a result, $\mathcal{A}_r$ aggregates all variance shells at the single scale r. The following deterministic lemma collects the properties of this class used in the proof.

Lemma B.2 (Finite-Envelope Rescaling Of An Unbounded Class)**.** Let $0 \in \mathcal{H} \subset \mathbb{R}^{[n]}$, let $\mathrm{Var}\,[h] \le V(h) < \infty$ for every $h \in \mathcal{H}$, and let ψ be a sub-root function satisfying (4.1). Then, for every $r > 0$,

$$\sup_{a\in\mathcal{A}_r} \mathrm{Var}\,[a] \le r, \qquad \max_{\eta\in\{+,-\}} \mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{A}_r) \le 2\psi(r), \qquad \mathfrak{R}^{+}_{N_{u,m}}(\mathcal{A}_r^{[2]}) \le 2\psi(r),$$
$$B_\infty(\mathcal{A}_r) \le \min\left\{B_\infty(\mathcal{H}), \sqrt{(n-1)r}\right\}. \tag{B.2}$$

In particular, $B_\infty(\mathcal{A}_r) < \infty$ for every $r > 0$, including when $B_\infty(\mathcal{H}) = \infty$.

Proof of Lemma B.2. For the fixed radius r, $w_r(h)$ is the first point of the grid $\{r, 4r, 16r, \ldots\}$ at least $V(h)$, and $\mathcal{A}_r = \{0\} \cup \{rh/w_r(h) : h \in \mathcal{H}\}$. Its square class is $\mathcal{A}_r^{[2]} = \{(a - P_n a)^2 : a \in \mathcal{A}_r\}$. We recall $R_p^{\pm} h = \pm(P_{A_p} - P_n)h$ and $\mathfrak{R}_p^{\eta}(\mathcal{H}) = \mathbb{E}_{A_p}\left[\sup_{h\in\mathcal{H}} R_p^{\eta} h\right]$. Concentration will be applied to the entire class $\mathcal{A}_r$, so the complexity bound must include all its variance shells.

Proof outline.

(1) Variance and envelope. We control every rescaled shell by its variance radius and the finite-population variance inequality.
(2) First- and second-order TC. We sum the shellwise first-order TCs and centered-square complexities using the sub-root scaling rule.

Step 1: Variance and envelope. The rescaling must make every centered coordinate finite while retaining control by the original variance surrogate. Since $V(h) < \infty$, the grid point $w_r(h)$ exists, and its multiplier $r/w_r(h)$ lies in $(0, 1]$. The zero member satisfies every bound. For a represented member with $w_r(h) = 4^k r$, we compute variance directly, without assuming that $\mathcal{H}$ is closed under rescaling or that V is homogeneous:

$$\mathrm{Var}\left[\frac{rh}{w_r(h)}\right] \overset{①}{=} 4^{-2k}\mathrm{Var}\,[h] \overset{②}{\le} 4^{-2k}V(h) \overset{③}{\le} 4^{-k}r \overset{④}{\le} r.$$

Step ① uses quadratic homogeneity of variance. Steps ②–④ use $\mathrm{Var}\,[h] \le V(h) \le 4^k r$ and $k \ge 0$.

Since $r/w_r(h) = 4^{-k} \le 1$, centering commutes with this rescaling and gives $B_\infty(\mathcal{A}_r) \le B_\infty(\mathcal{H})$. For any $a \in \mathcal{A}_r$, put $z = a - P_n a$ and $z_i = z(i)$. Since $\sum_i z_i = 0$, Cauchy–Schwarz gives

$$z_i^2 \overset{⑤}{=} \left(\sum_{j\ne i} z_j\right)^2 \overset{⑥}{\le} (n-1)\sum_{j\ne i} z_j^2, \qquad n z_i^2 \overset{⑦}{\le} (n-1)\sum_{j=1}^{n} z_j^2.$$

Steps ⑤–⑦ use centering, Cauchy–Schwarz, and addition of $(n-1)z_i^2$, respectively.

Therefore $z_i^2 \le (n-1)P_n z^2 \le (n-1)r$. Taking first the maximum over i and then the supremum over $a \in \mathcal{A}_r$ proves $B_\infty(\mathcal{A}_r) \le \sqrt{(n-1)r}$. The factor $n-1$ is attained by $z(1) = n-1$ and $z(i) = -1$ for $i > 1$, for which $P_n z^2 = n-1$ and $\max_i z(i)^2 = (n-1)^2$. Constant functions, including those of arbitrarily large magnitude, have zero centered coordinates and split deviation.

Step 2: First- and second-order TC. We control the union by summing its nonnegative shell suprema. For $k \ge 0$, put $S_k = \{h \in \mathcal{H} : w_r(h) = 4^k r\}$. Since $V(h) \le 4^k r$ on S_k, $\mathcal{A}_r \subseteq \bigcup_{k\ge 0}\{4^{-k}h : h \in \mathcal{H}(4^k r)\}$. Each class in this union contains zero, making its one-sided supremum nonnegative. For every sampled subset, the supremum over this union is at most the sum of those suprema. As a result, for each $\eta \in \{+, -\}$,

$$\mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{A}_r) \overset{⑧}{\le} \sum_{k\ge 0} 4^{-k}\mathfrak{R}^{\eta}_{N_{u,m}}\left(\mathcal{H}(4^k r)\right) \overset{⑨}{\le} \sum_{k\ge 0} 4^{-k}\psi(4^k r).$$

Step ⑧ sums the nonnegative shell suprema, and ⑨ applies (4.1) at $4^k r$.

For $h \in S_k$, centering and squaring the rescaled function gives $\left(4^{-k}h - P_n(4^{-k}h)\right)^2 = 16^{-k}(h - P_n h)^2$. Repeating the shell-union bound with this degree-two scaling gives

$$\mathfrak{R}^{+}_{N_{u,m}}(\mathcal{A}_r^{[2]}) \overset{①}{\le} \sum_{k\ge 0} 16^{-k}\mathfrak{R}^{+}_{N_{u,m}}\left(\mathcal{H}(4^k r)^{[2]}\right) \overset{②}{\le} \sum_{k\ge 0} 16^{-k}\psi(4^k r) \overset{③}{\le} \psi(r)\sum_{k\ge 0} 8^{-k}.$$

Steps ①–③ use the shell bound, the centered-square TC hypothesis, and $\psi(4^k r) \le 2^k \psi(r)$, respectively.

The first-order and square-class series are therefore bounded by $\psi(r)\sum_{k\ge 0} 2^{-k} = 2\psi(r)$ and $\psi(r)\sum_{k\ge 0} 8^{-k} = (8/7)\psi(r)$, respectively. Together with Step 1, this proves (B.2). □

Proof of Theorem 4.1. We recall that $A \ge 1$ is the coefficient of the variance penalty $V(h)/A$, whereas $x \ge 1$ specifies the one-sided failure-probability bound $\exp(-x)$ over the random split. The

split gap is $g(h) = (P_{\mathcal{U}} - P_{\mathcal{M}})h$, while $V(h) \geq \mathrm{Var}[h]$ is a property of the fixed function. We put $t := x/N_{u,m}$ and first bound $g(h)$. Step 4 treats $-g(h)$ by reflecting the rescaled class. The two TC quantities in (4.1) are controlled by ψ, whose fixed point r^* is finite by Lemma A.5. In the concentration input, $v^2(\mathcal{A}_r) = \sup_{a\in\mathcal{A}_r} \mathrm{Var}[a]$ and $B_\infty(\mathcal{A}_r) = \max_i \sup_{a\in\mathcal{A}_r} |a(i) - P_n a|$. We use the rescaling $\mathcal{A}_r = \{0\} \cup \{rh/w_r(h) : h \in \mathcal{H}\}$ from Lemma B.2, and the numerical constant $c_1 = 64(4 + c_0)^2$ from the theorem.

Proof outline.

(1) Geometric rescaling. Insert Lemma B.2 into the finite-envelope concentration inequality.
(2) Explicit fixed-point radius. Use $\psi(r) \leq \sqrt{rr^*}$ for $r \geq r^*$ and the two envelope bounds to choose one deterministic radius.
(3) Undo the rescaling. Bound the entire rescaled class on one concentration event, then recover a bound for each original function.
(4) Two orientations. Reflect the rescaled class to bound the reverse gap and intersect the two concentration events.

Step 1: Geometric rescaling. We use both localized TC bounds in (4.1). Lemma A.4 shows why the centered-square bound does not follow from the first-order bound under quadratic variance localization. By Lemma B.2, for every deterministic $r > 0$, $v^2(\mathcal{A}_r) \leq r$, both first-order TCs of $\mathcal{A}_r$ are at most $2\psi(r)$, and $\mathfrak{R}^+_{N_{u,m}}(\mathcal{A}_r^{[2]}) \leq 2\psi(r)$. Consequently, the complete right-hand side supplied by Theorem B.1 is at most

$$c_{\mathrm{split}} \max_{\eta\in\{+,-\}} \mathfrak{R}^{\eta}_{N_{u,m}}(\mathcal{A}_r) + c_0 v(\mathcal{A}_r)\sqrt{t} + \frac{c_0}{2}\mathfrak{R}^+_{N_{u,m}}(\mathcal{A}_r^{[2]}) + (16B_\infty(\mathcal{A}_r) + c_0/2)\, t$$
$$\overset{①}{\leq} (4 + c_0)\psi(r) + c_0\sqrt{rt} + 16B_\infty(\mathcal{A}_r)t + \frac{c_0}{2}t \overset{②}{\leq} (4 + c_0)\left[\psi(r) + \sqrt{rt} + B_\infty(\mathcal{A}_r)t + t\right].$$

Here ① substitutes the variance and TC bounds for $\mathcal{A}_r$. ② uses $4 + c_0 \geq 16$ to bound all four coefficients.

Lemma B.2 also gives $B_\infty(\mathcal{A}_r) \leq \min\left\{B_\infty(\mathcal{H}), \sqrt{(n-1)r}\right\}$.

Step 2: Explicit fixed-point radius. We choose a deterministic radius r for which the concentration bound from Step 1 is at most $r/(4A)$. The factor four matches the grid spacing in $w_r(h)$, allowing Step 3 to recover the penalty $V(h)/A$ when the rescaling is undone. Set $C^\circ := 256(4 + c_0)^2 = 4c_1$ and define the deterministic radius

$$r_0 := C^\circ\left[A^2(r^* + t) + \min\left\{A^2(n-1)t^2, AB_\infty(\mathcal{H})t\right\}\right].$$

The first entry $A^2(n-1)t^2$ is finite, so the minimum and r_0 are finite even when $B_\infty(\mathcal{H}) = \infty$. Moreover, $r_0 \geq C^\circ A^2(r^* + t) > 0$. Since $A \geq 1$ and $C^\circ \geq 1$, we have $r_0 \geq r^*$, and the sub-root property gives $\psi(r_0) \leq \sqrt{r_0 r^*}$. For $r^* > 0$, this follows by comparing $\psi(r_0)/\sqrt{r_0}$ with $\psi(r^*)/\sqrt{r^*} = \sqrt{r^*}$. If $r^* = 0$, the designated majorant is identically zero and this inequality still holds.

First suppose $A^2(n-1)t^2 \leq AB_\infty(\mathcal{H})t$. This case includes equality and $B_\infty(\mathcal{H}) = \infty$, and the minimum equals $A^2(n-1)t^2$. Using $B_\infty(\mathcal{A}_{r_0}) \leq \sqrt{(n-1)r_0}$, Step 1 gives the upper bound

$$(4 + c_0)\left[\sqrt{r_0}\left(\sqrt{r^*} + \sqrt{t} + \sqrt{n-1}\, t\right) + t\right].$$

The definition of r_0 gives $r^* \leq r_0/(C^\circ A^2)$, $t \leq r_0/(C^\circ A^2)$, and $(n-1)t^2 \leq r_0/(C^\circ A^2)$. Each of the three square-root products in the bound is therefore at most $r_0/(\sqrt{C^\circ}A)$. Also $t \leq r_0/(C^\circ A^2) \leq r_0/(C^\circ A)$ because $A \geq 1$. Adding these four contributions gives the upper bound

$$(4 + c_0)\left(\frac{3}{\sqrt{C^\circ}} + \frac{1}{C^\circ}\right)\frac{r_0}{A} \overset{③}{=} \left(\frac{3}{16} + \frac{1}{256(4 + c_0)}\right)\frac{r_0}{A} \overset{④}{<} \frac{r_0}{4A}.$$

Here ③ substitutes $C^\circ = 256(4 + c_0)^2$. ④ uses $1/[256(4 + c_0)] < 1/16$.

Now suppose that $AB_\infty(\mathcal{H})t < A^2(n-1)t^2$. The minimum then equals $AB_\infty(\mathcal{H})t$, which is necessarily finite. We use $B_\infty(\mathcal{A}_{r_0}) \leq B_\infty(\mathcal{H})$. The definition of r_0 yields $B_\infty(\mathcal{H})t \leq r_0/(C^\circ A)$.

Independently of the case, the common summand $C^\circ A^2(r^* + t)$ in r_0 gives $r^*, t \le r_0/(C^\circ A^2)$. The terms $\sqrt{r_0 r^*}$ and $\sqrt{r_0 t}$ each contribute at most $r_0/(\sqrt{C^\circ}A)$. The terms $B_\infty(\mathcal{H})t$ and t each contribute at most $r_0/(C^\circ A)$. As a result, the Step 1 bound is at most

$$(4 + c_0)\left(\frac{2}{\sqrt{C^\circ}} + \frac{2}{C^\circ}\right)\frac{r_0}{A} \overset{⑤}{=} \left(\frac{1}{8} + \frac{1}{128(4 + c_0)}\right)\frac{r_0}{A} \overset{⑥}{<} \frac{r_0}{4A}.$$

Here ⑤ substitutes the same value of C°. ⑥ uses $1/[128(4 + c_0)] < 1/8$.

As a result, the complete concentration bound is at most $r_0/(4A)$ in either case.

Step 3: Undo the rescaling. The radius r_0 is determined by the full sample, A, and x before the split. Applying Theorem B.1 to $\mathcal{A}_{r_0}$ and using Steps 1–2 gives $g(\mathcal{A}_{r_0}) \le r_0/(4A)$ with probability at least $1 - \exp(-x)$ over the random split. The bounds at other radii were deterministic verifications, so this application requires no union bound over radii.

On this event, every $h \in \mathcal{H}$ has $r_0 h/w_{r_0}(h) \in \mathcal{A}_{r_0}$. Linearity and the positive multiplier give $g(h) = \frac{w_{r_0}(h)}{r_0} g\left(\frac{r_0 h}{w_{r_0}(h)}\right) \overset{⑦}{\le} w_{r_0}(h)/(4A)$. Here ⑦ uses membership of the rescaled function in the class controlled by the one concentration event. If $V(h) \le r_0$, then $w_{r_0}(h) = r_0$ and $g(h) \le r_0/(4A)$. Otherwise write $w_{r_0}(h) = 4^k r_0$ with $k \ge 1$. The preceding grid point cannot be at least $V(h)$, by the minimality in the definition of w_{r_0}. Hence

$$4^{k-1} r_0 \overset{⑧}{<} V(h) \le 4^k r_0 = w_{r_0}(h), \qquad g(h) \overset{⑨}{\le} \frac{4^k r_0}{4A} \overset{①}{<} \frac{V(h)}{A}.$$

Step ⑧ uses minimality of $w_{r_0}(h)$. Step ⑨ undoes the rescaling, and ① uses $4^{k-1} r_0 < V(h)$.

In either case, adding the other nonnegative term gives $g(h) \le V(h)/A + r_0/(4A)$ simultaneously for all $h \in \mathcal{H}$, because the event controls the entire rescaled class. Dividing the definition of r_0 by $4A$ gives $(C^\circ/4)[A(r^* + t) + \min\{A(n-1)t^2, B_\infty(\mathcal{H})t\}]$. Since $C^\circ/4 = c_1$, this proves the one-sided assertion.

Step 4: Two orientations. To obtain an absolute deviation bound, we repeat the same construction after reflection without requiring symmetry of the original class. For the reverse gap, apply Theorem B.1 to $-\mathcal{A}_{r_0} = \{-a : a \in \mathcal{A}_{r_0}\}$. The class process $g(-\mathcal{A}_{r_0})$ is the reverse supremum. For each $a \in \mathcal{A}_{r_0}$, $\mathrm{Var}[-a] = \mathrm{Var}[a]$ and $R^{\pm}_{N_{u,m}}(-a) = R^{\mp}_{N_{u,m}} a$. Also, $B_\infty(-\mathcal{A}_{r_0}) = B_\infty(\mathcal{A}_{r_0})$ and $(-\mathcal{A}_{r_0})^{[2]} = \mathcal{A}^{[2]}_{r_0}$, because centering commutes with reflection and squaring removes its sign. Reflection therefore preserves every input bound in Steps 1–2, which already use the maximum of the two first-order TC orientations. The reflected class contains $-r_0 h/w_{r_0}(h)$ for every $h \in \mathcal{H}$. The identity $-g(h) = \frac{w_{r_0}(h)}{r_0} g\left(-\frac{r_0 h}{w_{r_0}(h)}\right)$ and Step 3 give the same upper bound for $-g(h)$. This argument requires neither $-h \in \mathcal{H}$ nor an extension of V. Both orientations use the same x and deterministic radius r_0, and each event has failure probability at most $\exp(-x)$ over the random split. On their intersection the common bound controls $\max\{g(h), -g(h)\} = |g(h)|$ for every $h \in \mathcal{H}$. The union bound proves (4.2) with failure probability at most $2\exp(-x)$ over the random split, without independence or a confidence shift. □

B.3 Bounded-loss specialization

The next lemma verifies the bounded comparison in Section 4.1. The localized hypothesis of Yang (2026, Theorems 4.2 and 4.4) concerns ordinary squares and radii above the corresponding fixed point. The following lemma converts that hypothesis to the centered-square condition (4.1) at every radius. This is only the bounded-class special case. Theorem 4.1 allows unbounded classes when $\mathrm{Var}[h] \le V(h) < \infty$ and (4.1) holds, with the envelope alternatives in (4.2).

Lemma B.3 (Bounded STLC Specialization)**.** Let $\mathcal{H}, V, \mathcal{H}(r)$ be as in Theorem 4.1, and suppose $|h(i)| \le H_0$ for every $h \in \mathcal{H}$ and $i \in [n]$, where $H_0 > 0$. Put $p = N_{u,m}$. Let ψ_b be a finite-valued sub-root function with positive fixed point r_b, and suppose, for every $r \ge r_b$,

$$\max\left\{\max_{\eta \in \{+,-\}} \mathfrak{R}^\eta_p(\mathcal{H}(r)), \mathfrak{R}^+_p\left(\{h^2 : h \in \mathcal{H}(r)\}\right)\right\} \le \psi_b(r). \tag{B.3}$$

Here h^2 is the ordinary pointwise square, without centering. Then $\psi(r) = (1 + 4H_0)\psi_b(r \vee r_b)$ is a finite-valued sub-root majorant satisfying (4.1) for all $r \geq 0$. Its positive fixed point satisfies $r_b \leq r^* \leq (1 + 4H_0)^2 r_b$. Consequently, for every $A, x \geq 1$, with probability at least $1 - \exp(-x)$ over the random split,

$$g(h) \leq \frac{V(h)}{A} + c_1 A(1 + 4H_0)^2 r_b + c_1(A + 2H_0)\frac{x}{p}, \qquad \forall h \in \mathcal{H}.$$

Proof of Lemma B.3. Recall that $\mathcal{H}(r) = \{h \in \mathcal{H} : V(h) \leq r\} \cup \{0\}$ and $\mathcal{H}(r)^{[2]} = \{(h - P_n h)^2 : h \in \mathcal{H}(r)\}$. For a uniform p-subset A_p, the two TC orientations are the expectations of the suprema of $(P_{A_p} - P_n)h$ and its negative. All classes below are deterministic and bounded on the finite sample.

Proof outline.

(1) Centered-square comparison. Expand the centered square and bound its extra term by the two first-order deviations.
(2) Extend the majorant to every radius. Use $\mathcal{H}(r) \subseteq \mathcal{H}(r_b)$ for $r < r_b$, then verify the sub-root property.
(3) Fixed point and risk bounds. Compare fixed points and apply the bounded-envelope branch of UTLC and its ERM specialization.

Step 1: Centered-square comparison. Centering a square creates a linear correction. The two first-order TC orientations will control this correction. Fix $r \geq 0$ and a subset A_p. For $h \in \mathcal{H}(r)$, the bound $|P_n h| \leq P_n |h| \leq H_0$ and cancellation of constants give

$$\begin{aligned}(P_{A_p} - P_n)(h - P_n h)^2 &\overset{①}{=} (P_{A_p} - P_n)h^2 - 2(P_n h)(P_{A_p} - P_n)h \\ &\overset{②}{\leq} (P_{A_p} - P_n)h^2 + 2H_0 \left|(P_{A_p} - P_n)h\right|.\end{aligned}$$

Steps ①–② expand the centered square, cancel constants, and use $|P_n h| \leq H_0$.

Since $0 \in \mathcal{H}(r)$, both signed suprema are nonnegative, and

$$\sup_{h \in \mathcal{H}(r)} \left|(P_{A_p} - P_n)h\right| \overset{③}{\leq} \sup_{h \in \mathcal{H}(r)} (P_{A_p} - P_n)h + \sup_{h \in \mathcal{H}(r)} (P_n - P_{A_p})h.$$

Here ③ bounds the maximum of the two nonnegative signed suprema by their sum.

Taking suprema in the first display and then averaging over A_p yields

$$\mathfrak{R}_p^+(\mathcal{H}(r)^{[2]}) \overset{④}{\leq} \mathfrak{R}_p^+\left(\{h^2 : h \in \mathcal{H}(r)\}\right) + 2H_0\left[\mathfrak{R}_p^+(\mathcal{H}(r)) + \mathfrak{R}_p^-(\mathcal{H}(r))\right].$$

Here ④ takes suprema in the preceding pointwise comparison and averages over A_p.

For $r \geq r_b$, (B.3) bounds both levels by $(1 + 4H_0)\psi_b(r)$.

Step 2: Extend the majorant to every radius. Put $c = 1 + 4H_0$. For $r < r_b$, nesting bounds both levels by their values at r_b, hence by $c\psi_b(r_b) = cr_b$. Therefore $\psi(r) = c\psi_b(r \vee r_b)$ satisfies (4.1) at every radius. It is finite and nondecreasing. Its ratio $\psi(r)/\sqrt{r}$ equals $cr_b/\sqrt{r}$ below r_b and $c\psi_b(r)/\sqrt{r}$ above it. Both expressions are nonincreasing and agree at r_b, proving the sub-root property.

Step 3: Fixed point and risk bounds. We compare the enlarged majorant with the identity at one explicit radius, then substitute that bound into UTLC. Because $c \geq 1$ and $\psi_b(r_b) = r_b$, the sub-root property gives

$$\psi(c^2 r_b) = c\psi_b(c^2 r_b) \overset{⑤}{\leq} c\sqrt{c^2 r_b / r_b}\,\psi_b(r_b) \overset{⑥}{=} c^2 r_b.$$

Steps ⑤–⑥ use sub-root scaling and $\psi_b(r_b) = r_b$.

Lemma A.5 therefore gives $r^* \leq c^2 r_b$. For $0 < r < r_b$, the completed majorant satisfies $\psi(r) = cr_b > r$, so no such r is a fixed point. Hence $r^* \geq r_b$. Also $B_\infty(\mathcal{H}) \leq 2H_0$. Substitution into the one-sided assertion of Theorem 4.1 proves the displayed bound. A surrogate dominating $P_n h^2$, as in the cited bounded STLC theorem, also dominates Var $[h]$, so it satisfies the UTLC surrogate requirement.

For the bounded ERM comparison, apply this construction to $\mathcal{H} = \mathcal{G}_\star$ with $V_\star(g_f) = BP_n g_f$ under Assumption 1 with $\theta = 1$. Suppose ψ_b, r_b satisfy (B.3) for this oracle class and its localization. The hypothesis of Yang (2026, Theorem 4.4) bounds both TC orientations of the ordinary-square oracle class at both subset sizes, so it implies this condition at $p = N_{u,m}$. If $|g_f(i)| \le H_0$, then $r_{\mathrm{exc}} \le c^2 r_b$ and $B_\infty(\mathcal{G}_\star) \le 2H_0$. Theorem 4.2, with $s = x + \log 2$, consequently gives

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) \overset{⑦}{\le} C_B \left[c^2 r_b + (1 + 2H_0)\frac{s}{p} \right]$$

Here ⑦ substitutes $r_{\mathrm{exc}} \le c^2 r_b$ and $B_\infty(\mathcal{G}_\star) \le 2H_0$ into Theorem 4.2.

This holds with probability at least $1 - \exp(-x)$ over the random split, under the attainment assumptions of Theorem 4.2, and proves the claimed bounded-loss recovery with constants depending on B, H_0. □

C Proof of the Generic ERM Theorem, Theorem 4.2

C.1 Preparatory lemmas

We use $g_f = \ell_f - \ell_{f_n^\star}$ for the oracle excess function and $g(g_f) = (P_{\mathcal{U}} - P_{\mathcal{M}})g_f$ for its random scalar split gap. The full-sample mean $P_n g_f$ is nonnegative. The oracle class is $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}\}$. For $r \ge 0$, its localization is $\mathcal{G}_\star(r) = \{g_f \in \mathcal{G}_\star : B(P_n g_f)^\theta \le r\}$, with surrogate localization functional $V_\star(g_f) = B(P_n g_f)^\theta$. The function $\psi_\star$ satisfies (4.1) for $\mathcal{G}_\star$ and $V_\star$, and r_{exc} denotes its designated fixed point.

Lemma C.1 (Finiteness Of The ERM Fixed Point)**.** Suppose Assumption 1 holds and $\psi_\star$ satisfies the localized TC condition (4.1) for $\mathcal{G}_\star$ and $V_\star$. The majorant $\psi_\star$ has a finite designated fixed point r_{exc}. For every $r_0 > 0$,

$$r_{\mathrm{exc}} \le \max\left\{ r_0, \frac{\psi_\star(r_0)^2}{r_0} \right\} < \infty.$$

The zero function has fixed point zero. Every nonzero majorant has a unique positive fixed point.

Proof of Lemma C.1. Proof outline.

(1) Apply the sub-root lemma. Verify its hypotheses for $\psi_\star$ and identify its fixed point with r_{exc}.

Step 1: Apply the sub-root lemma. Condition (4.1) assumes that $\psi_\star$ is a finite-valued sub-root function. Applying Lemma A.5 to its fixed point r_{exc} gives, for every $r_0 > 0$,

$$r_{\mathrm{exc}} \overset{①}{\le} \max\left\{ r_0, \psi_\star(r_0)^2 / r_0 \right\} < \infty.$$

Here ① uses the explicit reference-radius bound for a finite-valued sub-root function. The same lemma gives the zero case and uniqueness of the positive fixed point in the nonzero case. □

Lemma C.2 (Oracle Decomposition And The Limit Of Square Comparison)**.** Under Assumption 1, $V_\star$ is a finite surrogate localization functional on $\mathcal{G}_\star$, with $V_\star(0) = 0$. For any two subclasses $\mathcal{A}, \mathcal{B} \subseteq \mathcal{G}_\star$ containing zero, write $\mathcal{A} - \mathcal{B} = \{a - b : a \in \mathcal{A}, b \in \mathcal{B}\}$. For $1 \le p < n$ and $\eta \in \{+, -\}$, with $-\eta$ denoting the opposite sign,

$$\mathfrak{R}_p^\eta(\mathcal{A} - \mathcal{B}) = \mathfrak{R}_p^\eta(\mathcal{A}) + \mathfrak{R}_p^{-\eta}(\mathcal{B}).$$

In particular this equality applies to $\mathcal{A} = \mathcal{B} = \mathcal{G}_\star(r)$. There is no analogous universal multiplicative bound for the positive TC of centered squares by the two individual centered-square TCs.

Proof of Lemma C.2. Recall $g_f = \ell_f - \ell_{f_n^\star}$, $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}\}$, and $V_\star(g_f) = B(P_n g_f)^\theta$. The notation R_p^η denotes the signed subset deviation, with $R_p^+ h = (P_{A_p} - P_n)h$ and $R_p^- h = -R_p^+ h$. The class $\mathcal{H}^{[2]}$ consists of the functions $(h - P_n h)^2$ for $h \in \mathcal{H}$. All classes are fixed before the uniform subset A_p is drawn.

Proof outline.

(1) Variance and zero. Verify the surrogate directly from the Bernstein hypothesis and oracle optimality.
(2) Linear comparison. Separate the two suprema for each subset and then take expectation.
(3) Square obstruction. Construct two centered functions with constant squares and a difference with nonconstant square.

Step 1: Variance and zero. The surrogate must depend only on the excess function and must identify correctly what happens at zero radius. Oracle optimality gives $P_n g_f \geq 0$. Equal excess functions on $[n]$ have equal means, so $V_\star$ is well defined. Finite sample losses and Assumption 1(2) give $\mathrm{Var}\,[g_f] \overset{①}{\leq} B(P_n g_f)^\theta = V_\star(g_f) < \infty$. Here ① applies the assumed Bernstein inequality. The oracle represents zero and gives $V_\star(0) = 0$. Conversely, if $P_n g_f = 0$, then $P_n g_f^2 = \mathrm{Var}\,[g_f] \leq B(P_n g_f)^\theta = 0$. Nonnegativity of the squared coordinates implies $g_f = 0$ on $[n]$. Since $B, \theta > 0$, this proves $\mathcal{G}_\star(0) = \{0\}$ and shows that all full-sample minimizers have the same sample losses, even if their predictions are not unique.

Step 2: Linear comparison. Linear deviations can be separated before taking expectations, which is the comparison available without a new square-class assumption. For each fixed subset, linearity gives $R_p^\eta(a-b) \overset{②}{=} R_p^\eta a + R_p^{-\eta} b$. Here ② uses linearity of the signed subset functional. Independent choices of a, b separate the supremum into a sum. Approximating maximizers justify equality without attainment. If one supremum is infinite, fix the other function at zero. Both suprema are nonnegative because the classes contain zero, so averaging proves the equality even in the extended-real case.

Step 3: Square obstruction. To distinguish a pointwise square bound from a centered-process bound, we exhibit constant individual squares but a nonconstant squared difference. Take $n = 4$, $p = 2$, and centered functions $z, w : [4] \to \mathbb{R}$ whose evaluation vectors are $(1, 1, -1, -1)$ and $(1, -1, 1, -1)$, respectively. Consider the three loss functions $0, 1 + z, 1 + w$, with the zero function as full-sample oracle. The two nonzero excesses have mean one and variance one, so Assumption 1 holds with $B = 1, \theta = 1$. Their centered squares are both the constant function 1. As a result, $\mathcal{G}_\star(1)^{[2]} \overset{③}{=} \{0, 1\}$ has TC zero for every subset size. Here ③ centers the two nonzero excesses to obtain z, w and uses $z^2 = w^2 = 1$. However, $z - w$ belongs to $\mathcal{G}_\star(1) - \mathcal{G}_\star(1)$ and its centered square has values $0, 4, 4, 0$ at indices $1, 2, 3, 4$, respectively, with full-sample mean two. On the subset $\{2, 3\}$ its positive deviation is two. This subset has probability $1/6$ under the uniform choice of A_p, and zero belongs to the square class, so that class has positive TC at least $2/6 = 1/3$. A finite multiple of the individual square-class TCs is zero and cannot bound it. The pointwise inequality $(z - w)^2 \leq 2z^2 + 2w^2$ does not resolve this obstruction: $P_{A_p} - P_n$ has negative coefficients outside A_p and is not order preserving. The oracle-gap decomposition in Theorem 4.2 avoids this obstruction by using linear deviations before any squaring. □

To solve the excess inequalities from empirical optimality, we use the following scalar operation. For $d, q \geq 0$ and $0 < \theta \leq 1$, with the additional restriction $q < 1$ when $\theta = 1$, we define $\mathrm{root}_{\theta,q}(d) := \max\{r \geq 0 : r \leq d + qr^\theta\}$. The next lemma proves that this maximum exists and gives the bounds used in the proof of Theorem 4.2.

Lemma C.3 (Nonlinear Absorption)**.** Let $d, q \geq 0$ and $0 < \theta \leq 1$, with the additional restriction $q < 1$ when $\theta = 1$. The maximum defining $\mathrm{root}_{\theta,q}(d)$ exists, is finite, and is nondecreasing in d. If $\theta = 1$, then $\mathrm{root}_{1,q}(d) = d/(1-q)$. If $\theta < 1$, then $\mathrm{root}_{\theta,q}(d) \leq 2d + (2q)^{1/(1-\theta)}$. Consequently, every $r \geq 0$ satisfying $r \leq d + qr^\theta$ obeys $r \leq \mathrm{root}_{\theta,q}(d)$.

Proof of Lemma C.3. Recall that $\mathrm{root}_{\theta,q}(d)$ is the largest $r \geq 0$ satisfying $r \leq d + qr^\theta$, with $d, q \geq 0$, $0 < \theta \leq 1$, and $q < 1$ if $\theta = 1$. The feasible set contains zero and is closed by continuity.

Proof outline.

(1) Sublinear case. Bound feasible values by separating $r \leq 2d$ from $r > 2d$.
(2) Linear case. Solve the defining inequality.
(3) Monotonicity and absorption. Compare feasible sets and identify their maximum.

Step 1: Sublinear case. Suppose $\theta < 1$. If a feasible r exceeds $2d$, then $r \le d + qr^\theta < r/2 + qr^\theta$, hence $r^{1-\theta} < 2q$. This case is impossible for $q = 0$ and otherwise gives $r < (2q)^{1/(1-\theta)}$. Together with the case $r \le 2d$, this proves $\mathrm{root}_{\theta,q}(d) \le 2d + (2q)^{1/(1-\theta)}$. The nonempty closed feasible set is therefore bounded and attains its maximum.

Step 2: Linear case. For $\theta = 1$ and $q < 1$, feasibility is exactly $0 \le r \le d/(1-q)$, giving $\mathrm{root}_{1,q}(d) = d/(1-q)$.

Step 3: Monotonicity and absorption. Increasing d enlarges the feasible set, so its maximum is nondecreasing. Write $R = \mathrm{root}_{\theta,q}(d)$. Maximality and continuity give $R = d + qR^\theta$, since a strict inequality would persist for some $r > R$. Concavity then yields, for $0 \le t \le 1$,

$$d + q(tR)^\theta - tR \overset{①}{\ge} (1-t)d + t(d + qR^\theta - R) = (1-t)d \ge 0.$$

Step ① applies concavity to the endpoints $0, R$. Hence the feasible set is $[0, R]$, including $R = 0$. For $d = 0$, $q > 0$, and $\theta < 1$, its maximum is the positive root $q^{1/(1-\theta)}$. The absorption assertion follows from maximality. □

C.2 Proof of the ERM bounds

We first establish a parameterized bound from which (4.3) follows. This bound is also used in the kernel application. For $a, z \ge 1$, define the single UTLC remainder

$$R_\star(a, z) \coloneqq c_1 a \left(r_{\mathrm{exc}} + \frac{z}{N_{u,m}} \right) + c_1 \min \left\{ a(n-1)\frac{z^2}{N_{u,m}^2}, B_\infty(\mathcal{G}_\star)\frac{z}{N_{u,m}} \right\}.$$

Here c_1 is the constant in Theorem 4.1. For $a, z \ge 1$ and $t \in \{m, u\}$, with the additional restriction $a > Bt/n$ when $\theta = 1$, we define $Q_t(a, z) \coloneqq \mathrm{root}_{\theta, Bt/(an)}\left((t/n)R_\star(a, z)\right)$. Lemma C.3 ensures that $Q_t(a, z)$ is finite. Under the hypotheses of Theorem 4.2, fix deterministic $a, x_m, x_u \ge 1$, with $a > B\max\{u, m\}/n$ when $\theta = 1$. Then with probability at least $1 - \exp(-x_m) - \exp(-x_u)$ over the random split,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) \le R_\star(a, x_m) + R_\star(a, x_u) + \frac{B}{a}\left[Q_u(a, x_m)^\theta + Q_m(a, x_u)^\theta \right]. \tag{C.1}$$

Proof of Theorem 4.2. We fix the full sample, $a, x_m, x_u \ge 1$, and the majorant $\psi_\star$, with the additional restriction $a > B\max\{m, u\}/n$ when $\theta = 1$. The only randomness is the complementary split $(\mathcal{M}, \mathcal{U})$. The constants $B > 0$ and $0 < \theta \le 1$ are the Bernstein parameters in Assumption 1(2). Recall $g_f = \ell_f - \ell_{f_n^\star}$, $V_\star(g_f) = B(P_n g_f)^\theta$, and $g(h) = P_{\mathcal{U}}h - P_{\mathcal{M}}h$. The constant r_{exc} is the fixed point of $\psi_\star$, and $c_1 = 64(4 + c_0)^2$ is the numerical UTLC constant. For use throughout the proof, the deterministic remainder is

$$R_\star(a, z) = c_1 a \left(r_{\mathrm{exc}} + \frac{z}{N_{u,m}} \right) + c_1 \min \left\{ a(n-1)\frac{z^2}{N_{u,m}^2}, B_\infty(\mathcal{G}_\star)\frac{z}{N_{u,m}} \right\}.$$

For $t \in \{m, u\}$ and $z \ge 1$, the quantity $Q_t(a, z) = \mathrm{root}_{\theta, Bt/(an)}((t/n)R_\star(a, z))$ is the largest $r \ge 0$ satisfying $r \le (t/n)R_\star(a, z) + (Bt/(an))r^\theta$. Lemma C.3 makes this definition well posed, because our restriction on a gives $Bt/(an) < 1$ when $\theta = 1$. The prescribed estimators minimize their respective empirical losses over the same class $\mathcal{F}$.

Proof outline.

(1) Two signed oracle events. Apply Theorem 4.1 in opposite directions with total failure probability at most $\exp(-x_m) + \exp(-x_u)$ over the random split.
(2) Full-sample absorption. Combine empirical optimality with the averaging identities and Lemma C.3 to bound the two full-sample excesses by Q_u and Q_m.
(3) Reuse of the same events. Write $\ell_{\widehat{f}_m} - \ell_{\widehat{f}_u} = g_{\widehat{f}_m} - g_{\widehat{f}_u}$. The two uniform bounds from Step 1 control the corresponding deviations in opposite directions, yielding (C.1) with the single fixed point r_{exc} and no additional failure probability.
(4) Specialization for all Bernstein exponents. Choose a deterministically, use Lemma C.3 to bound the nonlinear terms, and set $x_m = x_u = x + \log 2$ to obtain (4.3).

Step 1: Two signed oracle events. The two estimators need opposite deviations, so we construct two uniform oracle events that can later be reused. The class $\mathcal{G}_\star$ contains zero, its functions are real-valued on $[n]$, and Lemma C.2 verifies the surrogate localization functional. The localized TC condition (4.1), with $\mathcal{H} = \mathcal{G}_\star$, $V = V_\star$, and $\psi = \psi_\star$, supplies the common majorant required by Theorem 4.1. Apply its one-sided assertion with sign $+1$ at x_m, and with sign -1 at x_u, using coefficient a in both. Their intersection, denoted by E, has probability at least $1 - \exp(-x_m) - \exp(-x_u)$ over the random split. On E, for every $f \in \mathcal{F}$,

$$g(g_f) \overset{①}{\le} \frac{B}{a}(P_n g_f)^\theta + R_\star(a, x_m), \qquad -g(g_f) \overset{②}{\le} \frac{B}{a}(P_n g_f)^\theta + R_\star(a, x_u). \tag{C.2}$$

Steps ①–② are the two one-sided UTLC bounds. Their intersection uses the union bound, without an independence assumption.

Step 2: Full-sample absorption. Empirical optimality converts each split deviation into an inequality for its full-sample excess, which is the quantity entering the variance surrogate. We work on E. Put $r_m = P_n g_{\widehat{f}_m}$ and $r_u = P_n g_{\widehat{f}_u}$. These are random full-sample excesses, whereas r_{exc} is deterministic. Both are nonnegative because $f_n^\star$ minimizes the full-sample risk. For every function $h : [n] \to \mathbb{R}$, the complementary averages give

$$P_n h \overset{③}{=} \frac{m}{n} P_{\mathcal{M}} h + \frac{u}{n} P_{\mathcal{U}} h \overset{④}{=} P_{\mathcal{M}} h + \frac{u}{n} g(h) \overset{⑤}{=} P_{\mathcal{U}} h - \frac{m}{n} g(h).$$

Steps ③–⑤ partition the full-sample average and substitute $g = P_{\mathcal{U}} - P_{\mathcal{M}}$.

The oracle is a feasible competitor in each empirical problem. Consequently $P_{\mathcal{M}} g_{\widehat{f}_m} \le 0$ and $P_{\mathcal{U}} g_{\widehat{f}_u} \le 0$. Combining these optimality inequalities with the averaging identities and then (C.2) gives

$$\begin{aligned} r_m &\overset{⑥}{\le} \frac{u}{n} g(g_{\widehat{f}_m}) \overset{⑦}{\le} \frac{u}{n}\left[\frac{B}{a} r_m^\theta + R_\star(a, x_m)\right], \\ r_u &\overset{⑧}{\le} -\frac{m}{n} g(g_{\widehat{f}_u}) \overset{⑨}{\le} \frac{m}{n}\left[\frac{B}{a} r_u^\theta + R_\star(a, x_u)\right]. \end{aligned}$$

Steps ⑥ and ⑧ use training and test optimality, respectively. Steps ⑦ and ⑨ use the corresponding uniform inequalities in (C.2).

When $\theta < 1$, Lemma C.3 applies for every finite nonnegative coefficient. When $\theta = 1$, the hypothesis on a gives $Bu/(an) < 1$ and $Bm/(an) < 1$, so both denominators are positive. In either case that lemma gives

$$r_m \overset{①}{\le} Q_u(a, x_m), \qquad r_u \overset{②}{\le} Q_m(a, x_u). \tag{C.3}$$

Steps ①–② apply Lemma C.3 to r_m and r_u, respectively.

Step 3: Reuse of the same events. The desired test excess is bounded by a difference of the two oracle gaps. This lets us use the same events without controlling a squared pairwise class. Set $h = \ell_{\widehat{f}_m} - \ell_{\widehat{f}_u}$. The test oracle is feasible in the training problem, so $P_{\mathcal{M}} h \le 0$. Training optimality and linearity therefore imply

$$0 \overset{③}{\le} \mathcal{E}_{\mathcal{U}}(\widehat{f}_m) = P_{\mathcal{U}} h \overset{④}{\le} g(h) \overset{⑤}{=} g(g_{\widehat{f}_m}) - g(g_{\widehat{f}_u}). \tag{C.4}$$

Steps ③–⑤ use test optimality, training optimality, and cancellation of the common oracle loss, respectively.

Combining (C.4) with the positive and negative bounds in (C.2) gives, still on E,

$$\begin{aligned} \mathcal{E}_{\mathcal{U}}(\widehat{f}_m) &\overset{⑥}{\le} R_\star(a, x_m) + R_\star(a, x_u) + \frac{B}{a}(r_m^\theta + r_u^\theta) \\ &\overset{⑦}{\le} R_\star(a, x_m) + R_\star(a, x_u) + \frac{B}{a}\left[Q_u(a, x_m)^\theta + Q_m(a, x_u)^\theta\right]. \end{aligned}$$

Step ⑥ adds the two oracle bounds on E. Step ⑦ uses (C.3) and monotonicity of $t \mapsto t^\theta$, proving (C.1).

Step 4: Specialization for all Bernstein exponents. Fix $0 < \theta \le 1$ and $x \ge 1$. Recall the deterministic quantities $s = x + \log 2$, $\delta_s = r_{\mathrm{exc}} + s/N_{u,m}$, $\tau_s = (n-1)s^2/N_{u,m}^2$, and $\Phi_\theta(v) = v \vee v^{1/(2-\theta)}$ from the statement. The choice of a balances the concentration remainder, whose first term grows linearly in a, with the Bernstein contribution multiplied by B/a. Define

$$W = \min\left\{\Phi_\theta(\delta_s + \tau_s), \Phi_\theta(\delta_s) + B_\infty(\mathcal{G}_\star)\frac{s}{N_{u,m}}\right\},$$
$$a = A_0 \max\left\{1, W^{\theta-1}\right\}, \qquad D = c_1 A_0 W, \qquad A_0 = 2(1 \vee B).$$

The theorem's constant is $C_B = 3c_1A_0$, so $3D = C_B W$. The first candidate in W is finite, and both candidates are at least $\Phi_\theta(\delta_s) > 0$. As a result, $0 < W < \infty$ even when $B_\infty(\mathcal{G}_\star) = \infty$, and all powers and divisions in these definitions are well defined. The coefficient a is finite, deterministic, and at least $A_0 \ge 2$. For every $0 < \theta \le 1$, it satisfies $a \ge A_0 \ge 2B$, so $Bt/(an) \le t/(2n) < 1$ for $t \in \{m, u\}$. As a result, both absorption roots are well defined, including at $\theta = 1$, when $a = A_0$. We use Steps 1–3 with this single choice of a and with $x_m = x_u = s$. We do not assert a simultaneous event for all a. We verify that the concentration remainder is at most the candidate risk scale D and that the normalized Bernstein coefficient $(B/a)D^{\theta-1}$ is at most $1/2$. These bounds reduce both scalar absorption inequalities to the same normalized inequality.

First we verify $R_\star(a, s) \le D$. For every $v > 0$, $v \max\left\{1, \Phi_\theta(v)^{\theta-1}\right\} = \Phi_\theta(v)$. Indeed, for $v \ge 1$ both sides equal v. For $0 < v \le 1$, the left-hand side equals $v^{1+(\theta-1)/(2-\theta)} = v^{1/(2-\theta)}$. If the first candidate defining W is no larger than the second, including equality, then $W = \Phi_\theta(\delta_s + \tau_s)$ and $R_\star(a, s) \overset{⑧}{\le} c_1 a(\delta_s + \tau_s) = c_1 A_0 W = D$. Here ⑧ chooses the variance-envelope branch of the remainder, after which the power identity gives equality with D. Otherwise the second candidate is strictly smaller and hence finite. In this case $W \ge \Phi_\theta(\delta_s)$ and $\theta - 1 \le 0$ give

$$a\delta_s \overset{⑨}{\le} A_0\delta_s \max\left\{1, \Phi_\theta(\delta_s)^{\theta-1}\right\} \overset{①}{=} A_0\Phi_\theta(\delta_s),$$
$$R_\star(a, s) \overset{②}{\le} c_1\left[A_0\Phi_\theta(\delta_s) + B_\infty(\mathcal{G}_\star)\frac{s}{N_{u,m}}\right] \overset{③}{\le} c_1 A_0 W = D.$$

Step ⑨ uses $W \ge \Phi_\theta(\delta_s)$ and $\theta - 1 \le 0$, and ① uses the preceding power identity. Steps ②–③ select the envelope branch and use $A_0 \ge 1$.

Next put $\kappa = (B/a)D^{\theta-1}$. Since $c_1A_0 \ge 1$ and $\theta - 1 \le 0$,

$$0 \le \kappa \overset{④}{\le} \frac{B}{A_0}\frac{W^{\theta-1}}{\max\{1, W^{\theta-1}\}} \overset{⑤}{\le} \frac{1}{2}.$$

Here ④ uses $(c_1A_0)^{\theta-1} \le 1$. ⑤ uses $B/A_0 \le 1/2$ and bounds the displayed ratio by one.

For $t \in \{m, u\}$, write $\alpha = t/n \in (0, 1)$ and $y = Q_t(a, s)/D$. The maximum defining $Q_t(a, s)$ is feasible by Lemma C.3. Division of its defining inequality by D, followed by $R_\star(a, s) \le D$, gives $y \le \alpha(1 + \kappa y^\theta)$. We claim that $y \le 2\alpha$. If $y > 2\alpha$, then $y > 0$ and

$$1 \overset{⑥}{\le} \frac{\alpha}{y} + \alpha\kappa y^{\theta-1} \overset{⑦}{<} \frac{1}{2} + \frac{\kappa}{2}(2\alpha)^\theta \overset{⑧}{\le} \frac{1}{2} + \kappa \overset{⑨}{\le} 1.$$

Steps ⑥–⑨ use division by $y > 0$, $y > 2\alpha$ with $\theta - 1 \le 0$, $(2\alpha)^\theta \le 2$, and $\kappa \le 1/2$, respectively. The strict inequality is a contradiction.

We have proved $Q_t(a, s) \le \frac{2t}{n}D, \qquad t \in \{m, u\}$. Since $u + m = n$, concavity gives $\frac{1}{2}[(u/n)^\theta + (m/n)^\theta] \overset{①}{\le} ((u + m)/(2n))^\theta = 2^{-\theta}$, also when $\theta = 1$. Here ① applies concavity of the power function to u/n and m/n. Consequently,

$$Q_u(a, s)^\theta + Q_m(a, s)^\theta \overset{②}{\le} 2^\theta D^\theta\left[(u/n)^\theta + (m/n)^\theta\right] \overset{③}{\le} 2D^\theta.$$

Here ② raises the two absorption bounds to the power θ. ③ uses concavity and $u + m = n$.

Therefore (C.1) is at most $2D + \frac{2B}{a}D^\theta = 2D(1 + \kappa) \overset{④}{\le} 3D = C_B W$. Here ④ uses $\kappa \le 1/2$ to complete the absorption. This proves (4.3), with a constant independent of θ. At $\theta = 1$, Φ_1 is the

identity and $a = A_0$, so the minimum simplifies to the stated linear-Bernstein expression without changing C_B. For later kernel applications, we also record the slightly more precise linear estimate. Whenever $\theta = 1$, $a \geq 2B$, and $a \geq 1$, Lemma C.3 gives

$$Q_t(a,s) \overset{⑤}{=} \frac{(t/n)R_\star(a,s)}{1 - Bt/(an)} \overset{⑥}{\leq} \frac{2t}{n} R_\star(a,s), \qquad t \in \{m,u\}. \tag{C.5}$$

Step ⑤ solves the linear absorption inequality. Step ⑥ uses $1 - Bt/(an) \geq 1 - t/(2n) > 1/2$.

Summing (C.5) and using $m + u = n$ gives $Q_u(a,s) + Q_m(a,s) \leq 2R_\star(a,s)$. As a result, (C.1) is at most $(2 + 2B/a)R_\star(a,s) \leq 3R_\star(a,s)$. This sharper use of the single remainder preserves the kernel constants below. Finally, the two events from Step 1 have total failure probability at most $2\exp(-s) = \exp(-x)$ over the random split. Selecting a deterministic branch of W and a deterministic a creates no additional event. □

C.3 Bernstein geometries and an unbounded squared-loss example

The oracle-existence and uncentered Bernstein assumptions appear in Tolstikhin et al. (2014, Assumptions 10(1)–(2)) and Yang (2025b; 2026, Assumption 1). At $\theta = 1$, their condition $P_n g_f^2 \leq B P_n g_f$ implies Assumption 1(2), since $\mathrm{Var}[g_f] = P_n g_f^2 - (P_n g_f)^2 \leq P_n g_f^2$. For general $0 < \theta \leq 1$, our assumption makes variance tend to zero with full-sample excess, which is the link needed for localization. This is a structural condition on the loss class, not a consequence of boundedness alone. The variance–mean formulation is already used in Bartlett et al. (2005, Theorem 3.3). For Assumption 1(1), continuity of the full-sample loss on a nonempty compact set of prediction vectors suffices for attainment. Finite coordinates alone do not guarantee attainment on an arbitrary class. In our unbounded TKL application, Lemma D.1 proves existence separately and derives the required Bernstein geometry from quadratic curvature on the convex hull of subset solutions.

The variance in Assumption 1(2), rather than the uncentered second moment, is essential to allow $B_\infty(\mathcal{G}_\star) = \infty$. Indeed, the stronger condition $P_n g_f^2 \leq B\,(P_n g_f)^\theta$ would imply by Jensen's inequality that $(P_n g_f)^2 \overset{①}{\leq} B(P_n g_f)^\theta$. If $P_n g_f > 0$, division by $(P_n g_f)^\theta$ gives $(P_n g_f)^{2-\theta} \overset{②}{\leq} B$, hence $P_n g_f \overset{③}{\leq} B^{1/(2-\theta)}$ because $2 - \theta > 0$. If $P_n g_f = 0$, the stronger condition directly gives $P_n g_f^2 = 0$. In both cases, substituting the mean bound back into that condition yields $P_n g_f^2 \overset{④}{\leq} B^{2/(2-\theta)}$. Since $\mathrm{Var}[g_f] \leq P_n g_f^2$, the finite-population inequality in Step 1 of the proof of Lemma B.2 would therefore give $B_\infty(\mathcal{G}_\star) \overset{⑤}{\leq} \sqrt{n-1}\,B^{1/(2-\theta)} < \infty$. As a result, no class with $B_\infty(\mathcal{G}_\star) = \infty$ can satisfy that stronger condition with one finite B. Steps ①–⑤ are justified in the preceding calculation by Jensen's inequality, division by the positive mean, taking a positive power, substitution, and the variance bound of Lemma B.2, respectively.

We first verify three sufficient empirical geometries, then give the unbounded example cited after Assumption 1.

Lemma C.4 (Three Sufficient Bernstein Geometries)**.** Suppose Assumption 1(1) holds, and write $g_f = \ell_f - \ell_{f_n^\star}$. The notation $\mathbb{1}_{\{f \neq f_n^\star\}}$ denotes the function on $[n]$ with value $\mathbb{1}_{\{f(\mathbf{x}_i) \neq f_n^\star(\mathbf{x}_i)\}}$ at index i. Each of the following conditions implies Assumption 1(2) with the indicated parameters.

(1) There is $K > 0$ such that $0 \leq g_f(i) \leq K$ for every $i \in [n]$ and $f \in \mathcal{F}$. Then $B = K$ and $\theta = 1$ suffice.

(2) The loss is squared loss, $y_i = f_n^\star(\mathbf{x}_i)$ for every i, and there is $\Lambda > 0$ such that
$$P_n(f - f_n^\star)^4 \leq \Lambda^2 P_n(f - f_n^\star)^2, \qquad f \in \mathcal{F}.$$
Then $B = \Lambda^2$ and $\theta = 1$ suffice. A sufficient condition for this moment inequality is $\max_i |f(\mathbf{x}_i) - f_n^\star(\mathbf{x}_i)| \leq \Lambda$ for every $f \in \mathcal{F}$.

(3) The labels and all predictions lie in $\{0,1\}$, the loss is zero–one loss, and there are $B > 0$ and $0 < \theta \leq 1$ such that
$$P_n \mathbb{1}_{\{f \neq f_n^\star\}} \leq B(P_n g_f)^\theta, \qquad f \in \mathcal{F}.$$
Then those same B and θ suffice.

In fact, each setting gives the stronger uncentered inequality $P_n g_f^2 \le B(P_n g_f)^\theta$.

Proof of Lemma C.4. Recall $g_f = \ell_f - \ell_{f_n^\star}$ and that P_n averages over the fixed sample. We use $\mathrm{Var}\,[g_f] \le P_n g_f^2$ in all three settings. Oracle optimality gives $P_n g_f \ge 0$, and the constants are uniform in f.

Proof outline.

(1) Nonnegative excess. Bound its square by a multiple of itself.
(2) Squared loss. Apply the fourth-to-second moment condition.
(3) Binary classification. Express the squared excess as disagreement.

Step 1: Nonnegative excess. The pointwise bound $0 \le g_f \le K$ implies $g_f^2 \le K g_f$. Averaging gives $\mathrm{Var}\,[g_f] \le P_n g_f^2 \le K P_n g_f$.

Step 2: Squared loss. When $y_i = f_n^\star(\mathbf{x}_i)$, the oracle loss is zero and $g_f = (f - f_n^\star)^2$. The moment hypothesis gives

$$\mathrm{Var}\,[g_f] \le P_n g_f^2 = P_n (f - f_n^\star)^4 \overset{①}{\le} \Lambda^2 P_n (f - f_n^\star)^2 = \Lambda^2 P_n g_f.$$

Step ① is the assumed fourth-to-second moment bound. The sufficient condition $|f(\mathbf{x}_i) - f_n^\star(\mathbf{x}_i)| \le \Lambda$ implies this bound pointwise before averaging, including when the second moment is zero.

Step 3: Binary classification. Under zero–one loss, agreeing binary predictions have identical losses, while disagreeing predictions give $g_f(i) \in \{-1, 1\}$. Consequently $g_f(i)^2 = \mathbb{1}_{\{f(\mathbf{x}_i) \ne f_n^\star(\mathbf{x}_i)\}}$, and the margin condition gives $\mathrm{Var}\,[g_f] \le P_n g_f^2 = P_n \mathbb{1}_{\{f \ne f_n^\star\}} \le B(P_n g_f)^\theta$. Each calculation proves both the second-moment and variance bounds. □

These geometries are empirical counterparts of the Bernstein and margin conditions in local-complexity fast rates (Bartlett et al., 2005). They control an uncentered second moment and hence force a finite envelope on a fixed full sample, as shown at the start of this subsection. The next example instead uses centered variance to allow $B_\infty(\mathcal{G}_\star) = \infty$.

Example C.5 (Unbounded Squared Loss With An Infinite Envelope)**.** Let n be even, let $\mathbf{z} = (z_i)_{i=1}^n$ satisfy $z_i \in \{-1, 1\}$ and $P_n \mathbf{z} = 0$, and put $\mathbf{x}_i = (1, z_i)^\top \in \mathbb{R}^2$ and $y_i = 0$. Consider

$$\mathcal{F} = \left\{ f_{\mathbf{w}} : f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \quad \mathbf{w} = (a, b)^\top \in \mathbb{R}^2, \quad \min\{|a|, |b|\} \le 1 \right\}$$

with squared loss $\ell_{f_{\mathbf{w}}}(i) = (a + b z_i)^2$. Both coefficients can vary and either can be arbitrarily large, although they cannot both exceed one in absolute value. The full-sample oracle is $f_n^\star = f_{\mathbf{0}}$, and every nonempty subset admits the empirical minimizer $f_{\mathbf{0}}$. Write $g_f = \ell_f - \ell_{f_{\mathbf{0}}}$ and $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}\}$. Assumption 1 holds with $B = 4$ and $\theta = 1$, while $B_\infty(\mathcal{G}_\star) = \infty$. For $V_\star(g_f) = 4 P_n g_f$, a sub-root majorant satisfying (4.1) and its fixed point are

$$\psi_\star(r) = \sqrt{\frac{r(n - N_{u,m})}{N_{u,m}(n-1)}}, \qquad r_{\mathrm{exc}} = \frac{n - N_{u,m}}{N_{u,m}(n-1)} \le \frac{1}{N_{u,m}}.$$

As a result, all hypotheses of Theorem 4.2 are compatible with $B_\infty(\mathcal{G}_\star) = \infty$, and this fixed point vanishes whenever $N_{u,m} \to \infty$. The constant choice of $f_{\mathbf{0}}$ on every subset has zero excess risk, so the example asserts compatibility rather than a risk lower bound. Removing the coefficient restriction makes Assumption 1(2) fail for every finite B and $0 < \theta \le 1$.

Proof of Example C.5. Recall the balanced vector $\mathbf{z} = (z_i)_{i=1}^n$, with $z_i \in \{-1, 1\}$ and $P_n \mathbf{z} = 0$. The inputs are $\mathbf{x}_i = (1, z_i)^\top$, and the labels are zero. The predictors are $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ with $\mathbf{w} = (a, b)^\top$ and $\min\{|a|, |b|\} \le 1$. As a result, $\ell_{f_{\mathbf{w}}}(i) = (a + b z_i)^2$ is squared loss, $f_n^\star = f_{\mathbf{0}}$, and $g_{f_{\mathbf{w}}} = \ell_{f_{\mathbf{w}}}$. The oracle class is $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}\}$, its surrogate is $V_\star(g_f) = 4 P_n g_f$, and $\mathcal{G}_\star(r) = \{g_f \in \mathcal{G}_\star : V_\star(g_f) \le r\}$.

Proof outline.

(1) Oracle geometry and localization. Verify existence of every required minimum, the centered Bernstein inequality, and the centered localized class.
(2) Exact sampling variance. We expand the squared subset average using the one-coordinate and two-coordinate inclusion probabilities.
(3) TC majorant and fixed point. We apply Cauchy–Schwarz to the first-order TC, observe that all centered squares are constant, and solve the oracle fixed-point equation.
(4) Unboundedness and the coefficient restriction. Exhibit a sequence with diverging centered coordinates, and show why the unrestricted two-dimensional coefficient class fails the Bernstein condition.

Step 1: Oracle geometry and localization. We first separate the mean and fluctuating parts of the squared loss to identify both curvature and localization. Every loss coordinate is finite for finite coefficients. On each nonempty subset, squared loss is nonnegative and the zero coefficient vector is feasible with loss zero. As a result, $f_{\mathbf{0}}$ attains every required empirical minimum. Its constant selection is measurable and has zero excess risk. In particular we may take $f_n^\star = f_{\mathbf{0}}$, so the excess loss equals the loss itself. Since $P_n\mathbf{z} = 0$ and $P_n\mathbf{z}^2 = 1$,

$$g_{f_{\mathbf{w}}}(i) \overset{①}{=} a^2 + b^2 + 2abz_i, \qquad P_n g_{f_{\mathbf{w}}} \overset{②}{=} a^2 + b^2,$$
$$\mathrm{Var}\left[g_{f_{\mathbf{w}}}\right] \overset{③}{=} 4a^2b^2 \overset{④}{\leq} 4(a^2+b^2) = V_\star(g_{f_{\mathbf{w}}}).$$

Steps ①–④ expand the loss and use $P_n\mathbf{z} = 0$, $P_n\mathbf{z}^2 = 1$, and $\min\{a^2, b^2\} \leq 1$, respectively.

For the inequality, $\min\left\{a^2, b^2\right\} \leq 1$ gives $a^2b^2 \leq \max\left\{a^2, b^2\right\} \leq a^2 + b^2$. This proves both parts of Assumption 1 with the stated parameters. The mean formula also shows that the full-sample minimum is attained only at the zero coefficient vector. Every centered member of $\mathcal{G}_\star(r)$ consequently has sample values cz_i with $|c| \leq \sqrt{r}$: its variance is c^2, bounded by its surrogate and hence by r. Only this containment is used. The localized class contains zero because the zero predictor is feasible.

Step 2: Exact sampling variance. The centered class depends on one balanced vector, so its exact subset variance supplies the needed complexity scale. Let A_p be a uniform p-subset of $[n]$, with $1 \leq p \leq n$. Under this choice of A_p, each index is included with probability p/n, and each pair of distinct indices with probability $p(p-1)/(n(n-1))$. Moreover, $\sum_i z_i^2 = n$ and $\sum_{i\neq j} z_i z_j = (\sum_i z_i)^2 - \sum_i z_i^2 = -n$. Expansion of the squared average therefore gives

$$\mathbb{E}_{A_p}\left[(P_{A_p}\mathbf{z})^2\right] \overset{⑤}{=} \frac{1}{p^2}\left(\frac{p}{n}\sum_i z_i^2 + \frac{p(p-1)}{n(n-1)}\sum_{i\neq j} z_i z_j\right) \overset{⑥}{=} \frac{n-p}{p(n-1)}.$$

Step ⑤ uses the single-index and pair inclusion probabilities. Step ⑥ uses $\sum_i z_i^2 = n$ and $\sum_{i\neq j} z_i z_j = -n$.

Also $\mathbb{E}_{A_p}\left[P_{A_p}\mathbf{z}\right] = 0$. As a result, the displayed second moment is its exact variance, including $p = n$.

Step 3: TC majorant and fixed point. The first-order process uses the subset variance, whereas constant centered squares make the second-order process vanish. Fix $r \geq 0$ and $1 \leq p < n$. For each sign η and every p-subset, Step 1 gives $\sup_{h\in\mathcal{G}_\star(r)} R_p^\eta h \leq \sqrt{r}\,\left|P_{A_p}\mathbf{z}\right|$. Taking expectations and using Step 2 yields

$$\max_{\eta\in\{+,-\}} \mathfrak{R}_p^\eta(\mathcal{G}_\star(r)) \overset{⑦}{\leq} \sqrt{r}\,\mathbb{E}_{A_p}\left[\left|P_{A_p}\mathbf{z}\right|\right] \overset{⑧}{\leq} \sqrt{\frac{r(n-p)}{p(n-1)}}.$$

Here ⑦ enlarges the centered localized class to $\{c\mathbf{z} : |c| \leq \sqrt{r}\}$. ⑧ uses Cauchy–Schwarz and the exact second moment from Step 2.

Each centered square equals c^2 at every coordinate, hence $\mathfrak{R}_p^+(\mathcal{G}_\star(r)^{[2]}) = 0$. At $r = 0$, the coefficient c is zero and both TC quantities vanish as well. As a result, taking $p = N_{u,m}$ proves (4.1) for $\mathcal{G}_\star$ and $V_\star$ at every radius required in the example. The chosen oracle majorant is explicitly $\psi_\star(r) = \sqrt{r(n-N_{u,m})/[N_{u,m}(n-1)]}$. This finite function is nondecreasing, and its ratio to $\sqrt{r}$

is constant for $r > 0$, so it is sub-root. It is nonzero because $N_{u,m} < n$. Writing $\psi_\star(r) = r$ and dividing by $\sqrt{r}$ for $r > 0$ therefore gives $r_{\mathrm{exc}} = (n - N_{u,m})/(N_{u,m}(n-1))$. This is at most $1/N_{u,m}$ and tends to zero whenever $N_{u,m} \to \infty$.

Step 4: Unboundedness and the coefficient restriction. We check separately that the envelope is infinite and that the coefficient restriction is essential to the Bernstein inequality. Take $a = 1$ and let $|b| \to \infty$ within the permitted parameter set. The centered excess has value $2bz_i$ at index i, with maximum absolute value $2|b|$. Hence $B_\infty(\mathcal{G}_\star) = \infty$. Finally, if the restriction is removed, the coefficients $a = b = t$ are permitted for every real t. The same moment calculation then gives $\mathrm{Var}[g_{f_{\mathbf{w}}}] = 4t^4$ and $P_n g_{f_{\mathbf{w}}} = 2t^2$. For $t \neq 0$ and any $0 < \theta \leq 1$,

$$\frac{\mathrm{Var}[g_{f_{\mathbf{w}}}]}{(P_n g_{f_{\mathbf{w}}})^\theta} \overset{⑨}{=} 2^{2-\theta}\,|t|^{4-2\theta} \overset{①}{\longrightarrow} \infty \quad \text{as } |t| \longrightarrow \infty.$$

Here ⑨ substitutes the variance and mean along $a = b = t$. ① uses $4 - 2\theta > 0$.

The exponent $4 - 2\theta$ is positive. Therefore no finite B can satisfy the centered Bernstein inequality on that unrestricted class. □

D Proofs for Unbounded Transductive Kernel Learning

D.1 Deterministic localization and localized TC

Recall from Section 4.3 that $\mathcal{F} = \mathcal{H}_{\mathbf{X}_n} = \mathrm{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. We use $\mathcal{F}$ for this finite-dimensional prediction space throughout, and $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. For $\lambda \geq 0$, the fitting criterion is $J_A(f) = P_A(f - y)^2 + \lambda\|f\|^2_{\mathcal{H}_K}$. The ordinary target is $\mathcal{E}_{\mathcal{U}}$ defined in Section 4.3, with test comparator $\widehat{f}_{u,0}$. Define $\ell_{\lambda,f}(i) = (f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|^2_{\mathcal{H}_K}$ and the intermediate regularized excess $\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) = J_{\mathcal{U}}(\widehat{f}_{m,\lambda}) - J_{\mathcal{U}}(\widehat{f}_{u,\lambda})$. The subset solution is $\widehat{f}_{A,\lambda}$, with minimum-RKHS-norm tie-breaking at $\lambda = 0$. Throughout this appendix, a relevant subset means $A \subseteq [n]$ with $|A| \in \{m, u, n\}$. We define $\Pi_\lambda : \mathcal{H}_K \to \mathcal{H}_K$ by

$$\Pi_\lambda f := \underset{h \in \mathrm{span}\left\{\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda} : A,B \subseteq [n],\ |A|,|B| \in \{m,u,n\}\right\}}{\arg\min} \|f - h\|^2_{\mathcal{H}_K}, \qquad f \in \mathcal{H}_K. \tag{D.1}$$

The span in (D.1) is finite dimensional and closed, so the nearest point exists uniquely. As a result, Π_λ is the RKHS orthogonal projection onto the span of differences between subset solutions. The projected Gram matrix has entries $(\mathbf{K}_\lambda)_{ij} = \langle \Pi_\lambda K(\cdot, \mathbf{x}_i), \Pi_\lambda K(\cdot, \mathbf{x}_j)\rangle_{\mathcal{H}_K}$, and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n \geq 0$ are the eigenvalues of $\mathbf{K}_\lambda/n$, counted with multiplicity. Unless a positive-eigenvalue restriction is stated, spectral sums run over $j \in [n]$. For an integer cutoff $Q \in \{0, \ldots, n\}$, the head consists of indices $1 \leq j \leq Q$ and the tail of indices $Q < j \leq n$. For general ridge, define the diameter D_λ and residual maximum L_λ by

$$D_\lambda = \max_{\substack{A,B \subseteq [n] \\ |A|,|B| \in \{m,u,n\}}} \left\|\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda}\right\|_{\mathcal{H}_K}, \qquad L_\lambda = \max_{\substack{A \subseteq [n],\ |A| \in \{m,u,n\} \\ i \in [n],\ (\mathbf{K}_\lambda)_{ii} > 0}} \left(\widehat{f}_{A,\lambda}(\mathbf{x}_i) - y_i\right)^2.$$

These definitions agree with D_0, L_0 in Section 4.3. The subsets in the diameter need not be disjoint or complementary. The diameter, residual maximum, and projected spectrum are fixed before the split. Empty residual maxima and spectral sums are zero, and the ratio $\widehat{\lambda}_j/(\widehat{\lambda}_j + \lambda)$ is zero when both entries vanish. Define the spectral fixed point in squared-loss units by

$$r_{K,\lambda} = \inf\left\{ r > 0 : \frac{L_\lambda}{N_{u,m}} \sum_{j=1}^{n} \min\left\{ r\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, D_\lambda^2 \widehat{\lambda}_j \right\} \leq r^2 \right\}.$$

It is finite, since the spectral sum is at most $D_\lambda^2 \sum_j \widehat{\lambda}_j < \infty$, and it is zero if $L_\lambda = 0$. The generic fixed point r_{exc} below is for losses divided by L_λ, so its comparison uses $r_{K,\lambda}/L_\lambda$ when $L_\lambda > 0$.

For the ridge restrictions, let R_λ be the largest norm of a subset solution,

$$R_\lambda^2 := \max_{\substack{A \subseteq [n] \\ |A| \in \{m,u,n\}}} \left\|\widehat{f}_{A,\lambda}\right\|^2_{\mathcal{H}_K}. \tag{D.2}$$

When $\mathbf{K} \neq \mathbf{0}$, let $K^{+}_{\min}$ be the smallest positive eigenvalue among $\mathbf{K}_A/|A|$ over relevant subsets. The collection is finite and contains a positive eigenvalue from $A = [n]$. The scales in (D.2) and above are determined before the split and may depend on all full-sample labels.

How projection sharpens the bound. Projection affects the spectrum controlling prediction differences and the residual maximum L_λ.

Spectral complexity. For fixed $\lambda, D_\lambda, L_\lambda$, each summand is nondecreasing in its eigenvalue. As a result, $\widehat{\lambda}_j \le \lambda_j(\mathbf{K}/n)$ can only decrease $r_{K,\lambda}$ and the cutoff bounds in Corollary D.4. For a cutoff $Q \in \{0, \dots, n\}$, the head term $\sum_{j \le Q} \widehat{\lambda}_j/(\widehat{\lambda}_j + \lambda)$ counts positive projected eigenvalues at zero ridge and weights them by numbers in $(0, 1)$ otherwise. It therefore measures only directions in the span of differences between subset solutions. If $L_\lambda > 0$ and $\mathbf{K}_\lambda \neq \mathbf{K}$, this span is a proper subspace of $\mathcal{F}$, and projection strictly decreases the full-spectrum fixed point with the same diameter and residual maximum. Step 4 of Lemma D.1 proves both fixed-point comparisons. The identity-kernel example in Section D.6 also gives a strictly faster rate.

Residual maximum. A coordinate with $(\mathbf{K}_\lambda)_{ii} = 0$ is exactly one at which all subset solutions agree, as Step 1 of Lemma D.1 proves from $(\mathbf{K}_\lambda)_{ii} = \|\Pi_\lambda K(\cdot, \mathbf{x}_i)\|^2_{\mathcal{H}_K}$. Their convex combinations agree there as well, so the common squared loss cancels from every comparison within the hull. Excluding such coordinates from L_λ can reduce both the fixed point and the confidence term $L_\lambda x/N_{u,m}$ in Theorem D.3. The ridge penalty remains in the fitting objective, but cancels from centered split deviations because it is constant across coordinates for each predictor. Both reductions preserve the learner and require no additional probability event.

In the full-span TKL setup above, the subset solution $\widehat{f}_{A,\lambda}$ exists and is uniquely specified whenever $|A| \in \{m, u, n\}$, as proved in Lemma D.1. We define

$$\mathcal{F}^{\mathrm{sel}}_\lambda := \left\{ \sum_{\substack{A \subseteq [n] \\ |A| \in \{m,u,n\}}} w_A \widehat{f}_{A,\lambda} : w_A \ge 0 \text{ for all such } A, \quad \sum_{\substack{A \subseteq [n] \\ |A| \in \{m,u,n\}}} w_A = 1 \right\} \tag{D.3}$$

as the convex hull of all these finitely many subset solutions. Restricting the analysis to this hull preserves the training, test, and full-sample subset solutions and does not change the learner. At zero ridge this assertion concerns the zero-ridge subset solutions, not all least-squares minimizers.

The following normalization is used only when $L_\lambda > 0$. Put $f^\star_n = \widehat{f}_{[n],\lambda}$, $\widetilde{\ell}_f = \ell_{\lambda,f}/L_\lambda$, and $\widetilde{g}_f = \widetilde{\ell}_f - \widetilde{\ell}_{f^\star_n}$. The sole loss class needed in Theorem 4.2 is

$$\mathcal{G}_\star = \left\{\widetilde{g}_f : f \in \mathcal{F}^{\mathrm{sel}}_\lambda\right\}, \quad V_\star(\widetilde{g}_f) = 4P_n\widetilde{g}_f, \quad \mathcal{G}_\star(r) = \left\{\widetilde{g}_f \in \mathcal{G}_\star : 4P_n\widetilde{g}_f \le r\right\}, \qquad r \ge 0.$$

The next lemma verifies the variance bound for $V_\star$ and the localized TC condition (4.1) for $\mathcal{G}_\star$. It requires no Bernstein assumption on the unrestricted span.

Lemma D.1 (Deterministic Localization Of Subset Solutions And Localized TC)**.** Consider the full-span TKL setup of Section 4.3. The relevant subset solutions exist and are uniquely specified. Their hull preserves the training, test, and full-sample subset solutions. The projected Gram matrix satisfies $\mathbf{0} \preceq \mathbf{K}_\lambda \preceq \mathbf{K}$, $\operatorname{rank}(\mathbf{K}_\lambda) = \dim \operatorname{range}(\Pi_\lambda) \le \operatorname{rank}(\mathbf{K})$, and $\widehat{\lambda}_j \le \lambda_j(\mathbf{K}/n)$. Equality $\mathbf{K}_\lambda = \mathbf{K}$ holds exactly when the differences between relevant subset solutions span $\mathcal{F}$. If $L_\lambda = 0$, all relevant subset solutions coincide and $\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) = 0$ for every split. If $L_\lambda > 0$, the normalized oracle excess class satisfies Assumption 1 with $B = 4$, $\theta = 1$, and $B_\infty(\mathcal{G}_\star) \le 2$. For $L_\lambda > 0$ and $C_0 = 256\sqrt{2}$, the sub-root function

$$\psi_\lambda(r) = C_0 \sqrt{\frac{1}{N_{u,m}} \sum_{j=1}^{n} \min\left\{ r\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \frac{D^2_\lambda}{L_\lambda}\widehat{\lambda}_j \right\}}, \qquad r \ge 0, \tag{D.4}$$

satisfies the localized TC condition

$$\max\left\{ \max_{\eta \in \{+,-\}} \mathfrak{R}^{\eta}_{N_{u,m}}\left(\mathcal{G}_\star(r)\right), \mathfrak{R}^{+}_{N_{u,m}}\left(\mathcal{G}_\star(r)^{[2]}\right) \right\} \le \psi_\lambda(r), \qquad r \ge 0. \tag{D.5}$$

With $\psi_\star = \psi_\lambda$, its fixed point obeys $r_{\mathrm{exc}} \le C_0^2 r_{K,\lambda}/L_\lambda$. These bounds concern normalized oracle excesses on the hull, rather than losses on the unrestricted optimization domain.

Proof of Lemma D.1. Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. For $L_\lambda > 0$, the normalized coordinate loss is $\widetilde{\ell}_f(i) = [(f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K}]/L_\lambda$, and its reference excess is $\widetilde{g}_f = \widetilde{\ell}_f - \widetilde{\ell}_{f_n^\star}$ with $f_n^\star = \widehat{f}_{[n],\lambda}$. Here $\mathcal{G}_\star = \{\widetilde{g}_f : f \in \mathcal{F}_\lambda^{\mathrm{sel}}\}$, $V_\star(\widetilde{g}_f) = 4P_n\widetilde{g}_f$, and $\mathcal{G}_\star(r) = \{\widetilde{g}_f \in \mathcal{G}_\star : V_\star(\widetilde{g}_f) \le r\}$. The hull contains convex combinations of all subset solutions on subsets of sizes m, u, n and is fixed before the split. We treat $L_\lambda = 0$ before using normalized losses.

Proof outline.

(1) Existence and localization. Derive the ridge and minimum-norm representer formulas and verify convergence of gradient descent from zero. Establish compactness and preservation of the subset solutions. Prove the projected matrix order, rank, and equality characterization. Identify coordinates with common predictions and handle the zero residual maximum.
(2) Curvature and envelopes. Expand the full-sample objective at its minimizer and cancel its linear term by the normal equation. Normalize to prove the oracle Bernstein condition with $B = 4$ and $B_\infty(\mathcal{G}_\star) \le 2$.
(3) Weighted spectral head–tail bound. Expand differences in the projected empirical eigenbasis. Control the spectral head by regularized curvature and the tail by the RKHS diameter, then choose the cutoff that recovers the coordinatewise spectral minimum.
(4) Localized TC and fixed points. Convert oracle localization to the prediction constraints. Contract the loss on the prediction intervals determined by subset solutions, compare sampling schemes, and contract centered squares. Track the constants, verify the sub-root property, and compare the single fixed point with the spectral fixed point. Finally prove the full-spectrum comparison with the diameter and residual maximum held fixed.

Step 1: Existence and localization. The analysis will use only the prescribed subset solutions. We first justify their existence and the finite geometry that contains all of them.

Subset solutions at positive and zero ridge. For nonempty $A \subseteq [n]$, we define the evaluation operator $T_A : \mathcal{F} \to \mathbb{R}^{|A|}$ by $T_A f = (f(\mathbf{x}_i))_{i \in A}$ and put $\mathbf{y}_A = (y_i)_{i \in A}$. The criterion is then $J_A(f) = |A|^{-1} \|T_A f - \mathbf{y}_A\|_2^2 + \lambda \|f\|^2_{\mathcal{H}_K}$. The coordinates in A are ordered increasingly throughout the matrix formulas. Let $T_A^* : \mathbb{R}^{|A|} \to \mathcal{F}$ denote the adjoint for the Euclidean and RKHS inner products. For $\mathbf{c} \in \mathbb{R}^{|A|}$, the reproducing property gives $T_A^*\mathbf{c} = \sum_{i \in A} c_i K(\cdot, \mathbf{x}_i)$, where c_i denotes the coordinate indexed by $i \in A$. Consequently $T_A T_A^* = \mathbf{K}_A$, where $\mathbf{K}_A = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j \in A}$.

Indeed, $\langle T_A f, \mathbf{c} \rangle = \langle f, \sum_{i \in A} c_i K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_K}$, and evaluating the sum on A gives $\mathbf{K}_A \mathbf{c}$. The reproducing inequality also gives $\|T_A f\|_2^2 \le \sum_{i \in A} K(\mathbf{x}_i, \mathbf{x}_i) \|f\|^2_{\mathcal{H}_K}$, so T_A is continuous.

Recall that $\mathbf{I}_p$ is the identity matrix of order p. We use the superscript $\dagger$ for the Moore–Penrose inverse. For the symmetric positive-semidefinite matrices used here, this inverse reciprocates each positive eigenvalue and sends each zero eigendirection to zero. We record the directional calculation used for both ridge regimes. For $f, v \in \mathcal{F}$ and $t \in \mathbb{R}$, expansion of the finite sums and the RKHS norm gives

$$J_A(f + tv) - J_A(f) \overset{①}{=} 2t \left[P_A[(f - y)v] + \lambda \langle f, v \rangle_{\mathcal{H}_K} \right] + t^2 \left[P_A v^2 + \lambda \|v\|^2_{\mathcal{H}_K} \right].$$

Here ① expands the data-fit square and the squared RKHS norm.

Both signs of t are feasible, so the first bracket vanishes at a minimizer. Conversely, its vanishing for every v proves minimality through the nonnegative quadratic term at $t = 1$.

When $\lambda > 0$, the squared-norm penalty makes J_A strictly convex and coercive. Continuity and finite dimension therefore give a unique minimizer. To identify it, set $\mathbf{c} = (\mathbf{K}_A + \lambda |A| \mathbf{I}_{|A|})^{-1} \mathbf{y}_A$. Adjointness and $T_A T_A^* = \mathbf{K}_A$ give, for every $v \in \mathcal{F}$,

$$P_A[(T_A^*\mathbf{c} - y)v] + \lambda \langle T_A^*\mathbf{c}, v \rangle_{\mathcal{H}_K} = \frac{1}{|A|} \left\langle (\mathbf{K}_A + \lambda |A| \mathbf{I}_{|A|})\mathbf{c} - \mathbf{y}_A, T_A v \right\rangle = 0.$$

The directional criterion just proved identifies $\widehat{f}_{A,\lambda} = T_A^*(\mathbf{K}_A + \lambda |A| \mathbf{I}_{|A|})^{-1} \mathbf{y}_A$.

For $\lambda = 0$, we identify the least-squares prediction through $\operatorname{range}(T_A) = \operatorname{range}(\mathbf{K}_A)$, where $\mathbf{K}_A = T_A T_A^*$. Orthogonal complements below are in $\mathbb{R}^{|A|}$, and $\operatorname{null}(S)$ denotes the null space of an operator S.

The identity $\langle \mathbf{c}, T_A T_A^* \mathbf{c} \rangle = \|T_A^* \mathbf{c}\|_{\mathcal{H}_K}^2$ shows that $\text{null}(T_A T_A^*) = \text{null}(T_A^*)$. For a linear map S between finite-dimensional inner-product spaces, adjointness gives $\text{range}(S)^\perp = \text{null}(S^*)$, and its range is closed. Applying this fact to T_A and the self-adjoint map $T_A T_A^*$ yields

$$\text{range}(T_A T_A^*) \overset{②}{=} \text{null}(T_A T_A^*)^\perp = \text{null}(T_A^*)^\perp = \text{range}(T_A).$$

Step ② uses the orthogonal-complement identity and the common null space. This includes singular and zero Gram matrices.

The Moore–Penrose formula makes $\mathbf{K}_A \mathbf{K}_A^\dagger$ the orthogonal projection onto $\text{range}(T_A)$. This identifies the least-squares predictions without assuming interpolation. For every $f \in \mathcal{F}$,

$$\|T_A f - \mathbf{y}_A\|_2^2 \overset{③}{=} \left\| T_A f - \mathbf{K}_A \mathbf{K}_A^\dagger \mathbf{y}_A \right\|_2^2 + \left\| (\mathbf{I}_{|A|} - \mathbf{K}_A \mathbf{K}_A^\dagger) \mathbf{y}_A \right\|_2^2.$$

Here ③ applies Pythagoras to the projection onto $\text{range}(T_A)$.

The second term is fixed. The first vanishes for $\widehat{f}_{A,0} = T_A^* \mathbf{K}_A^\dagger \mathbf{y}_A$, since $T_A \widehat{f}_{A,0} = \mathbf{K}_A \mathbf{K}_A^\dagger \mathbf{y}_A$. Every minimizer therefore has the form $\widehat{f}_{A,0} + v$ with $T_A v = 0$.

Adjointness gives $\widehat{f}_{A,0} \perp \text{null}(T_A)$. Hence $\left\| \widehat{f}_{A,0} + v \right\|_{\mathcal{H}_K}^2 = \left\| \widehat{f}_{A,0} \right\|_{\mathcal{H}_K}^2 + \|v\|_{\mathcal{H}_K}^2$ for $v \in \text{null}(T_A)$, which proves uniqueness of the minimum-norm selection. The Gram norm identity $\|T_A^* \mathbf{c}\|_{\mathcal{H}_K}^2 = \mathbf{c}^\top \mathbf{K}_A \mathbf{c}$ and $\mathbf{K}_A^\dagger \mathbf{K}_A \mathbf{K}_A^\dagger = \mathbf{K}_A^\dagger$ give

$$\left\| \widehat{f}_{A,0} \right\|_{\mathcal{H}_K}^2 = \mathbf{y}_A^\top \mathbf{K}_A^\dagger \mathbf{y}_A,$$
$$\left\| \widehat{f}_{A,\lambda} \right\|_{\mathcal{H}_K}^2 = \mathbf{y}_A^\top (\mathbf{K}_A + \lambda |A| \mathbf{I}_{|A|})^{-1} \mathbf{K}_A (\mathbf{K}_A + \lambda |A| \mathbf{I}_{|A|})^{-1} \mathbf{y}_A, \qquad \lambda > 0.$$

Computing the zero-ridge subset solution by gradient descent. Set $\lambda = 0$. The directional calculation above identifies the RKHS gradient as $\nabla J_A(f) = 2 |A|^{-1} T_A^* (T_A f - \mathbf{y}_A)$. Recall that $\mathbf{K}_A = T_A T_A^*$ is the subset Gram matrix. When $\mathbf{K}_A \neq \mathbf{0}$, let $\lambda_{\max}(\mathbf{K}_A) > 0$ denote its largest eigenvalue and choose a constant step size $0 < \eta < |A| / \lambda_{\max}(\mathbf{K}_A)$. Starting from the zero function, for each integer $t \geq 0$ set

$$f^{(0)} = 0, \qquad f^{(t+1)} = f^{(t)} - \frac{2\eta}{|A|} \sum_{i \in A} \left(f^{(t)}(\mathbf{x}_i) - y_i \right) K(\cdot, \mathbf{x}_i). \tag{D.6}$$

Every iterate in (D.6) belongs to $\text{range}(T_A^*)$, as does $\widehat{f}_{A,0} = T_A^* \mathbf{K}_A^\dagger \mathbf{y}_A$. The least-squares normal equation gives $T_A^* (T_A \widehat{f}_{A,0} - \mathbf{y}_A) = 0$. Subtracting it from the update yields

$$f^{(t+1)} - \widehat{f}_{A,0} \overset{④}{=} f^{(t)} - \widehat{f}_{A,0} - \frac{2\eta}{|A|} T_A^* T_A \left(f^{(t)} - \widehat{f}_{A,0} \right).$$

Here ④ subtracts the least-squares normal equation from the gradient update.

To verify convergence, let $\mathbf{v}$ be a Euclidean unit eigenvector of $\mathbf{K}_A$ with eigenvalue $\mu > 0$. Then $e = T_A^* \mathbf{v} / \sqrt{\mu}$ has RKHS norm one, since $\|T_A^* \mathbf{v}\|_{\mathcal{H}_K}^2 = \mathbf{v}^\top \mathbf{K}_A \mathbf{v} = \mu$, and $T_A^* T_A e = \mu e$. An orthonormal eigenbasis for the positive eigenspaces of $\mathbf{K}_A$ therefore gives an RKHS orthonormal basis of $\text{range}(T_A^*)$. Indeed, the same inner-product calculation proves orthogonality, and $\text{null}(\mathbf{K}_A) = \text{null}(T_A^*)$ shows that the remaining eigenvectors have zero image under T_A^*. Write $(e_j)_{j=1}^{\text{rank}(\mathbf{K}_A)}$ for this RKHS orthonormal basis and $\mu_j > 0$ for its eigenvalues under $T_A^* T_A$. Taking the inner product of the error recursion with e_j multiplies that coefficient by $1 - 2\eta\mu_j / |A|$ at every iteration. The initial error is $-\widehat{f}_{A,0}$ because $f^{(0)} = 0$. Iteration and orthonormality therefore give

$$\left\| f^{(t)} - \widehat{f}_{A,0} \right\|_{\mathcal{H}_K}^2 \overset{⑤}{=} \sum_{j=1}^{\text{rank}(\mathbf{K}_A)} \left(1 - \frac{2\eta\mu_j}{|A|} \right)^{2t} \left| \left\langle \widehat{f}_{A,0}, e_j \right\rangle_{\mathcal{H}_K} \right|^2 \xrightarrow{⑥} 0.$$

Step ⑤ iterates the orthonormal coefficient recursion. In ⑥, the step-size restriction gives $|1 - 2\eta\mu_j / |A|| < 1$ for every term in the finite sum.

There is no orthogonal error component because both functions belong to range(T_A^*). If $\mathbf{K}_A = \mathbf{0}$, then $T_A^* = 0$, and every iterate and the subset solution are zero for any $\eta > 0$. This also covers singular Gram matrices and labels that cannot be interpolated.

For either ridge regime, the directional calculation gives $P_A[(\widehat{f}_{A,\lambda} - y)d] + \lambda \left\langle \widehat{f}_{A,\lambda}, d \right\rangle_{\mathcal{H}_K} = 0$ for every $d \in \mathcal{F}$. In operator form this is $|A|^{-1} T_A^*(T_A \widehat{f}_{A,\lambda} - \mathbf{y}_A) + \lambda \widehat{f}_{A,\lambda} = 0$. Cancelling the linear term in the same expansion gives

$$J_A(\widehat{f}_{A,\lambda} + d) - J_A(\widehat{f}_{A,\lambda}) \overset{⑦}{=} P_A d^2 + \lambda \|d\|_{\mathcal{H}_K}^2, \qquad d \in \mathcal{F}. \tag{D.7}$$

Step ⑦ uses the normal equation. This identity does not require interpolation, including at zero ridge.

A compact domain preserving all relevant subset solutions. The finite collection of subset solutions with subset size in $\{m, u, n\}$ and its convex hull $\mathcal{F}_\lambda^{\mathrm{sel}}$ are determined by the full sample. The hull is compact: it is the continuous image of the finite-dimensional compact simplex of nonnegative weights summing to one. It is convex by construction. For each relevant A, the hull contains $\widehat{f}_{A,\lambda}$ and is contained in the original optimization domain. Therefore

$$\inf_{f \in \mathcal{F}_\lambda^{\mathrm{sel}}} J_A(f) \overset{⑧}{=} J_A(\widehat{f}_{A,\lambda}) \overset{⑨}{=} \inf_{f \in \mathcal{F}} J_A(f).$$

Steps ⑧, ⑨ use the subset solution's membership in the hull and its optimality over $\mathcal{F}$.

At zero ridge every minimizer in the hull is also an original least-squares minimizer, whose norm is at least that of $\widehat{f}_{A,0}$. As a result, restriction preserves the specified minimum-norm selection as well. It need not preserve the entire set of original least-squares minimizers.

To verify the diameter claim, write $f = \sum_A a_A \widehat{f}_{A,\lambda}$ and $f' = \sum_B b_B \widehat{f}_{B,\lambda}$, where both coefficient families are nonnegative and sum to one. Both families are indexed by subsets of sizes in $\{m, u, n\}$.

Since both weight sums equal one,

$$f - f' = \sum_{A,B} a_A b_B (\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda}), \qquad \|f - f'\|_{\mathcal{H}_K} \le \sum_{A,B} a_A b_B D_\lambda = D_\lambda.$$

The first identity places every hull difference in the selected-solution difference span. The inequality is convexity of the norm.

A generating pair attains D_λ, because there are finitely many pairs, and both members belong to the hull. As a result, D_λ is exactly the hull's diameter. Finally, the subset solution is measurable as a function of the split because the split space is finite.

Projected kernel, eigenvalues, and rank. Recall that Π_λ projects onto the span of differences of relevant subset solutions. The projected kernel section at $\mathbf{x}_i$ is $\Pi_\lambda K(\cdot, \mathbf{x}_i)$, the component of the RKHS function $\mathbf{x} \mapsto K(\mathbf{x}, \mathbf{x}_i)$ in this span. The matrix $\mathbf{K}_\lambda$ is the Gram matrix of these n projected sections. The displayed expression for $f - f'$ and the definition (D.1) show that $\Pi_\lambda(f - f') = f - f'$. Consequently $f - \Pi_\lambda f = f' - \Pi_\lambda f'$ for any two hull members. For every $d \in \mathrm{range}(\Pi_\lambda)$, we have $\Pi_\lambda d = d$. Self-adjointness and the reproducing property therefore give

$$d(\mathbf{x}_i) \overset{①}{=} \langle d, K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_K} \overset{②}{=} \langle d, \Pi_\lambda K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_K}, \qquad i \in [n]. \tag{D.8}$$

Steps ①, ② use the reproducing property, self-adjointness of Π_λ, and $\Pi_\lambda d = d$.

The range of Π_λ is contained in $\mathcal{F}$. Since the kernel sections span that space, their projections span the range of Π_λ.

For $\boldsymbol{a} \in \mathbb{R}^n$, put $h = \sum_i a_i K(\cdot, \mathbf{x}_i)$. The Gram identity and the orthogonal decomposition $h = \Pi_\lambda h + (I - \Pi_\lambda)h$ give

$$\boldsymbol{a}^\top \mathbf{K}_\lambda \boldsymbol{a} \overset{③}{=} \|\Pi_\lambda h\|_{\mathcal{H}_K}^2 \ge 0, \qquad \boldsymbol{a}^\top (\mathbf{K} - \mathbf{K}_\lambda) \boldsymbol{a} \overset{④}{=} \|h - \Pi_\lambda h\|_{\mathcal{H}_K}^2 \ge 0. \tag{D.9}$$

Steps ③, ④ use the Gram identity and Pythagoras.

Equation (D.9) proves $\mathbf{0} \preceq \mathbf{K}_\lambda \preceq \mathbf{K}$, without requiring an inverse or unique coefficient representations.

For completeness, let E range over linear subspaces of $\mathbb{R}^n$. The variational formula for the jth largest eigenvalue gives

$$\widehat{\lambda}_j \overset{⑤}{=} \min_{\dim E = n-j+1} \max_{\substack{\boldsymbol{a}\in E \\ \|\boldsymbol{a}\|_2=1}} \frac{\boldsymbol{a}^\top \mathbf{K}_\lambda \boldsymbol{a}}{n} \overset{⑥}{\leq} \min_{\dim E = n-j+1} \max_{\substack{\boldsymbol{a}\in E \\ \|\boldsymbol{a}\|_2=1}} \frac{\boldsymbol{a}^\top \mathbf{K} \boldsymbol{a}}{n} \overset{⑦}{=} \lambda_j(\mathbf{K}/n).$$

Steps ⑤ and ⑦ use the eigenvalue variational formula. Step ⑥ applies (D.9) on each common subspace, without aligning the two eigenbases.

This is the Courant–Fischer variational formula, valid also for repeated and zero eigenvalues. Applying (D.9) to coordinate vectors gives $0 \leq (\mathbf{K}_\lambda)_{ii} \leq K(\mathbf{x}_i, \mathbf{x}_i)$.

For the rank identities, we apply rank–nullity to each Gram matrix and its coefficient-to-function map, whose null spaces coincide.

The original kernel sections.

For $A = [n]$, the map $T^*_{[n]} : \mathbb{R}^n \to \mathcal{F}$ is onto because $T^*_{[n]}\boldsymbol{a} = \sum_i a_i K(\cdot, \mathbf{x}_i)$ and these sections span $\mathcal{F}$. The common null-space identity above and rank–nullity give $\dim\mathcal{F} = \operatorname{rank}(\mathbf{K})$.

For the projected sections, linearity gives $\Pi_\lambda T^*_{[n]}\boldsymbol{a} = \sum_i a_i \Pi_\lambda K(\cdot, \mathbf{x}_i)$. Surjectivity of $T^*_{[n]}$ and $\operatorname{range}(\Pi_\lambda) \subseteq \mathcal{F}$ imply $\operatorname{range}(\Pi_\lambda T^*_{[n]}) = \operatorname{range}(\Pi_\lambda)$. The Gram identity $\boldsymbol{a}^\top \mathbf{K}_\lambda \boldsymbol{a} = \left\|\Pi_\lambda T^*_{[n]}\boldsymbol{a}\right\|^2_{\mathcal{H}_K}$ gives $\operatorname{null}(\mathbf{K}_\lambda) = \operatorname{null}(\Pi_\lambda T^*_{[n]})$. Rank–nullity on their common domain $\mathbb{R}^n$ now yields

$$\operatorname{rank}(\mathbf{K}_\lambda) \overset{⑧}{=} \dim\operatorname{range}(\Pi_\lambda T^*_{[n]}) = \dim\operatorname{range}(\Pi_\lambda) \leq \dim\mathcal{F} = \operatorname{rank}(\mathbf{K}).$$

Step ⑧ uses the common null space. The argument allows linearly dependent sections and zero-dimensional spans.

If $\mathbf{K}_\lambda = \mathbf{K}$, (D.9) implies $h = \Pi_\lambda h$ for every $h \in \mathcal{F}$, since every such h is a linear combination of the sections. Hence the differences between subset solutions span $\mathcal{F}$. Conversely, if they span that space, the projection fixes every section and $\mathbf{K}_\lambda = \mathbf{K}$. In particular, $\mathbf{K}_\lambda \neq \mathbf{K}$ means the projected rank is strictly smaller. At the other extreme, $\mathbf{K}_\lambda = \mathbf{0}$ exactly when $\Pi_\lambda = 0$, since the projected sections span its range. This is equivalent to $D_\lambda = 0$.

Residual maximum and the zero case. The coordinates entering L_λ are precisely those where some two relevant subset solutions disagree. Here $A, B \subseteq [n]$ have $|A|, |B| \in \{m, u, n\}$, and

$$\operatorname{range}(\Pi_\lambda) = \operatorname{span}\left\{\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda} : A, B \subseteq [n],\ |A|, |B| \in \{m, u, n\}\right\}.$$

An arbitrary $d \in \operatorname{range}(\Pi_\lambda)$ is a finite linear combination of these differences, and need not itself be one difference. Fix $i \in [n]$. The Gram diagonal identity gives $(\mathbf{K}_\lambda)_{ii} = \|\Pi_\lambda K(\cdot, \mathbf{x}_i)\|^2_{\mathcal{H}_K}$. If it is zero, (D.8) gives $d(\mathbf{x}_i) = 0$ for every d in the range, so all subset solutions agree there. Conversely, agreement of all subset solutions makes evaluation vanish on their difference span by linearity. Testing this functional at $d = \Pi_\lambda K(\cdot, \mathbf{x}_i)$ gives

$$0 = d(\mathbf{x}_i) \overset{⑨}{=} \|\Pi_\lambda K(\cdot, \mathbf{x}_i)\|^2_{\mathcal{H}_K} = (\mathbf{K}_\lambda)_{ii}.$$

Step ⑨ uses (D.8) and the membership of this projected section in the range. Consequently, $(\mathbf{K}_\lambda)_{ii} > 0$ exactly when some relevant pair disagrees there. Agreement extends to the convex hull by linearity, and the converse holds because the hull contains every subset solution.

On a coordinate with positive projected diagonal, a hull member $f = \sum_A w_A \widehat{f}_{A,\lambda}$ satisfies

$$|f(\mathbf{x}_i) - y_i| \leq \sum_A w_A \left|\widehat{f}_{A,\lambda}(\mathbf{x}_i) - y_i\right| \leq \sqrt{L_\lambda},$$

because $w_A \geq 0$, $\sum_A w_A = 1$, and every residual in the sum is bounded by $\sqrt{L_\lambda}$.

If $L_\lambda = 0$, all relevant subset solutions equal y_i there and agree with each other on every remaining coordinate. Their full-sample evaluation vectors therefore coincide, though this common vector need

not equal $\mathbf{y}$. The evaluation map is injective on $\mathcal{F}$. Indeed, if $f = \sum_i \alpha_i K(\cdot, \mathbf{x}_i)$ and $f(\mathbf{x}_i) = 0$ for every i, then $\|f\|^2_{\mathcal{H}_K} \overset{①}{=} \sum_i \alpha_i f(\mathbf{x}_i) = 0$. Here ① uses the finite representer expansion and the reproducing property. Hence all relevant subset solutions coincide, $D_\lambda = 0$, and $\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) = 0$ for every split. Ordinary excess will be compared with the unregularized test oracle in Lemma D.2. Conversely, $D_\lambda = 0$ makes $\mathbf{K}_\lambda = \mathbf{0}$, so $L_\lambda = 0$ under our convention that the supremum of a nonnegative function over an empty set is 0. As a result, $L_\lambda = 0$ and $D_\lambda = 0$ are equivalent. Their unnormalized loss-difference classes therefore consist only of zero, and both TC quantities vanish. No division by L_λ is made in this case. In what follows we assume $L_\lambda > 0$, which also gives $D_\lambda > 0$ and at least one positive eigenvalue of $\mathbf{K}_\lambda$.

Step 2: Curvature and envelopes. The normal equation removes the linear loss term, leaving quadratic prediction control from which the Bernstein inequality follows.

Curvature and the reference excess. We put $f_n^\star := \widehat{f}_{[n],\lambda}$. For $f \in \mathcal{F}_\lambda^{\mathrm{sel}}$, let $d = f - f_n^\star$.

The directional calculation in Step 1, with $A = [n]$, gives

$$\frac{1}{2}\frac{d}{dt} J_{[n]}(f_n^\star + td)\Big|_{t=0} = P_n[(f_n^\star - y)d] + \lambda \langle f_n^\star, d\rangle_{\mathcal{H}_K} = 0.$$

Both signs of t are feasible in $\mathcal{F}$, even when $f_n^\star$ lies on the boundary of the hull. Writing $g_f = \ell_{\lambda,f} - \ell_{\lambda,f_n^\star} = L_\lambda \widetilde{g}_f$, we obtain directly from (D.7)

$$P_n g_f = J_{[n]}(f_n^\star + d) - J_{[n]}(f_n^\star) \overset{②}{=} P_n d^2 + \lambda \|d\|^2_{\mathcal{H}_K}. \tag{D.10}$$

Step ② is the full-sample specialization of the subset quadratic identity.

Hence $0 \le P_n d^2 \le P_n g_f = L_\lambda P_n \widetilde{g}_f$, which converts prediction control to mean-excess control below. Finiteness on the hull and inclusion of $f_n^\star$ verify Assumption 1(1).

To control variance, factor the data-fit excess as $d(\mathbf{x}_i)(f(\mathbf{x}_i) + f_n^\star(\mathbf{x}_i) - 2y_i)$. The residual bound of Step 1 applies when $(\mathbf{K}_\lambda)_{ii} > 0$, and $d(\mathbf{x}_i) = 0$ otherwise. Hence, for every $i \in [n]$,

$$|d(\mathbf{x}_i)(f(\mathbf{x}_i) + f_n^\star(\mathbf{x}_i) - 2y_i)| \le 2\sqrt{L_\lambda}\,|d(\mathbf{x}_i)|\,.$$

Let $\widetilde{\ell}_f := \ell_{\lambda,f}/L_\lambda$ and $\widetilde{g}_f := \widetilde{\ell}_f - \widetilde{\ell}_{f_n^\star}$. The ridge penalty is constant across coordinates and disappears upon centering. Indeed, subtracting the full-sample mean leaves the pointwise identity on $[n]$

$$\widetilde{g}_f - P_n \widetilde{g}_f \overset{③}{=} \frac{d(f + f_n^\star - 2y) - P_n[d(f + f_n^\star - 2y)]}{L_\lambda}.$$

Here ③ cancels the coordinate-constant ridge term upon centering.

Squaring and averaging therefore gives $\mathrm{Var}\,[\widetilde{g}_f] = L_\lambda^{-2}\mathrm{Var}\,[d(f + f_n^\star - 2y)]$. Using $\mathrm{Var}\,[b] \le P_n b^2$ for the data-fit difference and then the pointwise bound gives

$$\mathrm{Var}\,[\widetilde{g}_f] \overset{④}{\le} \frac{1}{L_\lambda^2} P_n\left[d(f + f_n^\star - 2y)\right]^2 \overset{⑤}{\le} \frac{4}{L_\lambda} P_n d^2 \overset{⑥}{\le} 4 P_n \widetilde{g}_f.$$

Steps ④, ⑤, ⑥ use the second-moment bound, residual control, and (D.10), respectively.

As a result, the Bernstein inequality in Assumption 1(2) holds on $\mathcal{F}_\lambda^{\mathrm{sel}}$ with $B = 4$ and $\theta = 1$.

The oracle surrogate and envelope. The identity just proved gives $V_\star(\widetilde{g}_f) = 4P_n\widetilde{g}_f = 4[P_n(f - f_n^\star)^2 + \lambda\|f - f_n^\star\|^2_{\mathcal{H}_K}]/L_\lambda$. It is finite, nonnegative, and depends only on the normalized excess function on $[n]$. It vanishes at the oracle and satisfies $\mathrm{Var}\,[\widetilde{g}_f] \le V_\star(\widetilde{g}_f)$ for every hull member f. To bound $B_\infty(\mathcal{G}_\star)$, define $b : [n] \to \mathbb{R}$ by $b(i) = [(f(\mathbf{x}_i) - y_i)^2 - (f_n^\star(\mathbf{x}_i) - y_i)^2]/L_\lambda$. On a coordinate with positive projected diagonal, the two normalized squares lie in $[0, 1]$, so $b(i) \in [-1, 1]$. On every other coordinate their predictions agree, so $b(i) = 0$. The difference $\widetilde{g}_f(i) - b(i)$ is the ridge constant $\lambda(\|f\|^2_{\mathcal{H}_K} - \|f_n^\star\|^2_{\mathcal{H}_K})/L_\lambda$, independent of i. Therefore $|\widetilde{g}_f(i) - P_n\widetilde{g}_f| = |b(i) - P_n b| \le 2$, uniformly over the hull and coordinates. This proves $B_\infty(\mathcal{G}_\star) \le 2$.

Step 3: Weighted spectral head–tail bound. We use the spectral head–tail argument of Yang (2026, Lemma C.11 and proof of Theorem 5.2), with the projected spectrum and the regularized quadratic excess.

Coordinates adapted to the projected kernel. Let $q(d) \coloneqq P_n d^2 + \lambda \|d\|_{\mathcal{H}_K}^2$ be the quadratic form in (D.10). Because $g_f = L_\lambda \widetilde{g}_f$, this identity reads $q(f - f_n^\star) = L_\lambda P_n \widetilde{g}_f$. In this step we consider every $d \in \mathrm{range}(\Pi_\lambda)$, and $\mathbf{d} = (d(\mathbf{x}_i))_{i=1}^n$ is its evaluation vector. All coefficient sums in this step run over $j \in [n]$ with $\widehat{\lambda}_j > 0$. For each such eigenvalue of $\mathbf{K}_\lambda/n$, choose a Euclidean unit eigenvector $\mathbf{v}_j$, with these vectors mutually orthogonal, and put $\boldsymbol{\phi}_j = \sqrt{n}\,\mathbf{v}_j$. Define the RKHS element

$$e_j \coloneqq \frac{\sum_i (\mathbf{v}_j)_i \Pi_\lambda K(\cdot, \mathbf{x}_i)}{\sqrt{n\widehat{\lambda}_j}}.$$

The Gram identity and the eigenvalue equation give

$$\langle e_j, e_k \rangle_{\mathcal{H}_K} = \frac{\mathbf{v}_j^\top \mathbf{K}_\lambda \mathbf{v}_k}{n\sqrt{\widehat{\lambda}_j \widehat{\lambda}_k}} = \mathbb{1}_{\{j=k\}}, \qquad e_j(\mathbf{x}_i) = \sqrt{\widehat{\lambda}_j}(\boldsymbol{\phi}_j)_i.$$

There are $\mathrm{rank}(\mathbf{K}_\lambda) = \mathrm{dimrange}(\Pi_\lambda)$ such elements, so they form an RKHS orthonormal basis of that range, also for repeated positive eigenvalues. For $d \in \mathrm{range}(\Pi_\lambda)$, set $a_j = \sqrt{\widehat{\lambda}_j}\,\langle d, e_j \rangle_{\mathcal{H}_K}$. Then $\mathbf{d} = \sum_j a_j \boldsymbol{\phi}_j$, and the empirical and RKHS orthogonality relations give

$$P_n d^2 = \sum_j a_j^2, \qquad \|d\|_{\mathcal{H}_K}^2 = \sum_j \frac{a_j^2}{\widehat{\lambda}_j}, \qquad q(d) = \sum_j a_j^2 \left(1 + \frac{\lambda}{\widehat{\lambda}_j}\right).$$

All sums are over positive eigenvalues. The coefficients also satisfy $a_j = n^{-1} \sum_i d(\mathbf{x}_i)(\boldsymbol{\phi}_j)_i$ because $n^{-1}\boldsymbol{\phi}_j^\top \boldsymbol{\phi}_k = \mathbb{1}_{\{j=k\}}$.

We will use $q(d)$ to control the leading eigendirections and the RKHS norm to control the remaining directions. The latter constraint becomes more restrictive as $\widehat{\lambda}_j$ decreases, because its coefficient weight is $1/\widehat{\lambda}_j$. The Cauchy–Schwarz bounds below use these constraints on the coefficient sums.

Second moments of the random spectral coefficients. Fix an integer $1 \le p < n$. Let $J_1, \ldots, J_p$ be independent uniform indices in $[n]$, and let $\varepsilon_1, \ldots, \varepsilon_p$ be independent Rademacher variables, independent also of those indices. Write $J = (J_k)_{k=1}^p$ and $\varepsilon = (\varepsilon_k)_{k=1}^p$. For each positive eigendirection put $Z_j \coloneqq \frac{1}{p} \sum_{k=1}^p \varepsilon_k \left[\boldsymbol{\phi}_j\right]_{J_k}$.

The identities $\mathbb{E}_\varepsilon[\varepsilon_k \varepsilon_l] = \mathbb{1}_{\{k=l\}}$ and $n^{-1}\|\boldsymbol{\phi}_j\|_2^2 = 1$ give

$$\mathbb{E}_{J,\varepsilon}\left[Z_j^2\right] = \frac{1}{p^2} \sum_{k=1}^p \frac{1}{n} \sum_{i=1}^n (\boldsymbol{\phi}_j)_i^2 = \frac{1}{p}.$$

For notational convenience, set $Z_j = 0$ when $\widehat{\lambda}_j = 0$. Only positive eigendirections have positive weights below. Their individual second moments suffice, without independence between the Z_j.

For $f, f' \in \mathcal{F}_\lambda^{\mathrm{sel}}$, their difference $d = f - f'$ belongs to $\mathrm{range}(\Pi_\lambda)$ and satisfies $\|d\|_{\mathcal{H}_K} \le D_\lambda$ by the diameter bound proved in Step 1. We now consider all $d \in \mathrm{range}(\Pi_\lambda)$ satisfying $\|d\|_{\mathcal{H}_K} \le D_\lambda$ and $q(d) \le L_\lambda t$, where $t \ge 0$ is deterministic. This enlarged set is nonempty because it contains zero. It is closed and bounded in the finite-dimensional range of Π_λ, hence compact. Its linear prediction suprema are consequently finite and attained for every realization of J, ε.

If $t = 0$, then $q(d) = 0$ forces every sample evaluation to vanish. Injectivity on $\mathcal{F}$ gives $d = 0$, proving both prediction bounds below in this case. We now take $t > 0$ and recall $D_\lambda > 0$ from Step 1.

A bound for every cutoff. Fix an integer $Q \in \{0, \ldots, n\}$. In coefficient sums, the head consists of the positive eigendirections with $j \le Q$, and the tail consists of those with $j > Q$. Zero eigenvalues

are omitted from coefficient expansions and contribute zero to spectral sums. For each realization of the indices and Rademacher variables, the signed prediction average is

$$\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k d(\mathbf{x}_{J_k}) \overset{⑦}{=} \sum_{j:\widehat{\lambda}_j>0} a_j Z_j \overset{⑧}{=} \sum_{\substack{j\le Q\\ \widehat{\lambda}_j>0}} a_j Z_j + \sum_{\substack{j>Q\\ \widehat{\lambda}_j>0}} a_j Z_j.$$

Steps ⑦, ⑧ substitute the spectral expansion and separate its head and tail.

For the head, insert the weights supplied by $q(d)$ and apply Cauchy–Schwarz:

$$\left|\sum_{\substack{j\le Q\\ \widehat{\lambda}_j>0}} a_j Z_j\right| \overset{⑨}{\le} \left(\sum_{\substack{j\le Q\\ \widehat{\lambda}_j>0}} a_j^2\left(1+\frac{\lambda}{\widehat{\lambda}_j}\right)\right)^{1/2}\left(\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}Z_j^2\right)^{1/2}$$
$$\overset{①}{\le} \sqrt{L_\lambda t}\left(\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}Z_j^2\right)^{1/2}.$$

Steps ⑨, ① apply weighted Cauchy–Schwarz and $q(d)\le L_\lambda t$.

For the tail, use instead the weights of the squared RKHS norm:

$$\left|\sum_{\substack{j>Q\\ \widehat{\lambda}_j>0}} a_j Z_j\right| \overset{②}{\le} \left(\sum_{\substack{j>Q\\ \widehat{\lambda}_j>0}}\frac{a_j^2}{\widehat{\lambda}_j}\right)^{1/2}\left(\sum_{j>Q}\widehat{\lambda}_j Z_j^2\right)^{1/2} \overset{③}{\le} D_\lambda\left(\sum_{j>Q}\widehat{\lambda}_j Z_j^2\right)^{1/2}.$$

Steps ②, ③ apply weighted Cauchy–Schwarz and $\|d\|_{\mathcal{H}_K}\le D_\lambda$.

The triangle inequality combines the head and tail bounds. Their right-hand sides do not depend on the feasible coefficients a_j, so the same bounds hold after taking the supremum over d. We then average, using Cauchy–Schwarz and $\mathbb{E}\left[Z_j^2\right]=1/p$:

$$\mathbb{E}_{J,\varepsilon}\left[\left(\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}Z_j^2\right)^{1/2}\right] \overset{④}{\le} \sqrt{\frac{1}{p}\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}}, \quad \mathbb{E}_{J,\varepsilon}\left[\left(\sum_{j>Q}\widehat{\lambda}_j Z_j^2\right)^{1/2}\right] \overset{⑤}{\le} \sqrt{\frac{1}{p}\sum_{j>Q}\widehat{\lambda}_j}.$$

Steps ④, ⑤ use $\mathbb{E}\left[\sqrt{X}\right]\le\sqrt{\mathbb{E}\left[X\right]}$ and $\mathbb{E}\left[Z_j^2\right]=1/p$ in the head and tail, respectively.

It follows that

$$\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{d\in\text{range}(\Pi_\lambda),\ q(d)\le L_\lambda t\\ \|d\|_{\mathcal{H}_K}\le D_\lambda}}\left|\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k d(\mathbf{x}_{J_k})\right|\right] \overset{⑥}{\le} \sqrt{\frac{L_\lambda t}{p}\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}} + D_\lambda\sqrt{\frac{1}{p}\sum_{j>Q}\widehat{\lambda}_j}. \quad \text{(D.11)}$$

Here ⑥ combines the head and tail bounds before taking the supremum and expectation.

We now choose the cutoff so that each eigendirection uses the smaller of the curvature and diameter bounds:

$$Q = \left|\left\{j\in[n]:\widehat{\lambda}_j>0,\quad L_\lambda t\le D_\lambda^2(\widehat{\lambda}_j+\lambda)\right\}\right|. \quad \text{(D.12)}$$

Because $\widehat{\lambda}_j$ is nonincreasing in j, the positive indices satisfying $L_\lambda t\le D_\lambda^2(\widehat{\lambda}_j+\lambda)$ form an initial segment of the positive spectrum. As a result, their number Q indexes precisely that segment. Then $j\le Q$ selects the curvature term on every positive eigendirection satisfying the comparison, including equality, and $j>Q$ selects the diameter term on the remaining positive eigendirections. As a result,

$$L_\lambda t\sum_{j\le Q}\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda} + D_\lambda^2\sum_{j>Q}\widehat{\lambda}_j \overset{⑦}{=} \sum_{j=1}^{n}\min\left\{L_\lambda t\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}, D_\lambda^2\widehat{\lambda}_j\right\}.$$

Here ⑦ uses the chosen cutoff to select the smaller term in each positive eigendirection.

For nonnegative A, B, the inequality $(\sqrt{A}+\sqrt{B})^2 \le 2(A+B)$ follows from $2\sqrt{AB} \le A+B$. Applying it to (D.11) at the chosen Q in (D.12) proves

$$\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{d\in\mathrm{range}(\Pi_\lambda),\ q(d)\le L_\lambda t\\ \|d\|_{\mathcal{H}_K}\le D_\lambda}}\left|\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k d(\mathbf{x}_{J_k})\right|\right] \overset{⑧}{\le} \sqrt{\frac{2}{p}\sum_{j=1}^{n}\min\left\{L_\lambda t\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j+\lambda}, D_\lambda^2\widehat{\lambda}_j\right\}}. \tag{D.13}$$

Here ⑧ uses (D.11) at that cutoff and $(\sqrt{A}+\sqrt{B})^2 \le 2(A+B)$.

The cutoff is deterministic. The calculation also applies at zero ridge and, by the initial argument, at $t=0$.

Step 4: Localized TC and fixed points. Fix $p \in \{m,u\}$ and $r \ge 0$. We first contract the loss on the convex hull of subset solutions, where the residual bound holds, and then enlarge the linear prediction class to the constraint set of Step 3. This order matters because the larger set need not satisfy the residual bound used for loss contraction. Theorem A.3 and Lemma A.1 transfer the resulting Rademacher bound to the two TC quantities.

From oracle localization to prediction constraints. Recall $q(d) = P_n d^2 + \lambda\|d\|_{\mathcal{H}_K}^2$ from Step 3. For $f \in \mathcal{F}_\lambda^{\mathrm{sel}}$, equation (D.10) and the definition of the surrogate give $V_\star(\widetilde{g}_f) = 4P_n\widetilde{g}_f = \frac{4}{L_\lambda}q(f-f_n^\star)$. Since $L_\lambda > 0$, the localization condition is exactly $\widetilde{g}_f \in \mathcal{G}_\star(r) \quad \overset{⑨}{\Longleftrightarrow} \quad q(f-f_n^\star) \le L_\lambda r/4$. Here ⑨ uses the curvature identity and $L_\lambda > 0$ to rewrite the surrogate constraint. Every such difference lies in the range of Π_λ and has RKHS norm at most D_λ, since both predictors belong to the hull. We use the surrogate condition $V_\star \le r$, not merely the inequality $\mathrm{Var}[\widetilde{g}_f] \le r$. The latter would not by itself imply the displayed bound on $q(f-f_n^\star)$. The quadratic sublevel set in this compact hull is closed, hence compact, and its continuous excess-loss image is $\mathcal{G}_\star(r)$. The supremum can consequently be taken over the predictors in this deterministic sublevel set, and their continuous images give the same supremum over $\mathcal{G}_\star(r)$. The prediction bound (D.13) applies with $t = r/4$. At $r = 0$, injectivity of full-sample evaluation forces every admissible difference to vanish. Its normalized excess is then zero, so both TC quantities are zero. Henceforth let $r > 0$.

For this step, define the normalized data-fit excess by

$$b_f(i) := \frac{(f(\mathbf{x}_i)-y_i)^2 - (f_n^\star(\mathbf{x}_i)-y_i)^2}{L_\lambda}, \qquad i \in [n],$$

and let $\mathcal{B}_r := \{b_f : f \in \mathcal{F}_\lambda^{\mathrm{sel}},\ q(f-f_n^\star) \le L_\lambda r/4\}$. This class contains zero and is compact as the continuous image of the localized hull. Step 2 gives $b_f(i) \in [-1,1]$ on every coordinate, including the coordinates with zero projected diagonal, where $b_f(i) = 0$. As a result, $|b_f(i) - P_n b_f| \le 2$. We bound $\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r)$ with the independent uniform indices J_k and Rademacher variables ε_k of Step 3. As defined in Section A, its supremum contains no absolute value. The same RC bound will control both TC quantities.

Loss contraction on the prediction intervals of subset solutions. TC and centered-square TC are unchanged when a constant, including the ridge penalty, is subtracted separately from each class member. Indeed, if $h(i) = v(i) + c$ for all i, then $(P_{A_p} - P_n)h = (P_{A_p} - P_n)v$ and $h - P_n h = v - P_n v$. The scalar c may depend on the class member. For the present class this constant is $c_f = \lambda(\|f\|_{\mathcal{H}_K}^2 - \|f_n^\star\|_{\mathcal{H}_K}^2)/L_\lambda$, so $\widetilde{g}_f(i) = b_f(i) + c_f$ for every $i \in [n]$. In contrast to the TC cancellation, a fixed Rademacher realization gives

$$\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k\widetilde{g}_f(J_k) \overset{①}{=} \frac{1}{p}\sum_{k=1}^{p}\varepsilon_k b_f(J_k) + c_f\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k.$$

Here ① separates the member-dependent ridge constant from the Rademacher sum.

The last sum has mean zero, but c_f varies with the predictor over which the supremum is taken. Its zero mean therefore does not justify removing it inside that supremum. We first cancel the constant in TC and only then compare with the raw RC of the data-fit class $\mathcal{B}_r$. The generating predictors retain the localization constraint for the full regularized loss, expressed through $q(d) = P_n d^2 + \lambda\|d\|_{\mathcal{H}_K}^2$.

The ridge curvature therefore remains in the spectral bound. We write $d_i = f(\mathbf{x}_i) - f_n^\star(\mathbf{x}_i)$ and $\rho_i = f_n^\star(\mathbf{x}_i) - y_i$. When $(\mathbf{K}_\lambda)_{ii} > 0$, the coordinate map $z \mapsto [(\rho_i + z)^2 - \rho_i^2]/L_\lambda$ vanishes at zero. Since the hull is compact and convex and evaluation is linear, the differences at coordinate i form a compact interval. It contains zero because $f_n^\star$ belongs to the hull, and every z in it satisfies $|\rho_i + z| \le \sqrt{L_\lambda}$. For any two values z, z' in this interval, factorization gives

$$\frac{\left|(\rho_i + z)^2 - (\rho_i + z')^2\right|}{L_\lambda} \overset{②}{=} \frac{|z - z'|\,|(\rho_i + z) + (\rho_i + z')|}{L_\lambda} \overset{③}{\le} \frac{2}{\sqrt{L_\lambda}}\,|z - z'|\,.$$

Steps ②, ③ factor the difference of squares and use the residual bounds on the prediction interval.

Write the admissible interval as $[\alpha_i, \beta_i]$, with $\alpha_i \le 0 \le \beta_i$. Its metric projection $z \mapsto \min\{\beta_i, \max\{\alpha_i, z\}\}$ is 1-Lipschitz and fixes zero. Composing the loss-difference map with this projection therefore gives a globally $2/\sqrt{L_\lambda}$-Lipschitz extension that agrees with the original map on every admissible prediction.

When $(\mathbf{K}_\lambda)_{ii} = 0$, use the identically zero map, since every admissible d_i is zero. The coordinate maps therefore agree with the loss differences on the localized hull, where we apply contraction.

One Rademacher bound for the data-fit class. We apply Theorem A.2 to the localized prediction differences with the coordinate maps just constructed. Their common Lipschitz constant is $2/\sqrt{L_\lambda}$. The indexing predictors belong to the localized hull, so each difference $d = f - f_n^\star$ satisfies $d \in \mathrm{range}(\Pi_\lambda)$, $q(d) \le L_\lambda r/4$, and $\|d\|_{\mathcal{H}_K} \le D_\lambda$. After contraction, enlarge the linear prediction supremum to this constraint set and apply (D.13). The full chain is

$$\begin{aligned}
\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r) &\overset{④}{\le} \frac{2}{\sqrt{L_\lambda}}\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{f \in \mathcal{F}_\lambda^{\mathrm{sel}} \\ q(f - f_n^\star) \le L_\lambda r/4}} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (f - f_n^\star)(\mathbf{x}_{J_k})\right] \\
&\overset{⑤}{\le} \frac{2}{\sqrt{L_\lambda}}\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{d \in \mathrm{range}(\Pi_\lambda),\ q(d) \le L_\lambda r/4 \\ \|d\|_{\mathcal{H}_K} \le D_\lambda}} \left|\frac{1}{p}\sum_{k=1}^{p} \varepsilon_k d(\mathbf{x}_{J_k})\right|\right] \\
&\overset{⑥}{\le} 2\sqrt{2}\sqrt{\frac{1}{p}\sum_{j=1}^{n} \min\left\{r\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \frac{D_\lambda^2}{L_\lambda}\widehat{\lambda}_j\right\}}.
\end{aligned} \tag{D.14}$$

Inequality ④ is contraction with Lipschitz constant $2/\sqrt{L_\lambda}$, conditional on the sampled indices J and then averaged over them. In ⑤, every localized prediction difference satisfies the displayed constraints. The enlarged constraint set is invariant under $d \mapsto -d$, so its linear supremum equals its absolute supremum for each (J, ε). As a result, the absolute value enters only after loss contraction. For ⑥, we use (D.13) with $t = r/4$. Dividing its spectral summands by L_λ gives $\min\left\{(r/4)\widehat{\lambda}_j/(\widehat{\lambda}_j + \lambda), (D_\lambda^2/L_\lambda)\widehat{\lambda}_j\right\}$. Increasing $r/4$ to r gives the displayed bound.

TC of the localized excess-loss class. We retain the deterministic data-fit class $\mathcal{B}_r$ just defined. The ridge penalty is constant across coordinates, so its cancellation gives

$$\mathfrak{R}_p^{\eta}(\mathcal{G}_\star(r)) \overset{⑦}{=} \mathfrak{R}_p^{\eta}(\mathcal{B}_r), \qquad \mathfrak{R}_p^{+}(\mathcal{G}_\star(r)^{[2]}) \overset{⑧}{=} \mathfrak{R}_p^{+}(\mathcal{B}_r^{[2]}).$$

Steps ⑦, ⑧ cancel the coordinate-constant ridge term in the subset deviation and in the centered square.

The cancellation applies to TC and centered-square TC. We apply the RC comparison to the resulting class $\mathcal{B}_r$, since member-dependent constants need not cancel from raw RC.

The class $\mathcal{B}_r$ is nonempty, deterministic, and compact, so its suprema are finite. The independent uniform indices and independent Rademacher variables have the laws required by Theorem A.3. Applying that comparison gives

$$\max_{\eta \in \{+,-\}} \mathfrak{R}_p^{\eta}(\mathcal{B}_r) \overset{⑨}{\le} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r).$$

Here ⑨ applies Theorem A.3 to the deterministic class $\mathcal{B}_r$.

The inductive RC is the one defined above, without an absolute value inside its supremum. Both TC orientations are covered because $\mathfrak{R}_p^-(\mathcal{B}_r) = \mathfrak{R}_p^+(-\mathcal{B}_r)$ and $\mathfrak{R}_p^{(\mathrm{ind})}(-\mathcal{B}_r) = \mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r)$. The second equality follows by replacing ε by $-\varepsilon$, which has the same law. It requires no symmetry of $\mathcal{B}_r$. Combining this inequality with (D.14) therefore bounds both signs of TC for $\mathcal{G}_\star(r)$.

TC of the centered-square class. Recall $\mathcal{B}_r^{[2]} = \{(h - P_n h)^2 : h \in \mathcal{B}_r\}$, so centering precedes squaring. This class is deterministic, compact, and contains zero. Since $h(i) \in [-1, 1]$, its centered coordinates lie in $[-2, 2]$. For z, z' in this interval, $|z^2 - (z')^2| = |z - z'|\,|z + z'| \overset{①}{\le} 4\,|z - z'|$. Here ① uses $z, z' \in [-2, 2]$ to bound the sum factor. Composing the square map with projection onto $[-2, 2]$ therefore gives a globally 4-Lipschitz map that vanishes at zero and agrees with squaring on all these coordinates. Using the same independent (J, ε) as above, we obtain

$$\mathfrak{R}_p^+(\mathcal{B}_r^{[2]}) \overset{②}{\le} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r^{[2]}) = 2\mathbb{E}_{J,\varepsilon}\left[\sup_{h\in\mathcal{B}_r} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (h(J_k) - P_n h)^2\right]$$
$$\overset{③}{\le} 8\mathbb{E}_{J,\varepsilon}\left[\sup_{h\in\mathcal{B}_r} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (h(J_k) - P_n h)\right] \overset{④}{\le} 16\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{B}_r).$$

Steps ②, ③, ④ apply Theorem A.3, contraction with constant four, and Lemma A.1, respectively.

All suprema are finite by compactness. The contraction in ③ is Theorem A.2, and none of these inductive RC suprema contains an absolute value.

The common majorant. Both TC quantities are now bounded by the same data-fit RC, with multipliers two and 16. Substituting the spectral estimate (D.14) gives the common multiplier $\max\{2, 16\} \cdot 2\sqrt{2} = 32\sqrt{2} \le C_0 = 256\sqrt{2}$. The identities with the oracle classes proved above now yield

$$\max\left\{\max_{\eta\in\{+,-\}} \mathfrak{R}_p^\eta(\mathcal{G}_\star(r)), \mathfrak{R}_p^+\left(\mathcal{G}_\star(r)^{[2]}\right)\right\} \overset{⑤}{\le} C_0\sqrt{\frac{1}{p}\sum_j \min\left\{r\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \frac{D_\lambda^2}{L_\lambda}\widehat{\lambda}_j\right\}}.$$

Here ⑤ combines the two RC comparisons with (D.14) and $32\sqrt{2} \le C_0$.

We take $p = N_{u,m} \in \{m, u\}$, which satisfies $2 \le N_{u,m} \le n/2 < n$. The right side is then exactly $\psi_\lambda(r)$, proving (D.5), hence (4.1) for this oracle class. We choose $\psi_\star = \psi_\lambda$ for the generic ERM theorem. It remains to verify its sub-root property and bound its fixed point r_{exc}.

The sub-root property. Each spectral minimum in the formula for $\psi_\lambda(r)$ is a nonnegative, nondecreasing, continuous function of r, bounded above by $(D_\lambda^2/L_\lambda)\widehat{\lambda}_j$. The finite sum shows that ψ_λ is finite, nondecreasing, and continuous on $[0, \infty)$, with $\psi_\lambda(0) = 0$. For $r > 0$, its sub-root ratio is

$$\frac{\psi_\lambda(r)}{\sqrt{r}} \overset{⑥}{=} C_0\sqrt{\frac{1}{N_{u,m}}\sum_{j=1}^{n} \min\left\{\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \frac{D_\lambda^2\widehat{\lambda}_j}{L_\lambda r}\right\}}.$$

Here ⑥ divides each spectral minimum by the positive radius r.

Each term inside this square root is nonincreasing in r. Hence $\psi_\lambda(r)/\sqrt{r}$ is nonincreasing, which is the remaining sub-root condition.

Comparison of the two fixed points. Put $\phi = \psi_\lambda/C_0$. Since $L_\lambda, D_\lambda > 0$ and at least one projected eigenvalue is positive, both $\phi(t)$ and $\psi_\lambda(t)$ are positive for $t > 0$. They are finite sub-root functions. Lemma A.5 gives each a unique positive fixed point and identifies the radii where the function is at most the radius itself.

Substituting $r = L_\lambda t$ in the definition of $r_{K,\lambda}$ and dividing by $L_\lambda^2 > 0$ makes its defining inequality equivalent to $\phi(t)^2 \le t^2$, or $\phi(t) \le t$. Consequently the fixed point of ϕ is $t_\star = r_{K,\lambda}/L_\lambda$. Since $C_0 \ge 1$, sub-root scaling gives

$$\psi_\lambda(C_0^2 t_\star) = C_0\phi(C_0^2 t_\star) \overset{⑦}{\le} C_0^2\phi(t_\star) = C_0^2 t_\star, \qquad r_{\mathrm{exc}} \overset{⑧}{\le} C_0^2 t_\star = C_0^2 r_{K,\lambda}/L_\lambda.$$

Step ⑦ uses monotonicity of $\phi(t)/\sqrt{t}$. Step ⑧ uses the fixed-point characterization in Lemma A.5.

Comparison with the original kernel spectrum. It remains to justify the full-spectrum comparison stated in the main text. We hold $\lambda, D_\lambda, L_\lambda$ fixed. For every $t > 0$, the function of $s \geq 0$ given by $\min\{ts/(s+\lambda), D_\lambda^2 s\}$ is nondecreasing, with the stated zero-eigenvalue convention at $\lambda = 0$. Step 1 therefore orders the spectral sums at every radius. The same argument applies with D_λ^2/L_λ in the normalized TC formula. In the definition of $r_{K,\lambda}$, both spectral sums are compared with the same r^2. If the full-spectrum sum is at most r^2, the smaller projected sum is also at most r^2. As a result, every radius admissible for the full-spectrum inequality is admissible for the projected one. Taking infima proves that projection cannot enlarge $r_{K,\lambda}$.

To verify strictness, suppose $L_\lambda > 0$ and $\mathbf{K}_\lambda \neq \mathbf{K}$. Step 1 gives $D_\lambda > 0$ and $\operatorname{rank}(\mathbf{K}_\lambda) < \operatorname{rank}(\mathbf{K})$. There is therefore an index j with $\widehat{\lambda}_j = 0 < \lambda_j(\mathbf{K}/n)$. Its projected summand is zero, while its full-spectrum summand is positive at every positive radius. All other summands obey the weak comparison, so the comparison of sums is strict at every positive radius. Define r_{full} in this paragraph only as the fixed point obtained by replacing $\widehat{\lambda}_j$ by $\lambda_j(\mathbf{K}/n)$ in the definition of $r_{K,\lambda}$. Both radii are positive fixed points, since the corresponding square-root spectral functions are nonzero and sub-root. At r_{full} the full spectral sum, including its factor $L_\lambda/N_{u,m}$, equals r_{full}^2, whereas the projected sum is strictly smaller. The projected spectral sum is continuous. Its strict inequality with r^2 at $r = r_{\mathrm{full}} > 0$ therefore persists at some smaller positive radius. That radius is admissible in the definition of $r_{K,\lambda}$, proving $r_{K,\lambda} < r_{\mathrm{full}}$.

□

D.2 Ordinary excess risk and spectral rates

D.2.1 Deterministic risk and ridge comparisons

We use two deterministic comparisons. The first converts regularized excess to ordinary excess by accounting for the prediction change between the ridge and zero-ridge test oracles. The second compares the ridge learner directly with the zero-ridge training subset solution. Since it holds for every ridge value on each split, it also permits a ridge chosen from training labels. For $\lambda \geq 0$, define the deterministic test-oracle prediction bias by

$$b_{\lambda,u} := \max_{\substack{A\subseteq[n]\\|A|=u}} P_A\left(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0}\right)^2.$$

This finite maximum is fixed before the random split and satisfies $b_{0,u} = 0$.

Lemma D.2 (Ridge Comparisons For Ordinary Squared-Loss Excess)**.** In the setup of Lemma D.1, on every split and for all $f, h \in \mathcal{F}$,

$$\mathcal{E}_{\mathcal{U}}(f) = P_{\mathcal{U}}\left(f - \widehat{f}_{u,0}\right)^2, \qquad \left|\sqrt{\mathcal{E}_{\mathcal{U}}(f)} - \sqrt{\mathcal{E}_{\mathcal{U}}(h)}\right| \leq \sqrt{P_{\mathcal{U}}(f-h)^2}. \tag{D.15}$$

Consequently, on every split and for every $\lambda \geq 0$,

$$\sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda})} \leq \sqrt{\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda})} + \sqrt{b_{\lambda,u}}, \qquad b_{\lambda,u} \leq \frac{\lambda R_0^2}{4}. \tag{D.16}$$

If $\mathbf{K} \neq \mathbf{0}$, then, simultaneously for all $\lambda \geq 0$,

$$\sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda})} \leq \sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,0})} + \sqrt{K_{\max}} R_0 \frac{\lambda}{K_{\min}^{+} + \lambda}. \tag{D.17}$$

If $L_0 > 0$, there exists $\lambda_1 > 0$, determined by the fixed full sample, such that $L_\lambda \geq L_0/2$ for $0 < \lambda \leq \lambda_1$. Consequently

$$\lambda R_0^2 \leq \frac{4L_\lambda}{N_{u,m}} \quad \text{whenever} \quad 0 < \lambda \leq \min\left\{\lambda_1, \frac{2L_0}{N_{u,m}R_0^2}\right\}. \tag{D.18}$$

If $\mathbf{K} = \mathbf{0}$, every subset solution is zero and its ordinary excess is zero for every $\lambda \geq 0$.

Proof of Lemma D.2. All comparisons in this proof are deterministic on the fixed full sample. Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$ and that a relevant subset has size m, u, or n. The quantity R_0 is the largest RKHS norm of the zero-ridge subset solutions over all relevant subsets. The test-oracle prediction change is measured by $b_{\lambda,u} = \max_{\substack{A\subseteq[n]\\|A|=u}} P_A(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0})^2$. The ordinary excess compares with $\widehat{f}_{u,0}$, whereas $\mathcal{E}_{\mathcal{U},\lambda}$ compares the two regularized objectives.

Proof outline.

(1) Ordinary excess as prediction distance. Specialize the subset quadratic identity to the unregularized test problem, then prove (D.15) by Cauchy–Schwarz.
(2) Spectral differences of subset solutions. Subtract the representer formulas, derive the exact prediction and RKHS norm identities, and treat singular subset matrices.
(3) Comparison with regularized excess. Bound the test-oracle bias and apply (D.15) with the ridge test subset solution as the comparison predictor.
(4) Comparison with the zero-ridge learner. Control the training prediction change on the test coordinates and apply (D.15) with the zero-ridge training subset solution.
(5) Existence of positive admissible ridge. Use spectral shrinkage to keep a prediction difference nonzero and a residual of magnitude at least $\sqrt{L_0/2}$ at a coordinate attaining L_0.

Step 1: Ordinary excess as prediction distance. The test normal equation converts ordinary excess into a squared distance. This makes perturbations of the learner accessible through the triangle inequality. Fix a split and $f \in \mathcal{F}$, and put $d = f - \widehat{f}_{u,0}$. Apply the subset quadratic identity (D.7) with $A = \mathcal{U}$ and $\lambda = 0$. Here $J_{\mathcal{U}}(h) = P_{\mathcal{U}}(h - y)^2$, and $\widehat{f}_{u,0}$ attains its minimum over the full linear space. Hence

$$\mathcal{E}_{\mathcal{U}}(f) = J_{\mathcal{U}}(\widehat{f}_{u,0} + d) - J_{\mathcal{U}}(\widehat{f}_{u,0}) \overset{①}{=} P_{\mathcal{U}} d^2 \geq 0.$$

Step ① is (D.7) at zero ridge. The normal equation cancels $2P_{\mathcal{U}}[(\widehat{f}_{u,0} - y)d]$, without requiring interpolation or an invertible Gram matrix. This is the fixed-sample least-squares excess identity of Hsu et al. (2012, Proposition 21), including singular test matrices.

Any other test minimizer $\widetilde{f}_u$ satisfies $P_{\mathcal{U}}(\widetilde{f}_u - \widehat{f}_{u,0})^2 = \mathcal{E}_{\mathcal{U}}(\widetilde{f}_u) = 0$. Hence all test minimizers have the same test predictions, although their RKHS functions may differ.

For $f, h \in \mathcal{F}$, the reverse triangle inequality for the empirical L^2 seminorm gives

$$\left|\sqrt{\mathcal{E}_{\mathcal{U}}(f)} - \sqrt{\mathcal{E}_{\mathcal{U}}(h)}\right| \overset{②}{\leq} \sqrt{P_{\mathcal{U}}(f - h)^2}.$$

Step ② uses the ordinary-excess identity just proved. This proves (D.15) on every split, also for predictors selected using that split.

Step 2: Spectral differences of subset solutions. Using one eigenbasis for both ridge regimes isolates their difference as scalar shrinkage in each positive eigendirection. Fix a relevant subset A, write $p = |A|$ and $\mathbf{y}_A = (y_i)_{i\in A}$, and recall the evaluation operator $T_A f = (f(\mathbf{x}_i))_{i\in A}$ from Lemma D.1. It satisfies $T_A T_A^* = \mathbf{K}_A$. Choose Euclidean orthonormal eigenvectors $\mathbf{v}_{A,j}$ of $\mathbf{K}_A/p$, indexed by $j \in \{1, \ldots, p\}$, with eigenvalues $\rho_{A,j} \geq 0$, and write $c_{A,j} = \mathbf{v}_{A,j}^\top \mathbf{y}_A$. All sums in this step are over the indices with $\rho_{A,j} > 0$. For such an index put $e_{A,j} = \frac{T_A^* \mathbf{v}_{A,j}}{\sqrt{p\rho_{A,j}}}$.

The Gram identities give $\langle e_{A,j}, e_{A,k}\rangle_{\mathcal{H}_K} = \mathbb{1}_{\{j=k\}}$ and $T_A e_{A,j} = \sqrt{p\rho_{A,j}}\mathbf{v}_{A,j}$, by the same calculation as in Step 1 of Lemma D.1.

In this orthonormal RKHS family, the representer expansions are

$$\widehat{f}_{A,0} \overset{③}{=} \sum_{j:\rho_{A,j}>0} \frac{c_{A,j}}{\sqrt{p\rho_{A,j}}} e_{A,j}, \qquad \widehat{f}_{A,\lambda} \overset{④}{=} \sum_{j:\rho_{A,j}>0} \frac{\rho_{A,j}}{\rho_{A,j} + \lambda} \frac{c_{A,j}}{\sqrt{p\rho_{A,j}}} e_{A,j}.$$

Here ③ expands the minimum-norm representer formula in positive eigendirections. ④ expands the ridge formula in the same eigendirections.

The pseudoinverse and ridge inverse multiply positive spectral coefficients by $(p\rho_{A,j})^{-1}$ and $[p(\rho_{A,j} + \lambda)]^{-1}$, respectively. Applying $T_A^* \mathbf{v}_{A,j} = \sqrt{p\rho_{A,j}} e_{A,j}$ gives the displayed shrinkage factors (Hsu

et al., 2012, Proposition 22). Zero eigendirections contribute nothing because $\|T_A^* \mathbf{v}_{A,j}\|_{\mathcal{H}_K}^2 = \mathbf{v}_{A,j}^\top \mathbf{K}_A \mathbf{v}_{A,j} = 0$.

Subtracting these expansions multiplies the zero-ridge coefficient by $-\lambda/(\rho_{A,j}+\lambda)$. RKHS orthonormality and the evaluation formula therefore give

$$\left\|\widehat{f}_{A,\lambda} - \widehat{f}_{A,0}\right\|_{\mathcal{H}_K}^2 \overset{⑤}{=} \frac{1}{p} \sum_{j:\rho_{A,j}>0} \frac{\lambda^2 c_{A,j}^2}{\rho_{A,j}(\rho_{A,j}+\lambda)^2},$$
$$P_A(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0})^2 \overset{⑥}{=} \frac{1}{p} \sum_{j:\rho_{A,j}>0} \frac{\lambda^2 c_{A,j}^2}{(\rho_{A,j}+\lambda)^2}. \tag{D.19}$$

Step ⑤ sums the squared RKHS coefficients. Step ⑥ uses $T_A e_{A,j} = \sqrt{p\rho_{A,j}}\mathbf{v}_{A,j}$ and divides the squared Euclidean norm by p.

At $\lambda = 0$ both sums vanish. If $\mathbf{K}_A = \mathbf{0}$, the sums are empty and both subset solutions are zero. The formulas are independent of the eigenbasis because each multiplier is constant on an eigenspace and $\sum_{j:\rho_{A,j}=\rho} c_{A,j}^2$ is the squared norm of the projection of $\mathbf{y}_A$ onto that eigenspace.

Step 3: Comparison with regularized excess. To pass from regularized to ordinary excess, we separate the learner–ridge-oracle distance from the displacement of the test oracle. The test-oracle bias is obtained from (D.19):

$$b_{\lambda,u} \overset{⑦}{=} \max_{\substack{A\subseteq[n]\\|A|=u}} \frac{1}{u} \sum_{j:\rho_{A,j}>0} \frac{\lambda^2 c_{A,j}^2}{(\rho_{A,j}+\lambda)^2}.$$

Here ⑦ maximizes the prediction identity in (D.19) over all test-sized subsets.

This formula is the empirical counterpart of the squared prediction bias in Hsu et al. (2012, Proposition 23). We use the identity just proved, so no random-design or nonsingularity assumption from that setting is required.

The scalar inequality $4\rho\lambda \le (\rho+\lambda)^2$ gives $\lambda^2/(\rho+\lambda)^2 \le \lambda/(4\rho)$ for $\rho > 0$.

The inequality also holds at zero ridge. Applying it to (D.19) gives

$$P_A(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0})^2 \overset{⑧}{\le} \frac{\lambda}{4p} \sum_{j:\rho_{A,j}>0} \frac{c_{A,j}^2}{\rho_{A,j}} \overset{⑨}{=} \frac{\lambda}{4}\left\|\widehat{f}_{A,0}\right\|_{\mathcal{H}_K}^2 \overset{①}{\le} \frac{\lambda R_0^2}{4}.$$

Here ⑧ applies the scalar bias bound in each positive eigendirection. ⑨ uses the minimum-norm spectral identity. ① uses the definition of R_0.

Taking the maximum over the subsets of size u proves the bound on $b_{\lambda,u}$. Apply (D.15) with $f = \widehat{f}_{m,\lambda}$ and $h = \widehat{f}_{u,\lambda}$. The same equation identifies $\mathcal{E}_{\mathcal{U}}(h)$ with $P_{\mathcal{U}}(\widehat{f}_{u,\lambda} - \widehat{f}_{u,0})^2$, giving

$$\sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda})} \overset{②}{\le} \sqrt{P_{\mathcal{U}}(\widehat{f}_{m,\lambda} - \widehat{f}_{u,\lambda})^2} + \sqrt{P_{\mathcal{U}}(\widehat{f}_{u,\lambda} - \widehat{f}_{u,0})^2} \overset{③}{\le} \sqrt{\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda})} + \sqrt{b_{\lambda,u}}.$$

Step ② uses (D.15). For ③, subset curvature identifies regularized excess as $P_{\mathcal{U}}d^2 + \lambda\|d\|_{\mathcal{H}_K}^2 \ge P_{\mathcal{U}}d^2$, where $d = \widehat{f}_{m,\lambda} - \widehat{f}_{u,\lambda}$, and the other squared distance is at most $b_{\lambda,u}$. This proves (D.16).

Step 4: Comparison with the zero-ridge learner. For simultaneous ridge control, it is more useful to compare each ridge learner directly with the single zero-ridge learner. Suppose $\mathbf{K} \neq \mathbf{0}$. Recall $K_{\max} = \max_{i\in[n]} K(\mathbf{x}_i, \mathbf{x}_i)$ and that $K_{\min}^+$ is the smallest positive eigenvalue of $\mathbf{K}_A/|A|$ over all relevant subsets A. This finite collection is nonempty because it includes the positive spectrum of $\mathbf{K}/n$, so $K_{\min}^+ > 0$. Since every positive $\rho_{A,j}$ is at least $K_{\min}^+$, the first identity in (D.19) implies

$$\left\|\widehat{f}_{A,\lambda} - \widehat{f}_{A,0}\right\|_{\mathcal{H}_K}^2 \overset{④}{\le} \left(\frac{\lambda}{K_{\min}^+ + \lambda}\right)^2 \frac{1}{p} \sum_{j:\rho_{A,j}>0} \frac{c_{A,j}^2}{\rho_{A,j}} \overset{⑤}{\le} R_0^2 \left(\frac{\lambda}{K_{\min}^+ + \lambda}\right)^2.$$

Here ④ uses $\rho_{A,j} \ge K_{\min}^+$ in the exact spectral difference formula. ⑤ identifies the remaining sum with the squared norm of a subset solution and bounds it by R_0^2.

For every $i \in [n]$, the reproducing property and Cauchy–Schwarz give

$$\left|(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0})(\mathbf{x}_i)\right| \overset{⑥}{\leq} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)}\left\|\widehat{f}_{A,\lambda} - \widehat{f}_{A,0}\right\|_{\mathcal{H}_K} \overset{⑦}{\leq} \sqrt{K_{\max}} R_0 \frac{\lambda}{K^+_{\min} + \lambda}.$$

Here ⑥ uses the reproducing property and Cauchy–Schwarz. ⑦ uses the preceding norm bound and $K(\mathbf{x}_i, \mathbf{x}_i) \leq K_{\max}$.

For the training subset solution this implies

$$P_{\mathcal{U}}(\widehat{f}_{m,\lambda} - \widehat{f}_{m,0})^2 \overset{⑧}{\leq} K_{\max} R_0^2 \left(\frac{\lambda}{K^+_{\min} + \lambda}\right)^2.$$

Here ⑧ squares and averages the uniform coordinate bound over $\mathcal{U}$.

Apply (D.15) with $f = \widehat{f}_{m,\lambda}$ and $h = \widehat{f}_{m,0}$, and insert this bound. The result is (D.17). Every calculation holds for every $\lambda \geq 0$ on every split, which proves the asserted simultaneity. If $\mathbf{K} = \mathbf{0}$, the full span is $\{0\}$, proving the separate zero-kernel assertion.

Step 5: Existence of positive admissible ridge. A positive ridge is admissible only if the residual maximum does not collapse immediately. We retain both a differing coordinate and a nonzero residual there. Assume $L_0 > 0$. Lemma D.1 gives $D_0 > 0$, so $R_0 > 0$ and $\mathbf{K} \neq \mathbf{0}$. The finite maximum defining L_0 is attained at some relevant subset A_* and coordinate i_* with $(\mathbf{K}_0)_{i_* i_*} > 0$. There are relevant subsets B_*, C_* whose zero-ridge predictions differ at i_*, by the zero-diagonal characterization in that lemma. Fix these subsets and this coordinate using the full sample, before considering the random split. We give an explicit positive neighborhood. Put

$$\delta_* = \left|\widehat{f}_{B_*,0}(\mathbf{x}_{i_*}) - \widehat{f}_{C_*,0}(\mathbf{x}_{i_*})\right| > 0,$$

$$\varepsilon_* = \min\left\{\frac{\delta_*}{4}, (1 - 2^{-1/2})\sqrt{L_0}\right\} > 0, \qquad \lambda_1 = \frac{K^+_{\min}\varepsilon_*}{\sqrt{K_{\max}} R_0} > 0.$$

The denominator is positive because $R_0 > 0$ and $\mathbf{K} \neq \mathbf{0}$. For $0 < \lambda \leq \lambda_1$, Step 4 gives, for every relevant A and every $i \in [n]$,

$$\left|(\widehat{f}_{A,\lambda} - \widehat{f}_{A,0})(\mathbf{x}_i)\right| \overset{⑨}{\leq} \sqrt{K_{\max}} R_0 \frac{\lambda}{K^+_{\min} + \lambda} \overset{①}{\leq} \frac{\sqrt{K_{\max}} R_0 \lambda}{K^+_{\min}} \overset{②}{\leq} \varepsilon_*.$$

Here ⑨ uses Step 4's uniform evaluation bound. ① decreases the positive denominator. ② uses $\lambda \leq \lambda_1$.

Applying the reverse triangle inequality to the predictions of the two subset solutions, and then to the residual attaining L_0, gives

$$\left|\widehat{f}_{B_*,\lambda}(\mathbf{x}_{i_*}) - \widehat{f}_{C_*,\lambda}(\mathbf{x}_{i_*})\right| \overset{③}{\geq} \delta_* - 2\varepsilon_* \overset{④}{\geq} \delta_*/2 > 0,$$

$$\left|\widehat{f}_{A_*,\lambda}(\mathbf{x}_{i_*}) - y_{i_*}\right| \overset{⑤}{\geq} \sqrt{L_0} - \varepsilon_* \overset{⑥}{\geq} \sqrt{L_0/2}.$$

Steps ③ and ⑤ use the reverse triangle inequality. Steps ④ and ⑥ use the two defining bounds on ε_*.

The first line keeps coordinate i_* in the defining maximum for L_λ, since the zero-diagonal characterization gives $(\mathbf{K}_\lambda)_{i_* i_*} > 0$. The second line keeps its squared residual at least $L_0/2$. Together they prove $L_\lambda \geq L_0/2$. This argument controls evaluations of the subset solutions directly and does not require continuity of the projection or constancy of its rank. If also $\lambda \leq 2L_0/(N_{u,m} R_0^2)$, multiplying by $R_0^2 > 0$ and using $L_\lambda \geq L_0/2$ gives

$$\lambda R_0^2 \overset{⑦}{\leq} \frac{2L_0}{N_{u,m}} \overset{⑧}{\leq} \frac{4L_\lambda}{N_{u,m}}.$$

Here ⑦ uses the additional upper bound on λ. ⑧ uses $L_\lambda \geq L_0/2$.

This is precisely the ridge restriction (D.20). All denominators are positive by the nondegeneracy already checked. This proves (D.18) and completes the proof. □

D.2.2 THE FIXED-RIDGE EXCESS-RISK THEOREM

Theorem D.3 (Ordinary Excess Risk For Unbounded Transductive Kernel Learning)**.** In the full-span TKL setup above, let $x \geq 1$. There is a universal constant $C > 0$ such that the following assertions hold.

(1) Fixed ridge. For each fixed $\lambda \geq 0$ satisfying

$$\lambda R_0^2 \leq \frac{4L_\lambda}{N_{u,m}}, \tag{D.20}$$

there is an event $\Omega_\lambda(x)$ of probability at least $1 - \exp(-x)$ over the random split on which

$$0 \leq \mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \leq C\left(r_{K,\lambda} + \frac{L_\lambda x}{N_{u,m}}\right). \tag{D.21}$$

(2) A simultaneous ridge range. If $L_0 > 0$, then on the event $\Omega_0(x)$ from part (1),

$$0 \leq \mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \leq C\left(r_{K,0} + \frac{L_0 x}{N_{u,m}}\right) \tag{D.22}$$

simultaneously for every

$$0 \leq \lambda \leq K_{\min}^{+}\sqrt{\frac{L_0}{N_{u,m}K_{\max}R_0^2}}. \tag{D.23}$$

The upper endpoint $K_{\min}^{+}\sqrt{L_0/(N_{u,m}K_{\max}R_0^2)}$ of (D.23) is finite and strictly positive. Indeed, $L_0 > 0$ implies $R_0 > 0$, $K_{\max} > 0$, and $K_{\min}^{+} > 0$, as verified in Step 3 of the proof. If $L_0 = 0$, (D.22) holds at $\lambda = 0$ on $\Omega_0(x)$.

Proof of Theorem D.3. Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. The target is the ordinary excess $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda})$, whose comparator is $\widehat{f}_{u,0}$. We first prove an intermediate bound for $\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) = J_{\mathcal{U}}(\widehat{f}_{\mathcal{M},\lambda}) - J_{\mathcal{U}}(\widehat{f}_{\mathcal{U},\lambda})$, where $J_A(f) = P_A(f-y)^2 + \lambda\|f\|_{\mathcal{H}_K}^2$. The class $\mathcal{F}_\lambda^{\mathrm{sel}}$ contains all relevant subset solutions, D_λ is its RKHS diameter, and L_λ is the residual maximum on coordinates where those solutions can differ.

We distinguish the two loss scales used below. The radius $r_{K,\lambda}$ is in original loss units. For generic ERM, when $L_\lambda > 0$, we use $\widetilde{\ell}_f = \ell_{\lambda,f}/L_\lambda$ and the oracle $f_n^\star = \widehat{f}_{[n],\lambda}$. The normalized excess is $\widetilde{g}_f = \widetilde{\ell}_f - \widetilde{\ell}_{f_n^\star}$, with surrogate $V_\star(\widetilde{g}_f) = 4P_n\widetilde{g}_f$. Its majorant $\psi_\star = \psi_\lambda$ from (D.4) has fixed point r_{exc}. We use $c_1 = 64(4+c_0)^2$ and $C_0 = 256\sqrt{2}$.

Throughout this proof the full sample is fixed. A split is specified by a uniformly chosen m-element subset $\mathcal{M} \subseteq [n]$, with $\mathcal{U} = [n] \setminus \mathcal{M}$. As a result, the split space is finite.

Proof outline.

(1) Localize and normalize the losses. Restrict the analysis to the deterministic hull, identify the nonnegative excess risk, handle zero residual maximum, and normalize the entire regularized loss.
(2) Apply the generic ERM bound. Insert $B = 4$, the bound on r_{exc}, and $B_\infty(\mathcal{G}_\star) \leq 2$ into the two-event ERM theorem. Track absorption and normalization explicitly.
(3) Ordinary excess and positive ridge. Apply the deterministic comparison lemma, absorb the test-oracle bias under (D.20), and derive the zero-ridge spectral rate for the specified positive-ridge range on the zero-ridge event.

Step 1: Localize and normalize the losses. The generic theorem needs a fixed class containing its minimizers. The convex hull of subset solutions supplies this class before normalization. Lemma D.1 shows that the training, test, and full-sample subset solutions all belong to $\mathcal{F}_\lambda^{\mathrm{sel}}$. Each remains a minimizer after restriction, including the minimum-norm selection at zero ridge. The hull is fixed by the full sample. Applying the subset quadratic identity from Step 1 of that lemma with $A = \mathcal{U}$ and $d = \widehat{f}_{m,\lambda} - \widehat{f}_{u,\lambda}$ gives

$$\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) \overset{①}{=} P_{\mathcal{U}}d^2 + \lambda\|d\|_{\mathcal{H}_K}^2 \geq 0.$$

Here ① applies the subset curvature identity with $A = \mathcal{U}$.

If $L_\lambda = 0$, all hull generators coincide and $D_\lambda = r_{K,\lambda} = \mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) = 0$. Take $\Omega_\lambda(x)$ to be the whole split space in this case. For $L_\lambda > 0$, normalize by $\widetilde{\ell}_f = \ell_{\lambda,f}/L_\lambda$. Since the hull contains each unrestricted subset minimizer, for every relevant A we have

$$P_A\widetilde{\ell}_f = \frac{J_A(f)}{L_\lambda}, \qquad \min_{f\in\mathcal{F}_\lambda^{\mathrm{sel}}} P_A\widetilde{\ell}_f \overset{②}{=} \frac{1}{L_\lambda}\min_{f\in\mathcal{F}} J_A(f) \overset{③}{=} \frac{J_A(\widehat{f}_{A,\lambda})}{L_\lambda}.$$

Here ② uses positive normalization and preservation of the unrestricted minimum. ③ uses attainment at the subset solution.

This verifies the minimizer requirements at $A = \mathcal{M}, \mathcal{U}, [n]$. The generic excess is the regularized excess divided by L_λ:

$$P_{\mathcal{U}}\left(\widetilde{g}_{\widehat{f}_{m,\lambda}} - \widetilde{g}_{\widehat{f}_{u,\lambda}}\right) \overset{④}{=} P_{\mathcal{U}}\left(\widetilde{\ell}_{\widehat{f}_{m,\lambda}} - \widetilde{\ell}_{\widehat{f}_{u,\lambda}}\right) \overset{⑤}{=} \frac{J_{\mathcal{U}}(\widehat{f}_{m,\lambda}) - J_{\mathcal{U}}(\widehat{f}_{u,\lambda})}{L_\lambda}.$$

Here ④ cancels the reference loss. ⑤ undoes the normalization by L_λ.

Step 2: Apply the generic ERM bound. All structural inputs are now deterministic, so the parameterized ERM bound (C.1) supplies the split event. We insert the deterministic conclusions of Lemma D.1 into that bound, in the same order as the spectral-majorant substitution in Yang (2026, proof of Theorem 5.2).

Verification of the inputs. Step 1 verifies that restriction and normalization preserve the training, test, and full-sample subset solutions. For $r \geq 0$, the class $\mathcal{G}_\star(r)$ consists of the excesses $\widetilde{g}_f$ with $f \in \mathcal{F}_\lambda^{\mathrm{sel}}$ and $4P_n\widetilde{g}_f \leq r$. On this deterministic hull, Lemma D.1 gives

$$B = 4, \qquad \theta = 1, \qquad B_\infty(\mathcal{G}_\star) \overset{⑥}{\leq} 2, \qquad r_{\mathrm{exc}} \overset{⑦}{\leq} C_0^2\frac{r_{K,\lambda}}{L_\lambda}.$$

Here ⑥ uses the envelope conclusion of Lemma D.1. ⑦ uses that lemma's fixed-point comparison.

The same lemma proves (D.5), namely (4.1) with $\mathcal{H} = \mathcal{G}_\star$, $V = V_\star$, and $\psi = \psi_\lambda$, using Theorem A.3, loss contraction, and Lemma A.1. Its constant C_0 includes their factors. This verifies all inputs of Theorem 4.2. Choose $a = 8$ and $x_m = x_u = s = x + \log 2$. Since $x \geq 1$, these parameters are at least one. The extra condition for $\theta = 1$ is satisfied because $\frac{B\max\{u,m\}}{n} = \frac{4\max\{u,m\}}{n} < 4 < 8 = a$. All these choices are deterministic for the fixed full sample and fixed λ. The parameterized bound (C.1), proved in Section C.2, therefore holds on an event of probability at least $1-2\exp(-s) = 1-\exp(-x)$ over the random split. The remaining substitutions are deterministic on this event.

The scalar absorption and ERM substitution. Recall that $R_\star(8, s)$ is the UTLC remainder for the normalized oracle class. Here $B = 4$, $\theta = 1$, and $a = 8 = 2B$, so (C.5) applies with denominator $1 - Bt/(an) = 1 - t/(2n) > 1/2$ for $t \in \{m, u\}$. It gives $Q_t(8, s) \leq (2t/n)R_\star(8, s)$, and summing yields $Q_u(8, s) + Q_m(8, s) \leq 2R_\star(8, s)$. By Step 1 the generic normalized excess risk is the regularized excess risk divided by L_λ. Substituting $B/a = 1/2$ and $\theta = 1$ into (C.1) consequently gives, on the event just obtained,

$$\frac{\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda})}{L_\lambda} \overset{⑧}{\leq} 2R_\star(8, s) + \frac{1}{2}\left[Q_u(8, s) + Q_m(8, s)\right] \overset{⑨}{\leq} 3R_\star(8, s).$$

Here ⑧ applies (C.1) with $B/a = 1/2$. ⑨ uses the sum of the two linear absorption bounds.

Insertion of the spectral fixed point and restoration of units. The definition of $R_\star(8, s)$ and the candidate $B_\infty(\mathcal{G}_\star)s/N_{u,m}$ in its minimum give

$$\begin{aligned} R_\star(8, s) &= 8c_1\left(r_{\mathrm{exc}} + \frac{s}{N_{u,m}}\right) + c_1\min\left\{8(n-1)\frac{s^2}{N_{u,m}^2}, B_\infty(\mathcal{G}_\star)\frac{s}{N_{u,m}}\right\} \\ &\overset{①}{\leq} c_1\left(8r_{\mathrm{exc}} + 10\frac{s}{N_{u,m}}\right) \overset{②}{\leq} c_1\left(8C_0^2\frac{r_{K,\lambda}}{L_\lambda} + 10\frac{s}{N_{u,m}}\right). \end{aligned}$$

Here ① selects the envelope branch and uses $B_\infty(\mathcal{G}_\star) \leq 2$. ② inserts the spectral fixed-point comparison.

Multiplying the normalized excess bound by L_λ yields

$$\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) \overset{③}{\le} 24c_1C_0^2 r_{K,\lambda} + 30c_1\frac{L_\lambda s}{N_{u,m}}.$$

Here ③ multiplies the normalized excess bound by L_λ and inserts the preceding remainder estimate.

For $x \ge 1$, $s = x + \log 2 \le (1+\log 2)x$. Set $C_{\mathrm{reg}} = (80/3)c_1C_0^2$. The two coefficients satisfy $24c_1C_0^2 \le C_{\mathrm{reg}}, \qquad 30c_1(1+\log 2) \le 60c_1 \le C_{\mathrm{reg}}$. The last inequality uses $C_0^2 = 131072$. For definiteness, define the following event on the finite split space:

$$\Omega_\lambda(x) := \left\{ \mathcal{M} \subseteq [n] : |\mathcal{M}| = m, \quad \mathcal{E}_{[n]\setminus\mathcal{M},\lambda}(\widehat{f}_{\mathcal{M},\lambda}) \le C_{\mathrm{reg}}\left(r_{K,\lambda} + \frac{L_\lambda x}{N_{u,m}}\right)\right\}.$$

The preceding calculation shows that this event contains the event supplied by (C.1), so its probability is at least $1-\exp(-x)$ over the random split. Measurability follows from finiteness of the split space. By construction,

$$\mathcal{E}_{\mathcal{U},\lambda}(\widehat{f}_{m,\lambda}) \overset{④}{\le} C_{\mathrm{reg}}\left(r_{K,\lambda} + \frac{L_\lambda x}{N_{u,m}}\right) \quad \text{on } \Omega_\lambda(x). \tag{D.24}$$

Here ④ uses the defining bound of $\Omega_\lambda(x)$, whose probability was established above.

When $L_\lambda = 0$, Step 1 shows that both sides of this bound are zero on every split. The displayed definition then agrees with the whole-space event chosen in that step. For each fixed λ and x, the event depends on no subsequently chosen predictor or tolerance. At $\lambda = 0$, (D.24) is an ordinary excess bound for the minimum-norm learner on $\Omega_0(x)$.

Step 3: Ordinary excess and positive ridge. The regularized bound must be converted to the ordinary target. We absorb either the test-oracle bias or the prediction change from the zero-ridge learner. Recall the test-oracle prediction bias $b_{\lambda,u}$ defined before Lemma D.2, and let R_0 be the largest RKHS norm of the zero-ridge subset solutions on subsets of sizes m, u, n. For this proof only write $F_\lambda(x) = r_{K,\lambda} + L_\lambda x/N_{u,m}$. It is a nonnegative deterministic number and $F_\lambda(x) \ge L_\lambda/N_{u,m}$ because $x \ge 1$. Choose $C = 4C_{\mathrm{reg}}$, a universal number at least four. For any fixed $\lambda \ge 0$, Lemma D.2 and (D.24) give the unrestricted-ridge bound

$$\begin{aligned}\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) &\overset{⑤}{\le} \left(\sqrt{C_{\mathrm{reg}}F_\lambda(x)} + \sqrt{b_{\lambda,u}}\right)^2 \\ &\overset{⑥}{\le} 2C_{\mathrm{reg}}F_\lambda(x) + 2b_{\lambda,u} \overset{⑦}{\le} C\left(F_\lambda(x) + b_{\lambda,u}\right) \quad \text{on } \Omega_\lambda(x).\end{aligned} \tag{D.25}$$

Here ⑤ combines (D.16) with the regularized bound. ⑥ uses $(a+b)^2 \le 2a^2 + 2b^2$. ⑦ uses $C = 4C_{\mathrm{reg}}$ and $C_{\mathrm{reg}} \ge 1$.

For the fixed-ridge assertion, impose (D.20). Equation (D.16) then yields

$$0 \le b_{\lambda,u} \overset{⑧}{\le} \frac{\lambda R_0^2}{4} \overset{⑨}{\le} \frac{L_\lambda}{N_{u,m}} \overset{①}{\le} F_\lambda(x).$$

Here ⑧ uses the deterministic oracle-bias bound. ⑨ uses (D.20). ① uses $x \ge 1$ and $r_{K,\lambda} \ge 0$.

Inserting this estimate into the first inequality of (D.25) gives

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \overset{②}{\le} (\sqrt{C_{\mathrm{reg}}}+1)^2 F_\lambda(x) \overset{③}{\le} 4C_{\mathrm{reg}}F_\lambda(x) = CF_\lambda(x).$$

Here ② inserts $b_{\lambda,u} \le F_\lambda(x)$. ③ uses $\sqrt{C_{\mathrm{reg}}} + 1 \le 2\sqrt{C_{\mathrm{reg}}}$.

This proves (D.21). Condition (D.20) holds at $\lambda = 0$. When $L_0 > 0$, (D.18) supplies a positive interval of admissible ridge values.

For (D.22), use the event $\Omega_0(x)$ from (D.24) at zero ridge. When $L_0 > 0$, the localization lemma gives $D_0 > 0$. Since $D_0 \le 2R_0$, we have $R_0 > 0$. If $K_{\max} = 0$, all full-sample kernel sections have zero norm and the full span is $\{0\}$, contradicting $D_0 > 0$. As a result, $K_{\max} > 0$, the full Gram matrix is nonzero, and the definition of $K_{\min}^+$ gives $K_{\min}^+ > 0$. For every λ satisfying (D.23),

$$K_{\max}R_0^2\left(\frac{\lambda}{K_{\min}^+ + \lambda}\right)^2 \overset{④}{\le} K_{\max}R_0^2\frac{\lambda^2}{(K_{\min}^+)^2} \overset{⑤}{\le} \frac{L_0}{N_{u,m}}.$$

Here ④ decreases the positive denominator. ⑤ uses (D.23).

Equation (D.17) therefore gives, on $\Omega_0(x)$,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \overset{⑥}{\leq} \left(\sqrt{C_{\mathrm{reg}} F_0(x)} + \sqrt{L_0/N_{u,m}}\right)^2 \overset{⑦}{\leq} C F_0(x).$$

Here ⑥ uses the deterministic zero-ridge comparison on $\Omega_0(x)$. ⑦ uses $L_0/N_{u,m} \leq F_0(x)$ and $C = 4C_{\mathrm{reg}}$.

The comparison holds simultaneously for every allowed λ on $\Omega_0(x)$, hence also for any selection rule taking values in this interval. Such a rule is measurable on the finite split space, and no union bound over ridge values is needed. At zero ridge, the intermediate bound also covers $L_0 = 0$.

Finally, if $L_\lambda = 0$, the localization lemma gives $\widehat{f}_{m,\lambda} = \widehat{f}_{u,\lambda}$ on every split. The ordinary prediction-distance identity then gives

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \overset{⑧}{=} P_{\mathcal{U}}(\widehat{f}_{u,\lambda} - \widehat{f}_{u,0})^2 \overset{⑨}{\leq} b_{\lambda,u}.$$

Here ⑧ uses coincidence of the ridge subset solutions and the ordinary-excess identity. ⑨ uses the definition of $b_{\lambda,u}$.

Under (D.20), $L_\lambda = 0$ forces $b_{\lambda,u} = 0$, so the ordinary excess is zero, as required by (D.21). This includes $\lambda = 0$ without any division by R_0. Ordinary excess is always nonnegative by the same prediction-distance identity. The fixed-ridge assertion uses $\Omega_\lambda(x)$, whereas the simultaneous ridge-range assertion uses $\Omega_0(x)$. Each event has the stated probability over the random split separately, and neither conclusion requires their intersection.

□

D.2.3 Explicit spectral rates

Corollary D.4 (Explicit Spectral Rates, Including Zero Ridge)**.** Under the fixed-ridge condition (D.20), on $\Omega_\lambda(x)$ in Theorem D.3, and with the same universal constant C,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \leq C \min_{Q\in\{0,\dots,n\}} \left\{ \frac{L_\lambda}{N_{u,m}} \left(x + \sum_{j\leq Q} \frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda} \right) + D_\lambda \sqrt{\frac{L_\lambda}{N_{u,m}} \sum_{j>Q} \widehat{\lambda}_j} \right\}. \tag{D.26}$$

Replacing every $\widehat{\lambda}_j$ in (D.26) by $\lambda_j(\mathbf{K}/n)$ gives a valid, possibly larger bound on the same event, with D_λ, L_λ, C unchanged. For the separate conclusion on $\Omega_0(x)$, write $\widehat{\lambda}_j^{(0)} = \lambda_j(\mathbf{K}_0/n)$ for the zero-ridge projected eigenvalues. Without requiring (D.20) at a positive ridge, this event gives, for every ridge choice allowed in part (2) of Theorem D.3,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \leq C \min_{0\leq Q\leq \mathrm{rank}(\mathbf{K}_0)} \left\{ \frac{L_0(Q+x)}{N_{u,m}} + D_0 \sqrt{\frac{L_0}{N_{u,m}} \sum_{j>Q} \widehat{\lambda}_j^{(0)}} \right\}, \qquad Q \text{ integer.}$$

Its full-Gram counterpart replaces $\mathrm{rank}(\mathbf{K}_0)$ by $\mathrm{rank}(\mathbf{K})$ and $\widehat{\lambda}_j^{(0)}$ by $\lambda_j(\mathbf{K}/n)$, with the same D_0, L_0, C. If $\widehat{\lambda}_j^{(0)} \leq C_K j^{-2\beta}$ for $1 \leq j \leq n$, where $C_K > 0$ and $\beta > 1/2$, then on $\Omega_0(x)$, for the same ridge range,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \leq C_\beta \left[\min \left\{ \frac{L_0 \mathrm{rank}(\mathbf{K}_0)}{N_{u,m}}, (C_K D_0^2)^{1/(2\beta+1)} \left(\frac{L_0}{N_{u,m}} \right)^{2\beta/(2\beta+1)} \right\} + \frac{L_0 x}{N_{u,m}} \right]. \tag{D.27}$$

Here C_β depends only on β. A bound $\lambda_j(\mathbf{K}/n) \leq C_K j^{-2\beta}$ on the original spectrum also suffices, since it implies the displayed bound on the projected eigenvalues.

In (D.26), curvature controls the first Q directions through the sum $\sum_{j\leq Q} \widehat{\lambda}_j/(\widehat{\lambda}_j + \lambda)$. This weighted count is their regularized effective dimension. At zero ridge it counts their positive eigenvalues. The diameter D_λ controls the remaining spectral tail. For splits with $m \asymp u$, (D.27) gives $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \lesssim n^{-2\beta/(2\beta+1)} + x/n$ when $\beta > 1/2$ is fixed and C_K, D_0, L_0 stay uniformly bounded. The comparison constant depends only on β, these bounds, and the constants in $m \asymp u$.

Proof of Corollary D.4. Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$, and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$ are the eigenvalues of the projected matrix $\mathbf{K}_\lambda/n$. The quantities D_λ and L_λ are the RKHS diameter and residual maximum of the subset solutions. When $L_\lambda > 0$, the radius $r_{K,\lambda}$ is the unique positive solution of $r^2 = (L_\lambda/N_{u,m}) \sum_j \min\left\{r\widehat{\lambda}_j/(\widehat{\lambda}_j + \lambda), D_\lambda^2 \widehat{\lambda}_j\right\}$, as proved in Lemma D.1. When $L_\lambda = 0$, the radius is zero. At cutoff Q, the first Q ordered indices form the head and the indices $j > Q$ form the tail.

Under (D.20), Theorem D.3 bounds ordinary excess by $C(r_{K,\lambda} + L_\lambda x/N_{u,m})$ on $\Omega_\lambda(x)$. Its zero-ridge comparison uses $\Omega_0(x)$ and the zero-ridge diameter and residual maximum, even for the positive-ridge learners in (D.23). Our task is to evaluate $r_{K,\lambda}$, with this same C and without an additional probability event.

Proof outline.

(1) Exact cutoff identity. Express the spectral sum as the pointwise minimum over head–tail cutoffs. Solve each quadratic inequality using its larger root, then minimize over the cutoffs.
(2) Theorem substitution and zero ridge. We simplify the quadratic roots and insert the result into the ordinary-excess bounds of Theorem D.3 under the corresponding ridge restrictions. We compare the head and tail at every cutoff using the eigenvalue order, including the zero-eigenvalue convention at zero ridge, and then minimize. The constant and event remain unchanged.
(3) EDR. We bound the tail sum, balance the two terms, and check both endpoint cases.

Step 1: Exact cutoff identity. The ordered switch in (D.12) reduces the fixed-point inequality to a finite family of quadratics. For this proof only, put

$$a_Q := \frac{L_\lambda}{N_{u,m}} \sum_{j \leq Q} \frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \qquad b_Q := \frac{L_\lambda D_\lambda^2}{N_{u,m}} \sum_{j > Q} \widehat{\lambda}_j, \qquad Q \in \{0, \ldots, n\}.$$

These are nonnegative deterministic numbers. The coefficient a_Q measures the regularized spectral head and b_Q is the squared tail scale. We will prove

$$r_{K,\lambda} = \min_{Q \in \{0,\ldots,n\}} \frac{a_Q + \sqrt{a_Q^2 + 4b_Q}}{2}. \tag{D.28}$$

If $L_\lambda = 0$, all coefficients and both sides are zero. If $D_\lambda = 0$ or $\mathbf{K}_\lambda = \mathbf{0}$, the radius is zero and $a_0 = b_0 = 0$, which again proves the identity. In the remaining case $L_\lambda, D_\lambda > 0$ and the projected matrix has a positive eigenvalue.

Pointwise minimization over cutoffs. Fix $r > 0$. For any cutoff Q, the spectral sum is at most $a_Q r + b_Q$: this expression uses the curvature candidate on the first Q coordinates and the diameter candidate on the rest. For $\widehat{\lambda}_j > 0$, cancelling that eigenvalue shows that the curvature candidate is no larger exactly when $r \leq D_\lambda^2(\widehat{\lambda}_j + \lambda)$. The denominator $\widehat{\lambda}_j + \lambda$ is positive, so this equivalence holds also at zero ridge. Since the eigenvalues are nonincreasing, the qualifying indices form an initial segment. Choosing its length as Q, with equality assigned to the head, selects the smaller candidate in every positive eigendirection. Zero eigenvalues contribute zero to both candidates. As a result,

$$\frac{L_\lambda}{N_{u,m}} \sum_{j=1}^{n} \min\left\{r\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, D_\lambda^2 \widehat{\lambda}_j\right\} \overset{①}{=} \min_{Q \in \{0,\ldots,n\}} (a_Q r + b_Q). \tag{D.29}$$

Here ① chooses the cutoff selecting the smaller spectral candidate in every direction.

The minimizing cutoff in (D.29) may depend on r.

Solving the finite family of quadratic inequalities. For each Q, let $\rho_Q = (a_Q + \sqrt{a_Q^2 + 4b_Q})/2$. In the nondegenerate case under consideration, a_Q and b_Q cannot both vanish. Indeed, $b_0 > 0$, and $a_Q > 0$ whenever $Q \geq 1$ because the largest eigenvalue is positive. As a result, $\rho_Q > 0$.

The other quadratic root is nonpositive. Hence, for $r > 0$, $a_Q r + b_Q \leq r^2$ exactly when $r \geq \rho_Q$. The minimum in (D.29) is finite, so its admissible radii form $\bigcup_{Q=0}^{n} [\rho_Q, \infty) = [\min_Q \rho_Q, \infty)$. Taking the infimum proves (D.28).

Step 2: Theorem substitution and zero ridge. Simplifying each quadratic root gives the stated head–tail expression while preserving the event and constant.

Simplifying the quadratic roots.

For $a, b \geq 0$, let $\rho = (a + \sqrt{a^2 + 4b})/2$. The inequality $\sqrt{a^2 + 4b} \leq a + 2\sqrt{b}$ gives $\rho \leq a + \sqrt{b}$. If $\rho > 0$, write $v = a/\rho \in [0, 1]$. Since $b/\rho^2 = 1 - v$, $(a + \sqrt{b})/\rho = v + \sqrt{1 - v} = 5/4 - (\sqrt{1 - v} - 1/2)^2 \leq 5/4$. If $\rho = 0$, then $a = b = 0$, and both bounds hold as well. Apply these inequalities to each cutoff and use (D.28) to obtain

$$r_{K,\lambda} \leq \min_Q (a_Q + \sqrt{b_Q}) \leq \tfrac{5}{4} r_{K,\lambda}.$$

For the second inequality, evaluate the minimum at a cutoff attaining $\min_Q \rho_Q$.

Substitution on the theorem's event. We use only the first inequality in the risk bound, so the factor $5/4$ will not multiply its constant. Under (D.20), apply (D.21) on $\Omega_\lambda(x)$ to obtain

$$\begin{aligned}\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) &\overset{②}{\leq} C\left(r_{K,\lambda} + \frac{L_\lambda x}{N_{u,m}}\right) \\ &\overset{③}{\leq} C\left(\min_Q (a_Q + \sqrt{b_Q}) + \frac{L_\lambda x}{N_{u,m}}\right) \overset{④}{=} C \min_Q \left\{a_Q + \sqrt{b_Q} + \frac{L_\lambda x}{N_{u,m}}\right\}.\end{aligned}$$

Here ② applies (D.21). ③ bounds the fixed point by the minimum of the simplified roots. ④ moves the cutoff-independent confidence term inside the minimum.

Substituting a_Q, b_Q proves (D.26) with the same C. All cutoff bounds hold deterministically on the theorem's event, so minimization requires no union bound.

Replacement by the original Gram spectrum. For the full-Gram consequence of (D.26), we use $\widehat{\lambda}_j \leq \lambda_j(\mathbf{K}/n)$ from Lemma D.1. If $\lambda > 0$, the derivative of $t/(t + \lambda)$ is $\lambda/(t + \lambda)^2 \geq 0$ for $t \geq 0$. At $\lambda = 0$, the specified convention gives $t/(t + \lambda) = \mathbb{1}_{\{t>0\}}$, a nondecreasing function on $[0, \infty)$. Consequently, for every cutoff Q,

$$\sum_{j \leq Q} \frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda} \overset{⑤}{\leq} \sum_{j \leq Q} \frac{\lambda_j(\mathbf{K}/n)}{\lambda_j(\mathbf{K}/n) + \lambda}, \qquad \sum_{j > Q} \widehat{\lambda}_j \overset{⑥}{\leq} \sum_{j > Q} \lambda_j(\mathbf{K}/n).$$

Here ⑤ uses eigenvalue domination and monotonicity of the head ratio. ⑥ sums eigenvalue domination over the tail.

These inequalities hold at each cutoff, so they also compare the minima. The full-Gram bound follows on $\Omega_\lambda(x)$ with the same C, D_λ, L_λ and ridge restriction. No common eigenbasis is required.

Specialization to zero ridge. We now evaluate the deterministic spectral fixed point at zero ridge. On $\Omega_0(x)$, (D.22) controls ordinary excess for every learner in the stated ridge range by $C(r_{K,0} + L_0 x/N_{u,m})$. As a result, the following spectral calculation applies simultaneously to $\lambda = 0$ and, when $L_0 > 0$, every positive ridge satisfying (D.23). Throughout the remaining calculation we use $\widehat{\lambda}_j^{(0)} = \lambda_j(\mathbf{K}_0/n)$ for the zero-ridge spectrum. The learner's ridge parameter may still be positive. At zero ridge, the head ratio is one for a positive eigenvalue and zero otherwise. Its sum through any integer cutoff Q is $\min\{Q, \mathrm{rank}(\mathbf{K}_0)\}$, and the tail sum vanishes for $Q \geq \mathrm{rank}(\mathbf{K}_0)$. Restricting the minimum to $0 \leq Q \leq \mathrm{rank}(\mathbf{K}_0)$ therefore preserves its value and proves the projected zero-ridge rate. For the full-Gram bound the identical argument uses $\lambda_j(\mathbf{K}/n)$ and $\mathrm{rank}(\mathbf{K})$ instead, without changing D_0 or L_0. Choosing the projected rank as cutoff gives the sharper rank bound used below.

Step 3: EDR. We retain the zero-ridge spectrum, diameter, residual maximum, and ridge range from Step 2. The EDR assumption is $\widehat{\lambda}_j^{(0)} \leq C_K j^{-2\beta}$ with $C_K > 0$ and $\beta > 1/2$. This last restriction is what makes the infinite comparison tail summable. When $L_0 = 0$, the zero-ridge comparison invoked here asserts the bound for $\lambda = 0$ only. Its positive-ridge range requires $L_0 > 0$. The cases $L_0 = 0$ or $D_0 = 0$ have zero excess by Theorem D.3, so assume both are positive. Step 2 reduces the remaining task to bounding

$$\min_{\substack{0 \leq Q \leq \mathrm{rank}(\mathbf{K}_0) \\ Q \text{ integer}}} \left\{\frac{L_0 Q}{N_{u,m}} + D_0 \sqrt{\frac{L_0}{N_{u,m}} \sum_{j > Q} \widehat{\lambda}_j^{(0)}}\right\}.$$

The confidence term $L_0 x/N_{u,m}$ is independent of Q and will be added after this deterministic calculation.

Tail estimates. For an integer $Q \geq 1$, the function $s \mapsto s^{-2\beta}$ is decreasing, so

$$\sum_{j=Q+1}^{n} \widehat{\lambda}_j^{(0)} \overset{⑦}{\leq} C_K \sum_{j=Q+1}^{\infty} j^{-2\beta} \overset{⑧}{\leq} C_K \int_Q^{\infty} s^{-2\beta}\, ds \overset{⑨}{=} \frac{C_K Q^{1-2\beta}}{2\beta-1}.$$

Here ⑦ uses the EDR assumption and extends the nonnegative sum. ⑧ compares a decreasing series with its integral. ⑨ evaluates the convergent integral using $\beta > 1/2$.

At $Q = 0$, separating the first eigenvalue and applying the bound at cutoff one gives $\sum_{j=1}^{n} \widehat{\lambda}_j^{(0)} \leq C_K[1 + (2\beta-1)^{-1}]$.

The balancing scale. For $Q \leq \text{rank}(\mathbf{K}_0)$, the head is $L_0 Q/N_{u,m}$. Equating it to $D_0\sqrt{C_K L_0/N_{u,m}}\, Q^{(1-2\beta)/2}$ and dividing by the positive factors gives $Q^{(2\beta+1)/2} = D_0\sqrt{C_K N_{u,m}/L_0}$. Squaring therefore suggests the real cutoff $t = (C_K D_0^2 N_{u,m}/L_0)^{1/(2\beta+1)}$. The factor $(2\beta-1)^{-1/2}$ from the tail estimate remains in the final constant. We choose an admissible integer cutoff in each of the following three cases, which cover all $t > 0$ because the projected rank is a positive integer.

Case 1: The balancing cutoff is below one. If $t < 1$, we choose $Q = 0$. There is no head term, and the tail satisfies

$$D_0\sqrt{\frac{L_0}{N_{u,m}}\sum_{j=1}^{n}\widehat{\lambda}_j^{(0)}} \overset{①}{\leq} \sqrt{1+\frac{1}{2\beta-1}}\,\frac{L_0}{N_{u,m}} t^{(2\beta+1)/2} \overset{②}{\leq} \sqrt{1+\frac{1}{2\beta-1}}\,\frac{L_0 t}{N_{u,m}}.$$

Here ① uses the full spectral-sum bound and the definition of t. ② uses $0 < t < 1$ and $(2\beta+1)/2 > 1$.

Case 2: The balancing cutoff is between one and the rank. If $1 \leq t < \text{rank}(\mathbf{K}_0)$, we choose $Q = \lceil t \rceil$, the smallest integer at least t. Because the rank is an integer, $Q \leq \text{rank}(\mathbf{K}_0)$, while $t \leq Q < t+1 \leq 2t$. Hence the head satisfies $\frac{L_0 Q}{N_{u,m}} \leq \frac{2L_0 t}{N_{u,m}}$. Apply the tail estimate at the integer $Q \geq 1$. Since $Q \geq t$ and $(1-2\beta)/2 < 0$, it gives

$$\begin{aligned} D_0\sqrt{\frac{L_0}{N_{u,m}}\sum_{j>Q}\widehat{\lambda}_j^{(0)}} &\overset{③}{\leq} \frac{D_0\sqrt{C_K L_0/N_{u,m}}}{\sqrt{2\beta-1}}\, Q^{(1-2\beta)/2} \\ &\overset{④}{\leq} \frac{D_0\sqrt{C_K L_0/N_{u,m}}}{\sqrt{2\beta-1}}\, t^{(1-2\beta)/2} \overset{⑤}{=} \frac{L_0 t}{N_{u,m}\sqrt{2\beta-1}}. \end{aligned}$$

Here ③ uses the tail estimate under EDR. ④ uses $Q \geq t$ and the negative tail exponent. ⑤ uses the balancing identity for t.

Adding head and tail yields the multiplier $2 + (2\beta-1)^{-1/2}$.

Case 3: The balancing cutoff is at or above the rank. If $t \geq \text{rank}(\mathbf{K}_0)$, choose $Q = \text{rank}(\mathbf{K}_0)$. Then

$$\frac{L_0 Q}{N_{u,m}} + D_0\sqrt{\frac{L_0}{N_{u,m}}\sum_{j>Q}\widehat{\lambda}_j^{(0)}} \overset{⑥}{=} \frac{L_0 \text{rank}(\mathbf{K}_0)}{N_{u,m}} \overset{⑦}{\leq} \frac{L_0 t}{N_{u,m}}.$$

Here ⑥ uses vanishing eigenvalues beyond the rank. ⑦ uses $t \geq \text{rank}(\mathbf{K}_0)$.

Combining the cutoff and rank bounds. The three cases bound the optimized complexity by a constant depending only on β times $L_0 t/N_{u,m}$. In addition, the rank cutoff always gives the bound $L_0 \text{rank}(\mathbf{K}_0)/N_{u,m}$. To express the balancing scale in the original parameters, we compute

$$\frac{L_0 t}{N_{u,m}} = \frac{L_0}{N_{u,m}}\left(\frac{C_K D_0^2 N_{u,m}}{L_0}\right)^{1/(2\beta+1)} \overset{⑧}{=} (C_K D_0^2)^{1/(2\beta+1)} \left(\frac{L_0}{N_{u,m}}\right)^{2\beta/(2\beta+1)}.$$

Here ⑧ collects the powers of $L_0/N_{u,m}$ and $C_K D_0^2$.

Set $c = 2 + (2\beta - 1)^{-1/2}$ and $C_\beta = Cc$. This c dominates all three case constants because $\sqrt{1+s^2} \le 1 + s$ for $s \ge 0$. Combining the balancing and rank bounds and using $C \le C_\beta$ gives, on $\Omega_0(x)$,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda}) \overset{⑨}{\le} C_\beta \left[\min\left\{ \frac{L_0 \mathrm{rank}(\mathbf{K}_0)}{N_{u,m}}, \frac{L_0 t}{N_{u,m}} \right\} + \frac{L_0 x}{N_{u,m}} \right].$$

Here ⑨ combines the optimized cutoff bound with the separate rank bound on the same event.

This holds for $\lambda = 0$ and every positive ridge allowed by (D.23), as in Step 2. Inserting the displayed identity for $L_0 t / N_{u,m}$ proves (D.27). The constant C_β depends only on β and the universal constant C. No factor depending on D_0, L_0, C_K, or n has been absorbed into it, and no probability event has been added. □

D.2.4 Proof of Theorem 4.3

Proof of Theorem 4.3. We compare the chosen ridge learner with the minimum-norm learner on the zero-ridge event. Recall that $\mathcal{F} = \mathrm{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$, $\varepsilon > 0$ is the prediction tolerance, and $\lambda_{\mathcal{M}}$ is chosen by the two cases in Theorem 4.3. When $\mathbf{K}_{\mathcal{M}} \neq \mathbf{0}$, let $\kappa_{\mathcal{M}}$ denote the smallest positive eigenvalue of $\mathbf{K}_{\mathcal{M}}/m$. The rule is then $\lambda_{\mathcal{M}} = \kappa_{\mathcal{M}} \varepsilon / (\sqrt{K_{\max}} \left\| \widehat{f}_{m,0} \right\|_{\mathcal{H}_K} + \varepsilon)$, where $K_{\max} = \max_{i \in [n]} K(\mathbf{x}_i, \mathbf{x}_i)$. When $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$, any fixed positive ridge is prescribed instead.

The event $\Omega_0(x)$ is the zero-ridge event of Theorem D.3. Its fixed-ridge restriction (D.20) holds at zero ridge. The argument below uses this event and does not require that restriction at the training-dependent positive value $\lambda_{\mathcal{M}}$.

Proof outline.

(1) Training-subset shrinkage. Treat the zero training Gram matrix directly, then apply the spectral difference formula when it is nonzero.
(2) Prediction tolerance. Use the chosen ridge and the reproducing property on all test coordinates.
(3) Ordinary excess on the zero-ridge event. Apply (D.15) with the zero-ridge training subset solution as comparison predictor, then insert the zero-ridge bounds. Verify simultaneity over tolerances.

Step 1: Training-subset shrinkage. The ridge rule uses training observations, so its perturbation bound must use the training subset spectrum alone. We first check the zero-matrix branch without invoking a positive eigenvalue.

If $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$, then $T^*_{\mathcal{M}} = 0$ by Lemma D.1. Every training prediction is zero, so $J_{\mathcal{M}}(f) = P_{\mathcal{M}} y^2 + \lambda \|f\|^2_{\mathcal{H}_K}$. Its unique minimizer at positive ridge and minimum-norm selection at zero ridge are both zero. Consequently $\widehat{f}_{m,\lambda} = \widehat{f}_{m,0} = 0$ on the whole input domain, and their prediction change is zero for any fixed positive ridge. Neither $\kappa_{\mathcal{M}}$ nor (4.4) is used in this branch.

In the remaining case $\mathbf{K}_{\mathcal{M}} \neq \mathbf{0}$, positive semidefiniteness and finite dimension ensure $\kappa_{\mathcal{M}} > 0$. The first identity of (D.19), with $A = \mathcal{M}$, gives for every $\lambda \ge 0$

$$\left\| \widehat{f}_{m,\lambda} - \widehat{f}_{m,0} \right\|_{\mathcal{H}_K} \overset{①}{\le} \frac{\lambda}{\kappa_{\mathcal{M}} + \lambda} \left\| \widehat{f}_{m,0} \right\|_{\mathcal{H}_K}.$$

Here ① bounds every spectral shrinkage factor by its value at $\kappa_{\mathcal{M}}$ and takes square roots.

The rule uses no test labels. Indeed, with $\mathbf{y}_{\mathcal{M}} = (y_i)_{i \in \mathcal{M}}$ ordered as the training Gram matrix, $\left\| \widehat{f}_{m,0} \right\|^2_{\mathcal{H}_K} = \mathbf{y}_{\mathcal{M}}^\top \mathbf{K}^\dagger_{\mathcal{M}} \mathbf{y}_{\mathcal{M}}$. Both this norm and $\kappa_{\mathcal{M}}$ use training data, while $K_{\max}$ uses observed inputs. The numerator and denominator in (4.4) are positive.

Step 2: Prediction tolerance. The RKHS norm change controls every test prediction through evaluation. The ridge rule is chosen to keep that change within the prescribed tolerance. Step 1 already gives the tolerance bound when $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$. Otherwise put $M_{\mathcal{M}} = \sqrt{K_{\max}} \left\| \widehat{f}_{m,0} \right\|_{\mathcal{H}_K}$.

Substituting the ridge rule gives $\lambda_{\mathcal{M}}/(\kappa_{\mathcal{M}}+\lambda_{\mathcal{M}}) = \varepsilon/(M_{\mathcal{M}}+2\varepsilon)$. The reproducing inequality and Step 1 therefore give, for every $i \in [n]$,

$$\left|(\widehat{f}_{m,\lambda_{\mathcal{M}}} - \widehat{f}_{m,0})(\mathbf{x}_i)\right| \overset{②}{\leq} M_{\mathcal{M}} \frac{\lambda_{\mathcal{M}}}{\kappa_{\mathcal{M}}+\lambda_{\mathcal{M}}} = \frac{M_{\mathcal{M}}\varepsilon}{M_{\mathcal{M}}+2\varepsilon} \leq \varepsilon.$$

Step ② combines the evaluation bound with training-subset shrinkage. Squaring and averaging over $\mathcal{U}$ proves $P_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}} - \widehat{f}_{m,0})^2 \leq \varepsilon^2$.

Step 3: Ordinary excess on the zero-ridge event. The prediction perturbation can be applied on the zero-ridge event for every tolerance, avoiding a concentration argument at a random ridge. Apply (D.15) on each split with $f = \widehat{f}_{m,\lambda_{\mathcal{M}}}$ and $h = \widehat{f}_{m,0}$. Step 2 bounds their squared prediction difference by ε^2, including when the training Gram matrix is zero. In that case the two ordinary excess risks are equal, but need not vanish: the test minimizer may be nonzero. The condition $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$ does not imply $\mathbf{K} = \mathbf{0}$ or $L_0 = D_0 = 0$. We retain the zero-ridge risk bound on $\Omega_0(x)$, without excluding any splits with a zero training Gram matrix. As a result,

$$\sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}})} \overset{③}{\leq} \sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,0})} + \sqrt{P_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}} - \widehat{f}_{m,0})^2} \overset{④}{\leq} \sqrt{\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,0})} + \varepsilon.$$

Here ③ applies (D.15). ④ uses the prediction tolerance from Step 2.

Squaring and using $2ab \leq a^2 + b^2$ gives $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \leq 2\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,0}) + 2\varepsilon^2$. Choose the universal constant C from Theorem D.3 and C_β from Corollary D.4. The former, at $\lambda = 0$, gives on $\Omega_0(x)$ the intermediate fixed-point bound

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \overset{⑤}{\leq} 2C\left(r_{K,0} + \frac{L_0 x}{N_{u,m}}\right) + 2\varepsilon^2. \tag{D.30}$$

Here ⑤ uses $(a+b)^2 \leq 2a^2 + 2b^2$ and the zero-ridge theorem on $\Omega_0(x)$.

Step 2 of the proof of Corollary D.4 shows that

$$r_{K,0} \overset{⑥}{\leq} \min_{\substack{0 \leq Q \leq \operatorname{rank}(\mathbf{K}_0) \\ Q \text{ integer}}} \left\{ \frac{L_0 Q}{N_{u,m}} + D_0 \sqrt{\frac{L_0}{N_{u,m}} \sum_{j>Q} \widehat{\lambda}_j^{(0)}} \right\}.$$

Here ⑥ uses the cutoff bound proved in Corollary D.4.

Adding $L_0 x/N_{u,m}$, which does not depend on Q, inside this minimum and substituting into (D.30) proves (4.5). Eigenvalue domination in that corollary proves the full-Gram replacement with the same C, D_0, L_0. Its cutoff calculation under EDR (D.27), applied to $\widehat{f}_{m,0}$ before the deterministic comparison above, gives (4.6) with $2C_\beta$ and $2\varepsilon^2$. For the zero-ridge learner itself, no prediction perturbation is needed. The same corollary therefore proves both stated bounds with constants C, C_β and no tolerance term. These arguments also cover $L_0 = 0$ or zero projected rank, because the spectral expressions then vanish and the positive-ridge comparison retains the term $2\varepsilon^2$. All preceding deterministic comparisons hold for every $\varepsilon > 0$ on every split, while $\Omega_0(x)$ does not depend on ε. As a result, the same event proves the simultaneous assertion and has probability at least $1 - \exp(-x)$ over the random split. For fixed tolerance the rule is measurable on the finite split space. Only $\Omega_0(x)$ is used, without substituting a random ridge into $\Omega_\lambda(x)$. For $\varepsilon = N_{u,m}^{-1}$ the added term is $2N_{u,m}^{-2}$. To verify the stated rate consequences, retain the rank cutoff in Corollary D.4 for the first one and use its EDR bound for the second one. Under their respective uniform scale assumptions and fixed x, these give $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \lesssim N_{u,m}^{-1}$ and $\mathcal{E}_{\mathcal{U}}(\widehat{f}_{m,\lambda_{\mathcal{M}}}) \lesssim N_{u,m}^{-2\beta/(2\beta+1)}$, respectively, with the dependence specified in the main-text rate discussion. Since $N_{u,m} \geq 1$ and $0 < 2\beta/(2\beta+1) < 1$, the added term $2N_{u,m}^{-2}$ has no larger order than either bound. The exact finite-sample assertion remains (4.5), including its additive tolerance term when the zero-ridge bound vanishes.

□

D.3 Prediction error with sub-Gaussian label noise

Proof of Theorem 4.4. Recall that the inputs are independent with law ρ, the random split is independent of them and the noises, and $u \geq m \geq 2$, so $N_{u,m} = m$. The signal $f^* \in \mathcal{H}_K$ is fixed before

sampling. Write $\lambda = \lambda_m = m^{-\alpha}$, where $\alpha = 2\beta/(2\beta+1)$, and put $x = m\lambda = m^{1/(2\beta+1)}$ and $Q = \lceil x \rceil$. These are deterministic, $x \geq 1$, and $Q \leq 2x$. In this proof, C denotes a finite positive constant depending only on β, a_+, κ, whose value can increase from one bound to the next. Steps 1–7 treat the oracle ridge and track four failure events, each of probability at most e^{-x}. Step 8 proves the data-based alternative (4.10) with its separate confidence guarantee.

Proof outline.

(1) Population and empirical spectra. Bound the empirical spectral tails on the full and training samples using population EDR.
(2) UTLC for squared prediction differences. Verify the two localized TC bounds on normalized predictions and extend the resulting comparison to every RKHS difference by homogeneity.
(3) Fixed point and effective dimension. Evaluate both spectral quantities at the chosen ridge scale.
(4) Ridge normal equation and noise. Bound training prediction error plus the ridge-weighted squared error norm.
(5) Joint probability. Combine the uniform UTLC comparison and the ridge bound, accounting for all conditioning.
(6) Expected error. Control the exceptional event to prove (4.9).
(7) Critical population rate and an example. Identify the critical scale and verify a concrete family satisfying both bounds.
(8) Data-based ridge. Choose the ridge from the full empirical spectrum, control the training effective dimension conditionally on the inputs, and prove the same expected rate without knowing the population EDR.

Step 1: Population and empirical spectra. We first control the empirical spectrum without assuming uniformly bounded eigenfunctions. For a nonempty index set A, let $\widehat{T}_A h = |A|^{-1} \sum_{i \in A} h(\mathbf{x}_i) K_{\mathbf{x}_i}$ be its empirical integral operator on $\mathcal{H}_K$. Its ordered eigenvalues are $\widehat{\nu}_{A,j}$, extended by zeros. The nonzero ones are the eigenvalues of $\mathbf{K}_A/|A|$. Indeed, the evaluation map $h \mapsto (h(\mathbf{x}_i)/\sqrt{|A|})_{i \in A}$ has these two operators as its adjoint product and its product with its adjoint. The range and positive-eigenvalue argument in Section D.1 therefore applies. Write $\widehat{T}_n = \widehat{T}_{[n]}$ and $\widehat{\nu}_j = \widehat{\nu}_{[n],j}$. Every $h \in \mathcal{H}_K$ is a pointwise limit of finite combinations of measurable kernel sections, hence is measurable. Separability then makes $\mathbf{x} \mapsto K_{\mathbf{x}}$ strongly measurable, and $\|h(\mathbf{x})K_{\mathbf{x}}\|_{\mathcal{H}_K} \leq \kappa^2 \|h\|_{\mathcal{H}_K}$ ensures Bochner integrability. Reproduction gives $\langle h, T_\rho h' \rangle_{\mathcal{H}_K} = \int hh' \, d\rho$, so T_ρ is positive and self-adjoint. For an RKHS orthonormal basis (e_j),

$$\operatorname{tr}(T_\rho) \overset{①}{=} \sum_j \int e_j(\mathbf{x})^2 \, d\rho(\mathbf{x}) \overset{②}{=} \int \|K_{\mathbf{x}}\|_{\mathcal{H}_K}^2 \, d\rho(\mathbf{x}) \leq \kappa^2.$$

Step ① uses $\langle h, T_\rho h \rangle_{\mathcal{H}_K} = \int h^2 \, d\rho \geq 0$. Step ② interchanges nonnegative summands and the integral, then applies Parseval to $K_{\mathbf{x}}$. This proves that T_ρ is trace class and justifies its spectral decomposition. The same calculation with a projected section will identify the spectral tail below.

Let Π_Q^ρ be the orthogonal projection onto eigenvectors for the Q largest positive eigenvalues of T_ρ, or onto all its positive eigenspaces if its rank is less than Q. Choose an orthonormal basis within any repeated eigenspace. Define $z_Q(\mathbf{x}) = \left\| (I - \Pi_Q^\rho) K_{\mathbf{x}} \right\|_{\mathcal{H}_K}^2$ and $t_Q = \sum_{j>Q} \nu_j$, where I is the identity on $\mathcal{H}_K$. The spectral theorem and the kernel bound give

$$0 \leq z_Q(\mathbf{x}) \leq \kappa^2, \qquad \mathbb{E}_\rho[z_Q(\mathbf{x})] \overset{③}{=} \operatorname{tr}\left((I - \Pi_Q^\rho) T_\rho\right) \overset{④}{=} t_Q \overset{⑤}{\leq} \frac{a_+}{2\beta - 1} Q^{1-2\beta}.$$

Step ③ integrates the squared projected section, ④ removes the first Q eigenvalues, and ⑤ uses $\sum_{j>Q} j^{-2\beta} \leq \int_Q^\infty t^{-2\beta} \, dt$. The population projection need not commute with $\widehat{T}_A$. We compare their traces using projection weights, which avoids this restriction. Let Π be any orthogonal projection of rank at most Q, choose orthonormal eigenvectors v_j for the positive eigenvalues $\widehat{\nu}_{A,j}$, and put $w_j = \|\Pi v_j\|_{\mathcal{H}_K}^2$. Append zero weights when the empirical eigenvalues are zero. Then $0 \leq w_j \leq 1$. For an orthonormal basis (u_l) of range(Π), Bessel's inequality gives

$$\sum_j w_j = \sum_l \sum_j |\langle v_j, u_l \rangle_{\mathcal{H}_K}|^2 \overset{⑥}{\leq} \sum_l \|u_l\|_{\mathcal{H}_K}^2 = \operatorname{rank}(\Pi) \leq Q.$$

Step ⑥ applies Bessel's inequality for each u_l. Since the empirical eigenvalues are nonincreasing,

$$\operatorname{tr}\left(\Pi\widehat{T}_A\right) - \sum_{j\le Q}\widehat{\nu}_{A,j} = \sum_{j\le Q}\widehat{\nu}_{A,j}(w_j-1) + \sum_{j>Q}\widehat{\nu}_{A,j}w_j \overset{⑦}{\le} \widehat{\nu}_{A,Q}\left(\sum_j w_j - Q\right) \le 0.$$

In ⑦, replace each eigenvalue by $\widehat{\nu}_{A,Q}$, using $w_j - 1 \le 0$ in the head and $w_j \ge 0$ in the tail. Zero padding makes the calculation valid even when the empirical rank is less than Q. Apply it to $\Pi = \Pi_Q^\rho$ and subtract from the full trace to obtain

$$\sum_{j>Q}\widehat{\nu}_{A,j} \overset{⑧}{\le} \operatorname{tr}\left((I-\Pi_Q^\rho)\widehat{T}_A\right) \overset{⑨}{=} P_A z_Q.$$

Here ⑧ subtracts the projection trace from the full trace, and ⑨ expands $\widehat{T}_A$ into its rank-one summands.

For completeness, the scalar tail bound needed here follows directly from an exponential moment. Convexity gives $e^v \le 1+(e-1)v$ for $0 \le v \le 1$, and $1+w \le e^w$ for $w \ge 0$. Hence $\mathbb{E}_\rho\left[e^{z_Q/\kappa^2}\right] \le \exp((e-1)t_Q/\kappa^2)$. For k independent inputs, Markov's inequality and independence give

$$\Pr\left\{\frac{1}{k}\sum_{i=1}^k z_Q(\mathbf{x}_i) > 2t_Q + \frac{\kappa^2 x}{k}\right\} \overset{①}{\le} e^{-2kt_Q/\kappa^2 - x}\left(\mathbb{E}_\rho\left[e^{z_Q/\kappa^2}\right]\right)^k \overset{②}{\le} e^{-x}.$$

Step ① is Markov's inequality, and ② uses $e-1<2$. Apply this first to all n inputs and then to the m training inputs conditional on any fixed split. Such training inputs are independent with law ρ. Integrating over the random split and taking a union bound gives an event of probability at least $1-2e^{-x}$ over the inputs and random split on which

$$\sum_{j>Q}\widehat{\nu}_{A,j} \le 2t_Q + \frac{\kappa^2 x}{|A|}, \qquad A \in \{[n], \mathcal{M}\}. \tag{D.31}$$

No independence between the two tail events is asserted or needed.

Step 2: UTLC for squared prediction differences. The next step gives a comparison uniform over prediction differences. This uniformity is essential because the learner's difference depends on the split and training noise. We first normalize the directions, verify both localized TC bounds, and then undo the normalization. Fix the full input sample, and define

$$\mathcal{Q} = \left\{h_v : h_v(i) = v(\mathbf{x}_i)^2/\kappa^2,\ \|v\|_{\mathcal{H}_K} \le 1\right\}.$$

Use $V(h) = P_n h$ and $\mathcal{Q}(r) = \{h \in \mathcal{Q} : P_n h \le r\}$ for localization. The reproducing property gives $0 \le h \le 1$ on $[n]$. Hence $\operatorname{Var}[h] \le P_n h^2 \le P_n h = V(h)$ and $B_\infty(\mathcal{Q}) \le 1$. The class contains zero. Its evaluation image is compact. In fact, orthogonal projection onto the finite-dimensional closed span $\mathcal{H}_{\mathbf{X}_n}$ preserves all sample evaluations and cannot increase the norm. Indeed, its orthogonal residual is orthogonal to every $K_{\mathbf{x}_i}$ and therefore vanishes at every sample input by reproduction. The finite-dimensional unit ball is compact, and evaluation followed by squaring is continuous.

All suprema below are consequently finite and measurable. The same projection argument proves the finite-sample signal assertion in the main text.

We verify both parts of (4.1) explicitly. Put $e_j = \widehat{\nu}_j/\kappa^2$. Consider the linear prediction class

$$\mathcal{V}(r) = \left\{w = v/\kappa : \|v\|_{\mathcal{H}_K} \le 1,\ P_n w^2 \le r\right\}.$$

Since $\left\langle h, \widehat{T}_n h\right\rangle_{\mathcal{H}_K} = P_n h^2$, its null space consists exactly of functions vanishing on the sample, or equivalently of $\mathcal{H}_{\mathbf{X}_n}^\perp$. The positive-eigenvalue eigenvectors therefore form an orthonormal basis of $\mathcal{H}_{\mathbf{X}_n}$.

Choose these RKHS-orthonormal eigenvectors v_j, and set $\phi_j = v_j/\sqrt{\widehat{\nu}_j}$ on the positive eigendirections. The identity $P_n(hh') = \left\langle h, \widehat{T}_n h'\right\rangle_{\mathcal{H}_K}$ gives $P_n(\phi_j\phi_k) = \mathbb{I}_{\{j=k\}}$, while RKHS orthonormality gives $\langle\phi_j,\phi_k\rangle_{\mathcal{H}_K} = \mathbb{I}_{\{j=k\}}/\widehat{\nu}_j$. Replace w by its projection onto $\mathcal{H}_{\mathbf{X}_n}$, preserving the constraints

and every sample evaluation. Write $w = \sum_j b_j \phi_j$. The two localization constraints are now

$$\sum_j b_j^2 \overset{③}{=} P_n w^2 \le r, \qquad \sum_j \frac{b_j^2}{e_j} \overset{④}{=} \kappa^2 \|w\|_{\mathcal{H}_K}^2 \le 1.$$

Step ③ uses empirical orthonormality, and ④ uses RKHS orthogonality and $e_j = \widehat{\nu}_j/\kappa^2$. Both sums use positive eigendirections. If none exist, all sample predictions vanish. For $r > 0$, put $s_j = \min\{r, e_j\}$ on these positive directions. Then

$$\sum_j \frac{b_j^2}{s_j} \overset{⑤}{\le} \frac{\sum_j b_j^2}{r} + \sum_j \frac{b_j^2}{e_j} \overset{⑥}{\le} 2.$$

Step ⑤ uses $1/\min\{r, e_j\} \le 1/r + 1/e_j$, and ⑥ uses the two constraints. Let $J_1, \dots, J_m$ be independent uniform indices in $[n]$, and let $\sigma_1, \dots, \sigma_m$ be independent Rademacher variables, independent of those indices. Define $Z_j = m^{-1} \sum_{i=1}^m \sigma_i \phi_j(\mathbf{x}_{J_i})$.

As in Step 3 of Lemma D.1, independence of the Rademacher variables and uniform sampling give $\mathbb{E}_{J,\sigma}\left[Z_j^2\right] = m^{-1} P_n \phi_j^2 = 1/m$. The inductive RC defined in Section A therefore satisfies

$$\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{V}(r)) \overset{⑦}{\le} \mathbb{E}_{J,\sigma}\left[\sqrt{2\sum_j s_j Z_j^2}\right] \overset{⑧}{\le} \sqrt{\frac{2}{m}\sum_{j=1}^n \min\{r, e_j\}}.$$

Step ⑦ applies weighted Cauchy–Schwarz and the preceding coefficient bound. Step ⑧ uses Jensen's inequality and the second moments of Z_j. Appended zero eigenvalues contribute zero. At $r = 0$, every permitted prediction vanishes on the sample, so the same bound is zero without division by r.

The map $w \mapsto w^2$ is 2-Lipschitz on $[-1, 1]$ and vanishes at zero. Each $w \in \mathcal{V}(r)$ has values in this interval. Every member of $\mathcal{Q}(r)$ is the square of a member of $\mathcal{V}(r)$, and conversely. For either $\eta \in \{+, -\}$, the first-order TC therefore obeys

$$\mathfrak{R}_m^{\eta}(\mathcal{Q}(r)) \overset{⑨}{\le} 2\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{Q}(r)) \overset{①}{\le} 4\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{V}(r)) \overset{②}{\le} 4\sqrt{\frac{2}{m}\sum_{j=1}^n \min\{r, e_j\}}.$$

Step ⑨ is Theorem A.3, ① is Theorem A.2 for squaring, and ② is the preceding prediction RC bound. To verify the second level separately, observe that $h - P_n h \in [-1, 1]$ because $0 \le h \le 1$. A further contraction and Lemma A.1 give

$$\begin{aligned}\mathfrak{R}_m^{+}(\mathcal{Q}(r)^{[2]}) &\overset{③}{\le} 2\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{Q}(r)^{[2]}) \overset{④}{\le} 4\mathfrak{R}_m^{(\mathrm{ind})}(\{h - P_n h : h \in \mathcal{Q}(r)\}) \\ &\overset{⑤}{\le} 8\mathfrak{R}_m^{(\mathrm{ind})}(\mathcal{Q}(r)) \overset{⑥}{\le} 16\sqrt{\frac{2}{m}\sum_{j=1}^n \min\{r, e_j\}}.\end{aligned}$$

Here ③ is the TC comparison, ④ contracts squaring on $[-1, 1]$, ⑤ is ghost centering, and ⑥ uses the first squaring contraction and the prediction RC. Clipping the scalar maps to their stated intervals supplies globally Lipschitz extensions if needed. None of these RC definitions inserts an absolute value inside the supremum. Consequently a common majorant in (4.1) is

$$\psi(r) = 16\sqrt{2}\sqrt{\frac{1}{m}\sum_{j=1}^n \min\{r, \widehat{\nu}_j/\kappa^2\}}, \qquad r \ge 0. \tag{D.32}$$

The function is continuous and nondecreasing, and $\psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$ because every summand $\min\{1, e_j/r\}$ is nonincreasing. It is therefore sub-root. When $\mathbf{K} \neq \mathbf{0}$, its positive fixed point r^* exists and is finite by Lemma A.5. When $\mathbf{K} = \mathbf{0}$, take $r^* = 0$ and the comparison below holds directly.

Apply the one-sided part of Theorem 4.1 to $\mathcal{Q}$, with its absorption parameter equal to 2 and confidence exponent x. The envelope branch of the minimum gives an event of conditional probability at least $1 - e^{-x}$ over the random split on which

$$P_{\mathcal{U}} h - P_{\mathcal{M}} h \le \frac{1}{2} P_n h + 2c_1 r^* + 3c_1 x/m \qquad (h \in \mathcal{Q}).$$

Here c_1 is the universal constant in Theorem 4.1. To isolate the test average, substitute $P_n = (m/n)P_{\mathcal{M}} + (u/n)P_{\mathcal{U}}$ into the preceding inequality:

$$\left(1 - \frac{u}{2n}\right) P_{\mathcal{U}} h \overset{⑦}{\leq} \left(1 + \frac{m}{2n}\right) P_{\mathcal{M}} h + 2c_1 r^* + 3c_1 x/m,$$
$$P_{\mathcal{U}} h \overset{⑧}{\leq} 2P_{\mathcal{M}} h + 4c_1 r^* + 6c_1 x/m.$$

Step ⑦ collects the test terms. For ⑧, the denominator is $1 - u/(2n) \geq 1/2$ and $(1 + m/(2n))/(1 - u/(2n)) = (2 + m/n)/(1 + m/n) \leq 2$. The training average is nonnegative, so enlarging its coefficient preserves the inequality. For a nonzero $d \in \mathcal{H}_K$, substitute $v = d/\|d\|_{\mathcal{H}_K}$ and multiply by $\kappa^2 \|d\|^2_{\mathcal{H}_K}$. For $d = 0$ there is nothing to divide. We obtain, simultaneously for every $d \in \mathcal{H}_K$,

$$P_{\mathcal{U}} d^2 \leq 2P_{\mathcal{M}} d^2 + \kappa^2 (4c_1 r^* + 6c_1 x/m) \|d\|^2_{\mathcal{H}_K}. \tag{D.33}$$

The class uses only the full input sample. Uniformity in d permits $d = \widehat{f}_{m,\lambda} - f^*$ in (D.33), without independence from the split or labels. This is a jointly measurable event: continuity in d and separability reduce the inequality to a fixed countable dense subset of $\mathcal{H}_K$. Also, r^* is measurable, since the empirical eigenvalues are measurable and $r^* = \inf\{r \in \mathbb{Q} : r > 0,\ \psi(r) \leq r\}$.

Step 3: Fixed point and effective dimension. We now show that the norm term in (D.33) has the same scale as the ridge. When $r^* > 0$, its fixed-point equation in (D.32) gives

$$(r^*)^2 \overset{⑨}{=} \frac{512}{m} \sum_{j=1}^{n} \min\{r^*, \widehat{\nu}_j/\kappa^2\} \overset{①}{\leq} \frac{512Q}{m} r^* + \frac{512}{\kappa^2 m} \sum_{j>Q} \widehat{\nu}_j,$$
$$r^* \overset{②}{\leq} \frac{512Q}{m} + \sqrt{\frac{512}{\kappa^2 m} \sum_{j>Q} \widehat{\nu}_j}.$$

Step ⑨ squares the fixed-point identity, ① separates the head and tail, and ② uses $r^2 \leq ar + b \Rightarrow r \leq a + \sqrt{b}$ for nonnegative a, b, r, as in Section D.2.3. The bound also covers $r^* = 0$.

Put $A_\beta = a_+/(2\beta - 1)$. The choices of Q, x, λ imply

$$t_Q \overset{③}{\leq} A_\beta Q^{1-2\beta} \overset{④}{\leq} A_\beta m\lambda^2, \qquad Q/m \leq 2\lambda, \qquad \sqrt{\lambda/n} \leq \lambda.$$

Here ③ is Step 1. Step ④ uses $Q \geq x$, $1 - 2\beta < 0$, and $x^{1-2\beta} = m\lambda^2$. The last inequality follows from $n\lambda \geq m\lambda = x \geq 1$. On the full-sample event in (D.31), substitution into the fixed-point bound gives

$$r^* \overset{⑤}{\leq} 1024\lambda + \sqrt{\frac{512}{\kappa^2 m}\left(2A_\beta m\lambda^2 + \frac{\kappa^2 x}{n}\right)} \overset{⑥}{\leq} \left(1024 + \frac{32\sqrt{A_\beta}}{\kappa} + 16\sqrt{2}\right)\lambda.$$

Step ⑤ uses $Q/m \leq 2\lambda$ and the spectral-tail event. Step ⑥ uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, $x = m\lambda$, and $\sqrt{\lambda/n} \leq \lambda$. This displays the dependence on β, a_+, κ and proves $r^* \leq C\lambda$. Since $x/m = \lambda$, the uniform comparison (D.33) becomes

$$P_{\mathcal{U}} d^2 \leq C\left(P_{\mathcal{M}} d^2 + \lambda \|d\|^2_{\mathcal{H}_K}\right) \qquad (d \in \mathcal{H}_K). \tag{D.34}$$

The noise bound below uses the empirical effective dimension, defined by $\widehat{\mathcal{N}}_{\mathcal{M}}(\lambda) = \sum_{j \geq 1} \widehat{\nu}_{\mathcal{M},j}/(\widehat{\nu}_{\mathcal{M},j} + \lambda)$. All its denominators are positive, including in singular Gram matrices. On the training event in (D.31),

$$\widehat{\mathcal{N}}_{\mathcal{M}}(\lambda) \overset{⑦}{\leq} Q + \lambda^{-1} \sum_{j>Q} \widehat{\nu}_{\mathcal{M},j} \overset{⑧}{\leq} Q + 2t_Q/\lambda + \kappa^2 x/(m\lambda) \overset{⑨}{\leq} (2 + 2A_\beta + \kappa^2) m\lambda.$$

Step ⑦ bounds each head ratio by one and each tail ratio by its numerator divided by λ. Step ⑧ uses the training spectral-tail event. Step ⑨ uses $Q \leq 2m\lambda$, $t_Q \leq A_\beta m\lambda^2$, and $m\lambda \geq 1$.

Step 4: Ridge normal equation and noise. We control the two terms on the right of (D.34) together, rather than bounding the RKHS norm of the subset solution by the observed maximum label. The

projection argument in Section D.1 applies also to $\mathcal{H}_K$: projection onto the training-section span preserves training predictions and decreases the ridge penalty. Hence the same unique subset solution minimizes over $\mathcal{H}_K$ and $\mathcal{H}_{\mathbf{x}_n}$. Its coefficients are $(\mathbf{K}_\mathcal{M} + m\lambda \mathbf{I}_m)^{-1}\mathbf{y}_\mathcal{M}$, where $\mathbf{I}_m$ is the identity matrix and $\mathbf{y}_\mathcal{M} = (y_i)_{i\in\mathcal{M}}$. This formula proves measurability and includes singular or zero training Gram matrices. In the zero case the subset solution is zero.

Let $d = \widehat{f}_{m,\lambda} - f^*$, write $\varepsilon(i) = \varepsilon_i$ for the noise function on $[n]$, and set $\zeta_\mathcal{M} = m^{-1}\sum_{i\in\mathcal{M}} \varepsilon_i K_{\mathbf{x}_i}$. The directional calculation in Step 1 of Lemma D.1 applies here on $\mathcal{H}_K$. Optimality along both signs of an arbitrary $h \in \mathcal{H}_K$ gives $P_\mathcal{M}[(\widehat{f}_{m,\lambda} - y)h] + \lambda \left\langle \widehat{f}_{m,\lambda}, h \right\rangle_{\mathcal{H}_K} = 0$. Substituting $y = f^* + \varepsilon$ and $\widehat{f}_{m,\lambda} = f^* + d$ yields, for every $h \in \mathcal{H}_K$,

$$0 \overset{①}{=} P_\mathcal{M}[(d-\varepsilon)h] + \lambda \langle f^* + d, h\rangle_{\mathcal{H}_K} \overset{②}{=} \left\langle (\widehat{T}_\mathcal{M} + \lambda I)d - \zeta_\mathcal{M} + \lambda f^*, h \right\rangle_{\mathcal{H}_K}.$$

Step ① substitutes the model into the directional optimality condition. Step ② uses the definition of the empirical integral operator and the reproducing property in the definition of $\zeta_\mathcal{M}$. An RKHS vector orthogonal to every h is zero, so

$$(\widehat{T}_\mathcal{M} + \lambda I)d = \zeta_\mathcal{M} - \lambda f^*. \tag{D.35}$$

This identity includes components of f^* outside the sample span. Define $Z = \left\|(\widehat{T}_\mathcal{M} + \lambda I)^{-1/2}\zeta_\mathcal{M}\right\|^2_{\mathcal{H}_K}$. The operator $\widehat{T}_\mathcal{M} + \lambda I$ is bounded below by λI, so its inverse square root is a bounded operator. Apply it to (D.35) and take squared RKHS norms to obtain

$$\begin{aligned} P_\mathcal{M} d^2 + \lambda \|d\|^2_{\mathcal{H}_K} &\overset{③}{=} \left\|(\widehat{T}_\mathcal{M} + \lambda I)^{1/2} d\right\|^2_{\mathcal{H}_K} \\ &\overset{④}{\le} 2Z + 2\lambda^2 \left\langle f^*, (\widehat{T}_\mathcal{M} + \lambda I)^{-1} f^* \right\rangle_{\mathcal{H}_K} \overset{⑤}{\le} 2Z + 2\lambda \|f^*\|^2_{\mathcal{H}_K}. \end{aligned} \tag{D.36}$$

Step ③ expands the quadratic form of $\widehat{T}_\mathcal{M}$. Step ④ uses (D.35) and $\|a-b\|^2_{\mathcal{H}_K} \le 2\|a\|^2_{\mathcal{H}_K} + 2\|b\|^2_{\mathcal{H}_K}$. Step ⑤ uses $(\widehat{T}_\mathcal{M} + \lambda I)^{-1} \preceq \lambda^{-1} I$.

We next express Z as a finite-dimensional noise quadratic form. Extend the evaluation map to $T_\mathcal{M} : \mathcal{H}_K \to \mathbb{R}^m$ by the same formula $T_\mathcal{M} h = (h(\mathbf{x}_i))_{i\in\mathcal{M}}$, with indices ordered increasingly. Its adjoint is $T^*_\mathcal{M} \boldsymbol{a} = \sum_{i\in\mathcal{M}} a_i K_{\mathbf{x}_i}$, where a_i denotes the entry indexed by i. Consequently $T_\mathcal{M} T^*_\mathcal{M} = \mathbf{K}_\mathcal{M}$ and $\widehat{T}_\mathcal{M} = T^*_\mathcal{M} T_\mathcal{M}/m$. To keep the normalization explicit, observe that

$$\begin{aligned} (\widehat{T}_\mathcal{M} + \lambda I) T^*_\mathcal{M} &\overset{⑥}{=} T^*_\mathcal{M}(\mathbf{K}_\mathcal{M}/m + \lambda \mathbf{I}_m), \\ T_\mathcal{M}(\widehat{T}_\mathcal{M} + \lambda I)^{-1} T^*_\mathcal{M} &\overset{⑦}{=} \mathbf{K}_\mathcal{M}(\mathbf{K}_\mathcal{M}/m + \lambda \mathbf{I}_m)^{-1} = m\mathbf{K}_\mathcal{M}(\mathbf{K}_\mathcal{M} + m\lambda \mathbf{I}_m)^{-1}. \end{aligned}$$

Step ⑥ substitutes the two adjoint-product identities. Step ⑦ multiplies the first identity by the inverses on the appropriate sides, then by $T_\mathcal{M}$ on the left. Both inverses exist because $\lambda > 0$. Define $\mathbf{S}_\mathcal{M} = \mathbf{K}_\mathcal{M}(\mathbf{K}_\mathcal{M} + m\lambda \mathbf{I}_m)^{-1}$ and $\boldsymbol{\varepsilon}_\mathcal{M} = (\varepsilon_i)_{i\in\mathcal{M}}$. Since $\zeta_\mathcal{M} = T^*_\mathcal{M}\boldsymbol{\varepsilon}_\mathcal{M}/m$, we obtain

$$Z \overset{⑧}{=} m^{-1}\boldsymbol{\varepsilon}_\mathcal{M}^\top \mathbf{S}_\mathcal{M} \boldsymbol{\varepsilon}_\mathcal{M}, \qquad \operatorname{tr}(\mathbf{S}_\mathcal{M}) = \widehat{\mathcal{N}}_\mathcal{M}(\lambda), \qquad \mathbf{0} \preceq \mathbf{S}_\mathcal{M} \preceq \mathbf{I}_m.$$

Step ⑧ substitutes the preceding evaluation identity. The eigenvalues of $\mathbf{S}_\mathcal{M}$ are $\widehat{\nu}_{\mathcal{M},j}/(\widehat{\nu}_{\mathcal{M},j} + \lambda) \in [0,1]$, proving the trace and order assertions.

Conditional on all inputs and the split, the vector $\boldsymbol{\varepsilon}_\mathcal{M}$ satisfies $\mathbb{E}\left[\exp(\boldsymbol{a}^\top \boldsymbol{\varepsilon}_\mathcal{M}) \mid \mathbf{X}_n, \mathcal{M}\right] \le \exp(\sigma_0^2 \|\boldsymbol{a}\|_2^2/2)$ for every $\boldsymbol{a} \in \mathbb{R}^m$. This follows by multiplying the independent one-dimensional bounds in (4.7). Apply Hsu et al. (2012, Lemma 30) to the positive semidefinite matrix $\mathbf{S}_\mathcal{M}$. With conditional probability at least $1 - e^{-x}$ over the training noise,

$$Z \overset{⑨}{\le} \frac{\sigma_0^2}{m}\left[\operatorname{tr}(\mathbf{S}_\mathcal{M}) + 2\sqrt{x \operatorname{tr}(\mathbf{S}^2_\mathcal{M})} + 2x\|\mathbf{S}_\mathcal{M}\|_{\mathrm{OP}}\right] \overset{①}{\le} \frac{\sigma_0^2}{m}\left[2\widehat{\mathcal{N}}_\mathcal{M}(\lambda) + 3x\right].$$

Step ⑨ is the cited inequality. Step ① uses $\operatorname{tr}(\mathbf{S}^2_\mathcal{M}) \le \operatorname{tr}(\mathbf{S}_\mathcal{M})$, $\|\mathbf{S}_\mathcal{M}\|_{\mathrm{OP}} \le 1$, and $2\sqrt{ab} \le a + b$. For $\sigma_0 = 0$, the noise has variance zero and mean zero, so it vanishes almost surely and the same conclusion holds with $Z = 0$. Combining Step 3, $x = m\lambda$, and (D.36) gives

$$P_\mathcal{M} d^2 + \lambda \|d\|^2_{\mathcal{H}_K} \le C\lambda\left(\|f^*\|^2_{\mathcal{H}_K} + \sigma_0^2\right). \tag{D.37}$$

Step 5: Joint probability. Let G be the intersection of the two spectral-tail events in Step 1, the UTLC event in Step 2, and the conditional noise event in Step 4. The first two events together fail with probability at most $2e^{-x}$ over the inputs and random split. The UTLC event fails with conditional probability at most e^{-x} given all inputs. The noise event fails with conditional probability at most e^{-x} given the inputs and split. The tower property and the union bound therefore give $\Pr_{f^*}(G^c) \le 4e^{-x}$ over the inputs, random split, and label noise. These events need not be independent. Test noises are not used in the argument and can be integrated out. On G, substitute $d = \widehat{f}_{m,\lambda} - f^*$ into (D.34) and then apply (D.37). This proves (4.8). All constants are independent of u and of the chosen signal. No bound on the diameter or residual maximum of zero-ridge subset solutions has been assumed.

Step 6: Expected error. We control the exceptional event explicitly because a high-probability bound alone does not imply the claimed expectation bound. Put $\mu_* = \|f^*\|_{\mathcal{H}_K}$ and $W = \mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\lambda}, f^*)$. Comparison of $J_{\mathcal{M}}(\widehat{f}_{m,\lambda})$ with $J_{\mathcal{M}}(0)$ and the reproducing inequality give, for every realization,

$$\lambda \left\| \widehat{f}_{m,\lambda} \right\|_{\mathcal{H}_K}^2 \overset{②}{\le} P_{\mathcal{M}} y^2, \qquad W \overset{③}{\le} 2\kappa^2 \left(\lambda^{-1} P_{\mathcal{M}} y^2 + \mu_*^2 \right).$$

Here ② drops the nonnegative training residual, and ③ uses $|\widehat{f}_{m,\lambda}(\mathbf{x}) - f^*(\mathbf{x})|^2 \le 2\kappa^2 (\left\| \widehat{f}_{m,\lambda} \right\|_{\mathcal{H}_K}^2 + \mu_*^2)$. For $\sigma_0 > 0$, the sub-Gaussian moment bound gives $\Pr\{|\varepsilon_i| > t\} \le 2e^{-t^2/(2\sigma_0^2)}$ by applying Markov's inequality to both signs and minimizing in the exponential parameter. Integration yields $\mathbb{E}\left[\varepsilon_i^4\right] \le 8 \int_0^\infty t^3 e^{-t^2/(2\sigma_0^2)}\, dt = 16\sigma_0^4$. The same inequality holds for zero noise. Since $|f^*(\mathbf{x})| \le \kappa \mu_*$,

$$\mathbb{E}_{f^*}\left[(P_{\mathcal{M}} y^2)^2\right] \overset{④}{\le} \mathbb{E}_{f^*}\left[P_{\mathcal{M}} y^4\right] \overset{⑤}{\le} 8\kappa^4 \mu_*^4 + 128\sigma_0^4, \qquad \mathbb{E}_{f^*}\left[W^2\right] \overset{⑥}{\le} C\lambda^{-2}(\mu_*^2 + \sigma_0^2)^2.$$

Step ④ is Jensen's inequality for the finite average. Step ⑤ uses $|a+b|^4 \le 8(|a|^4 + |b|^4)$. Step ⑥ squares the deterministic bound on W, uses $(a+b)^2 \le 2a^2 + 2b^2$, and uses $0 < \lambda \le 1$. Cauchy–Schwarz and Step 5 now imply

$$\mathbb{E}_{f^*}\left[W \mathbb{I}_{\{G^c\}}\right] \overset{⑦}{\le} \sqrt{\mathbb{E}_{f^*}\left[W^2\right] \Pr_{f^*}(G^c)} \overset{⑧}{\le} C(\mu_*^2 + \sigma_0^2)\lambda^{-1} e^{-x/2} \overset{⑨}{\le} C(\mu_*^2 + \sigma_0^2)\lambda.$$

Step ⑦ is Cauchy–Schwarz, and ⑧ uses the second moment and failure probability. For ⑨, note that $\lambda^{-2} e^{-x/2} = x^{4\beta} e^{-x/2}$, whose supremum on $[1, \infty)$ is finite and depends only on β. On G, Step 5 bounds W by $C(\mu_*^2 + \sigma_0^2)\lambda$. Adding the two expectations proves (4.9) for all $u \ge m \ge 2$, after enlarging C.

Step 7: Critical population rate and an example. We finally explain why the selected ridge has the critical kernel order under the two-sided EDR used for minimax comparison. Define, for $r \ge 0$,

$$\mathcal{R}_m(r) = \sqrt{\frac{1}{m} \sum_{j \ge 1} \min\{r, \nu_j\}}, \qquad r_{\rho,m} = \inf\{r > 0 : \mathcal{R}_m(r) \le r\}.$$

The spectral sum is finite because $\sum_j \nu_j \le \kappa^2$. For small $r > 0$, put $k_+ = \lceil (a_+/r)^{1/(2\beta)} \rceil$ and $k_- = \lfloor (a_-/r)^{1/(2\beta)} \rfloor$. The upper bound separates the first k_+ directions from the summable tail. The lower bound retains the first k_- directions, each of which contributes r. For $r \le a_+$ and $r \le a_-/2^{2\beta}$, respectively,

$$\sum_{j \ge 1} \min\{r, \nu_j\} \overset{①}{\le} k_+ r + \frac{a_+}{2\beta - 1} k_+^{1-2\beta} \overset{②}{\le} \left(2 + \frac{1}{2\beta - 1}\right) a_+^{1/(2\beta)} r^{1 - 1/(2\beta)},$$

$$\sum_{j \ge 1} \min\{r, \nu_j\} \overset{③}{\ge} k_- r \overset{④}{\ge} \frac{1}{2} a_-^{1/(2\beta)} r^{1 - 1/(2\beta)}.$$

Step ① uses the integral tail bound from Step 1. Step ② uses $(a_+/r)^{1/(2\beta)} \le k_+ \le 2(a_+/r)^{1/(2\beta)}$ and $1 - 2\beta < 0$. Step ③ uses $\nu_j \ge r$ for $j \le k_-$. Step ④ uses $k_- \ge \frac{1}{2}(a_-/r)^{1/(2\beta)}$. Taking square roots after division by m gives constants $0 < b_- \le b_+ < \infty$, depending only on β, a_-, a_+, such that

$$b_- m^{-1/2} r^{-1/(2\alpha)} \overset{⑤}{\le} \frac{\mathcal{R}_m(r)}{r} \overset{⑥}{\le} b_+ m^{-1/2} r^{-1/(2\alpha)}$$

for all sufficiently small $r > 0$. Steps ⑤ and ⑥ divide the preceding spectral-sum bounds by mr^2 and take square roots, using $1/(2\alpha) = 1/2 + 1/(4\beta)$. Each summand $\min\{r, \nu_j\}/r^2$ is nonincreasing in $r > 0$. Hence $\mathcal{R}_m(r)/r$ is nonincreasing. It tends to infinity at zero by the lower bound, and to zero at infinity because $\mathcal{R}_m(r) \le \kappa/\sqrt{m}$. Continuity follows from dominated convergence with dominating summands ν_j.

To locate the crossing of $\mathcal{R}_m(r)/r$ with one, substitute $r = tm^{-\alpha}$ for a fixed $t > 0$. The preceding bounds become $b_- t^{-1/(2\alpha)} \le \mathcal{R}_m(r)/r \le b_+ t^{-1/(2\alpha)}$. For all sufficiently large m, both radii corresponding to $t = (b_-/2)^{2\alpha}$ and $t = (2b_+)^{2\alpha}$ are in the range of these inequalities. The ratio is at least two at the first radius and at most one half at the second. Monotonicity and the definition of $r_{\rho,m}$ now give

$$(b_-/2)^{2\alpha} m^{-\alpha} \overset{⑦}{\le} r_{\rho,m} \overset{⑧}{\le} (2b_+)^{2\alpha} m^{-\alpha}.$$

Step ⑦ excludes every radius below the first crossing bracket, and ⑧ exhibits a radius at which $\mathcal{R}_m(r) \le r$. Consequently $r_{\rho,m} \asymp \lambda_m$. The empirical fixed-point estimate in Step 3 is the upper comparison needed for the proof. No coordinatewise comparison of empirical and population eigenvalues is required.

A concrete example confirms that the common assumptions are attainable. Let ρ be uniform on $[0, 1]$, and use its orthonormal Fourier functions $\phi_1(t) = 1$, $\phi_{2k}(t) = \sqrt{2}\cos(2\pi kt)$, and $\phi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt)$ for $k \ge 1$. For the same fixed $\beta > 1/2$, set

$$K(t, s) = \sum_{j \ge 1} j^{-2\beta} \phi_j(t) \phi_j(s).$$

The series converges absolutely and uniformly since $|\phi_j(t)\phi_j(s)| \le 2$ and $\sum_j j^{-2\beta} < \infty$. Every finite partial sum is positive semidefinite, so its limit is a continuous positive semidefinite kernel. We verify its RKHS and eigenvalues of the integral operator directly to check both EDR assumptions for this same model. Consider functions $f = \sum_j b_j \phi_j$ with $\sum_j j^{2\beta} b_j^2 < \infty$. For any integer $J \ge 1$, Cauchy–Schwarz gives

$$\sup_{t \in [0,1]} \left| \sum_{j > J} b_j \phi_j(t) \right| \overset{⑨}{\le} \left(\sum_{j > J} j^{2\beta} b_j^2 \right)^{1/2} \left(2 \sum_{j > J} j^{-2\beta} \right)^{1/2} \longrightarrow 0.$$

Step ⑨ inserts the weights j^β and $j^{-\beta}$ and uses $|\phi_j(t)| \le \sqrt{2}$. The series therefore defines a continuous function by uniform convergence. Its coefficients are unique because $b_j = \int_0^1 f(t)\phi_j(t)\, dt$. Give these functions the inner product $\langle f, g \rangle = \sum_j j^{2\beta} b_j c_j$ when $g = \sum_j c_j \phi_j$. The coefficient map $f \mapsto (j^\beta b_j)_{j \ge 1}$ is an isometry onto the space of square-summable sequences, so this function space is a separable Hilbert space.

For fixed $s \in [0, 1]$, the coefficients of $K(\cdot, s)$ are $j^{-2\beta}\phi_j(s)$, whose squared weighted sum is $\sum_j j^{-2\beta}\phi_j(s)^2 < \infty$. Hence this section belongs to the function space, and

$$\langle f, K(\cdot, s) \rangle \overset{①}{=} \sum_j j^{2\beta} b_j j^{-2\beta} \phi_j(s) \overset{②}{=} f(s).$$

Step ① uses the weighted inner product, and ② uses the defining series. This proves the reproducing property and identifies the space with $\mathcal{H}_K$, with $\|f\|_{\mathcal{H}_K}^2 = \sum_j j^{2\beta} b_j^2$. The functions $e_j = j^{-\beta}\phi_j$ form a complete orthonormal RKHS basis. The integral identity from Step 1 gives

$$\langle e_k, T_\rho e_j \rangle_{\mathcal{H}_K} \overset{③}{=} \int_0^1 e_k(t) e_j(t)\, dt \overset{④}{=} j^{-2\beta} \mathbb{I}_{\{k=j\}}.$$

Step ③ is the integral identity, and ④ uses Fourier orthonormality. Completeness of (e_k) therefore gives $T_\rho e_j = j^{-2\beta} e_j$ for every j. Also $K(t, t) \le 2\sum_j j^{-2\beta}$ uniformly in t. Hence this family satisfies both EDR inequalities and the bounded-kernel assumption. Under Gaussian noise with any fixed positive variance and any fixed signal radius, Theorems 4.4 and D.5 give matching high-probability rates. Since $K(t, t) \ge 1$, every nonempty sample span contains a nonzero kernel section.

Scaling it gives arbitrarily large squared loss at its own sample input. The proof normalizes prediction differences for analysis, while the ridge learner still optimizes over the unbounded set $\mathcal{H}_{\mathbf{X}_n}$.

Step 8: Data-based ridge. We now replace the population exponent in the learner by an empirical spectral cutoff. Throughout this step, $x \geq 1$ is deterministic and $\widehat{\nu}_j$ are the eigenvalues of $\mathbf{K}/n$ already defined in Step 1. Define

$$\Delta_m = \min_{Q \in \{0,\dots,n\}} \left[\frac{Q}{m} + \sqrt{\frac{1}{m} \sum_{j>Q} \widehat{\nu}_j} \right], \qquad \widehat{\lambda}_m(x) = \frac{x}{m} + \Delta_m. \tag{D.38}$$

The choice in (4.10) is $\widehat{\lambda}_m(8 \log m)$. The eigenvalues and the finite minimum are measurable functions of the observed inputs. The ridge is positive even when $\mathbf{K} = \mathbf{0}$. In this step C_κ denotes a finite positive constant depending only on κ, allowed to increase. We first prove a bound conditional on the inputs, then use population EDR only to evaluate its rate.

Empirical fixed point. To relate the cutoff to kernel complexity precisely, define $\widehat{\mathcal{R}}_m(r) = \sqrt{m^{-1} \sum_{j=1}^n \min\{r, \widehat{\nu}_j\}}$. Let $\widehat{r}_m$ be its positive fixed point when $\mathbf{K} \neq \mathbf{0}$, and set $\widehat{r}_m = 0$ otherwise. The sub-root argument in Step 2 proves existence and finiteness. For any cutoff Q, its fixed-point identity gives

$$\widehat{r}_m^2 \overset{⑤}{\leq} \frac{Q}{m} \widehat{r}_m + \frac{1}{m} \sum_{j>Q} \widehat{\nu}_j, \qquad \widehat{r}_m \overset{⑥}{\leq} \Delta_m.$$

Step ⑤ separates the head and tail. Step ⑥ uses $r^2 \leq ar + b \Rightarrow r \leq a + \sqrt{b}$ and minimizes over Q. For the reverse comparison when $\widehat{r}_m > 0$, choose $q = |\{j : \widehat{\nu}_j > \widehat{r}_m\}|$. The fixed-point identity becomes $q\widehat{r}_m + \sum_{j>q} \widehat{\nu}_j = m\widehat{r}_m^2$. Consequently

$$\Delta_m \overset{⑦}{\leq} \frac{q}{m} + \sqrt{\frac{1}{m} \sum_{j>q} \widehat{\nu}_j} \overset{⑧}{\leq} 2\widehat{r}_m.$$

Step ⑦ evaluates the minimum at q. For ⑧, each nonnegative term on the left of the fixed-point identity is at most $m\widehat{r}_m^2$, so $q/m \leq \widehat{r}_m$ and the square root is at most $\widehat{r}_m$. If $\mathbf{K} = \mathbf{0}$, both Δ_m and $\widehat{r}_m$ vanish. The cutoff is therefore comparable to the empirical fixed point, with the normalization determined by m, not by n.

Conditional control at the selected cutoff. Fix all inputs, let $\widehat{Q}$ be the smallest minimizer in (D.38), and put $\widehat{t} = \sum_{j>\widehat{Q}} \widehat{\nu}_j$ and $\lambda = \widehat{\lambda}_m(x)$. These quantities are fixed conditional on the inputs, so the split remains uniform and the noise law is unchanged. We can therefore apply UTLC at this ridge without a union over its possible values. The definitions give

$$\widehat{Q} \leq m\lambda, \qquad \widehat{t} \leq m\lambda^2, \qquad x \leq m\lambda.$$

The fixed-point estimate in Step 3, applied with $Q = \widehat{Q}$, yields

$$r^* \overset{⑨}{\leq} \frac{512\widehat{Q}}{m} + \sqrt{\frac{512\widehat{t}}{\kappa^2 m}} \overset{①}{\leq} \left(512 + \frac{16\sqrt{2}}{\kappa} \right) \lambda.$$

Step ⑨ is that estimate for an arbitrary cutoff, and ① uses the preceding inequalities. Here r^* is the fixed point of the two-level majorant (D.32), whereas $\widehat{r}_m$ above is the unscaled kernel fixed point. Applying Step 2 at the present confidence x therefore gives $P_{\mathcal{U}} d^2 \leq C_\kappa (P_{\mathcal{M}} d^2 + \lambda \|d\|_{\mathcal{H}_K}^2)$ for every $d \in \mathcal{H}_K$, with conditional probability at least $1 - e^{-x}$ over the random split.

To control the training noise, we also need the training spectrum at this same ridge. Choose a projection $\widehat{\Pi}$ onto eigenvectors for the $\widehat{Q}$ largest positive eigenvalues of $\widehat{T}_n$, taking all positive eigendirections if their number is smaller. For $\widehat{Q} = 0$, take the zero projection. Put $z_i = \left\| (I - \widehat{\Pi}) K_{\mathbf{x}_i} \right\|_{\mathcal{H}_K}^2$. Then $0 \leq z_i \leq \kappa^2$ and $n^{-1} \sum_i z_i = \widehat{t}$ by the trace identities in Step 1. The same projection-trace comparison gives $\sum_{j>\widehat{Q}} \widehat{\nu}_{\mathcal{M},j} \leq m^{-1} \sum_{i \in \mathcal{M}} z_i$.

Although training indices are sampled without replacement, their exponential moment is no larger than that for independent uniform indices. The convex sampling comparison (Hoeffding, 1963, Theorem 4), also stated in Yang (2025a, Theorem A.6), gives

$$\mathbb{E}_{\mathcal{M}}\left[\exp\left(\kappa^{-2}\sum_{i\in\mathcal{M}} z_i\right) \mid \mathbf{X}_n\right] \overset{②}{\le} \left(\frac{1}{n}\sum_{i=1}^{n} e^{z_i/\kappa^2}\right)^m \overset{③}{\le} \exp\left(\frac{(e-1)m\widehat{t}}{\kappa^2}\right).$$

Step ② applies the comparison to the convex exponential of a sum of m scalars from the fixed population $(z_i)_{i=1}^n$. Step ③ uses $e^v \le 1+(e-1)v$ on $[0,1]$, then $1+w \le e^w$. Markov's inequality, as in Step 1, now implies

$$\Pr\left\{\sum_{j>\widehat{Q}} \widehat{\nu}_{\mathcal{M},j} > 2\widehat{t} + \kappa^2 x/m \mid \mathbf{X}_n\right\} \overset{④}{\le} e^{-x}.$$

For ④, multiply the exponential-moment bound by $\exp(-2m\widehat{t}/\kappa^2 - x)$ and use $e-1<2$. The displayed event depends only on matrix eigenvalues and is measurable. The auxiliary projection is chosen only after fixing the inputs, so no jointly measurable choice of eigenvectors is required. On its complement the effective dimension from Step 3 satisfies

$$\widehat{\mathcal{N}}_{\mathcal{M}}(\lambda) \overset{⑤}{\le} \widehat{Q} + \frac{2\widehat{t}}{\lambda} + \frac{\kappa^2 x}{m\lambda} \overset{⑥}{\le} 3m\lambda + \kappa^2 \overset{⑦}{\le} (3+\kappa^2)m\lambda.$$

Step ⑤ separates the head and tail of the effective dimension and uses the preceding tail event. Step ⑥ uses $\widehat{Q} \le m\lambda$, $\widehat{t} \le m\lambda^2$, and $x \le m\lambda$. Step ⑦ uses $m\lambda \ge x \ge 1$.

Ridge error and conditional probability. Conditional on the inputs and split, this positive ridge is fixed and independent of the noises. Every identity and conditional noise bound in Step 4 therefore applies with $\lambda = \widehat{\lambda}_m(x)$. In particular, with an additional conditional failure probability at most e^{-x},

$$Z \overset{⑧}{\le} \frac{\sigma_0^2}{m}\left[2\widehat{\mathcal{N}}_{\mathcal{M}}(\lambda) + 3x\right] \overset{⑨}{\le} (9+2\kappa^2)\sigma_0^2\lambda,$$

$$P_{\mathcal{M}}d^2 + \lambda\|d\|_{\mathcal{H}_K}^2 \overset{①}{\le} 2Z + 2\lambda\|f^*\|_{\mathcal{H}_K}^2 \overset{②}{\le} C_\kappa\lambda\left(\|f^*\|_{\mathcal{H}_K}^2 + \sigma_0^2\right), \qquad d = \widehat{f}_{m,\lambda} - f^*.$$

Step ⑧ is the noise inequality in Step 4, ⑨ substitutes the effective-dimension bound and $x \le m\lambda$, ① is (D.36), and ② collects terms. Combining this with the uniform UTLC comparison proves

$$\mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\widehat{\lambda}_m(x)}, f^*) \le C_\kappa\left(\|f^*\|_{\mathcal{H}_K}^2 + \sigma_0^2\right)\widehat{\lambda}_m(x) \tag{D.39}$$

with conditional probability at least $1-3e^{-x}$ over the random split and label noise, given all inputs. The three events are the UTLC event, the training spectral-tail event, and the noise event. A conditional union bound suffices, without independence between them or a union over cutoffs. This argument includes zero and singular Gram matrices.

Population rate and expectation. The algorithm above never uses β. For its analysis only, put $r_m = m^{-\alpha}$ and $Q = \lceil m^{1/(2\beta+1)}\rceil$. The full-sample argument in Step 1 is valid at any deterministic $x \ge 1$. With probability at least $1-e^{-x}$ over the inputs, it gives $\sum_{j>Q}\widehat{\nu}_j \le 2t_Q + \kappa^2 x/n$. Recall from Step 3 that $Q/m \le 2r_m$ and $t_Q \le A_\beta m r_m^2$, where $A_\beta = a_+/(2\beta-1)$. Evaluating the empirical minimum at this deterministic cutoff yields

$$\widehat{\lambda}_m(x) \overset{③}{\le} \frac{x}{m} + \frac{Q}{m} + \sqrt{\frac{2t_Q}{m} + \frac{\kappa^2 x}{mn}} \overset{④}{\le} (2+\sqrt{2A_\beta})r_m + \frac{x}{m} + \kappa\sqrt{\frac{x}{mn}} \overset{⑤}{\le} C\left(r_m + \frac{x}{m}\right).$$

Step ③ uses the spectral-tail event, ④ uses $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, and ⑤ uses $n \ge m$ and $x \ge 1$. Here and below C depends only on β, a_+, κ. The tower property combines this event with (D.39). As a result, for each deterministic $x \ge 1$, with probability at least $1-4e^{-x}$ over the inputs, random split, and label noise,

$$\mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\widehat{\lambda}_m(x)}, f^*) \le C\left(\|f^*\|_{\mathcal{H}_K}^2 + \sigma_0^2\right)\left(m^{-\alpha} + \frac{x}{m}\right). \tag{D.40}$$

This is a guarantee for the ridge recomputed at each chosen confidence, not a claim that one event covers every x simultaneously. Set $x = 8\log m \geq 1$. Since $\log m/m = o(m^{-\alpha})$ for each fixed $\beta > 1/2$, and $\sup_{m\geq 2}(\log m)m^{\alpha-1} < \infty$, (D.40) proves the asserted data-based version of (4.8), with failure probability $4m^{-8}$.

For the expected bound, write $W = \mathrm{MSE}_{\mathcal{U}}(\widehat{f}_{m,\widehat{\lambda}_m}, f^*)$ and $S = \|f^*\|^2_{\mathcal{H}_K} + \sigma_0^2$. Let G be the four-event intersection just used at $x = 8\log m$. The deterministic objective comparison in Step 6 remains valid for this random ridge. Since $\widehat{\lambda}_m \geq 1/m$, it gives $W \leq 2\kappa^2(mP_{\mathcal{M}}y^2 + \|f^*\|^2_{\mathcal{H}_K})$. The fourth-moment estimate in Step 6 then implies

$$\mathbb{E}_{f^*}\left[W^2\right] \overset{⑥}{\leq} C_\kappa m^2 S^2, \qquad \mathbb{E}_{f^*}\left[W\mathbb{I}_{\{G^c\}}\right] \overset{⑦}{\leq} \sqrt{\mathbb{E}_{f^*}\left[W^2\right]\Pr_{f^*}(G^c)} \overset{⑧}{\leq} C_\kappa S m^{-3}.$$

Step ⑥ squares the deterministic bound and uses $\mathbb{E}_{f^*}\left[(P_{\mathcal{M}}y^2)^2\right] \leq 8\kappa^4\|f^*\|^4_{\mathcal{H}_K} + 128\sigma_0^4$. Step ⑦ is Cauchy–Schwarz, and ⑧ uses $\Pr_{f^*}(G^c) \leq 4m^{-8}$. On G, the preceding bound is $W \leq CSm^{-\alpha}$. Adding the expectations and using $m^{-3} \leq m^{-\alpha}$ proves (4.9) for $\widehat{\lambda}_m$ as well. The same learner therefore attains the stated upper rate on every fixed signal ball, without using its radius or the population EDR exponent. Only an upper comparison with the population critical rate was needed. No empirical-to-population fixed-point equivalence is assumed.

□

D.4 A MINIMAX LOWER BOUND FOR NOISY TKL

The inductive lower exponent of Caponnetto & De Vito (2007, Theorem 2), with their parameters $b = 2\beta$ and $c = 1$, corresponds to two-sided EDR and a fixed RKHS signal radius. An upper eigenvalue bound alone does not exclude a finite-rank problem, and zero noise does not give the same statistical obstruction. We use these EDR conditions and a positive Gaussian noise variance below. Gaussian noise is a subclass of the independent sub-Gaussian model (4.7). We construct finitely many signals that are separated in population prediction error but difficult to distinguish from the training labels. A bounded kernel ensures that their separation persists on the test inputs when $u \geq m$. This gives a lower bound whose probability tends to one, while the Gaussian squared losses remain unbounded.

Recall the input distribution ρ, measurable bounded kernel K, separable RKHS $\mathcal{H}_K$, integral operator T_ρ, and eigenvalues ν_j from Section 4.4. For the signal ball $\mathcal{H}_K(\mu_0)$ defined in Section 4.4, we put $\mathcal{H}_{\mathbf{x}_n}(\mu_0) = \mathcal{H}_K(\mu_0) \cap \mathcal{H}_{\mathbf{x}_n}$. Write $\|h\|^2_{L^2(\rho)} = \int h(\mathbf{x})^2\, d\rho(\mathbf{x})$.

We fix $\mu_0, \sigma_0 > 0$, draw $\mathbf{x}_1, \ldots, \mathbf{x}_n$ independently from ρ, and let $y_i = f(\mathbf{x}_i) + \varepsilon_i$, where $f \in \mathcal{H}_K(\mu_0)$ and the ε_i are independent $N(0, \sigma_0^2)$ variables, independent of the inputs. Then a uniform random split on the full sample $\{\mathbf{x}_i\}_{i=1}^n$ generates the training index set $\mathcal{M}$ of size m and the test index set $\mathcal{U} = [n] \setminus \mathcal{M}$ with $u = n - m$. Recall that $\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f) = u^{-1}\sum_{i\in\mathcal{U}}(\widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2$. Theorem D.5 below gives the high-probability minimax lower bound for noisy TKL.

Theorem D.5 (High-Probability Minimax Lower Bound For Noisy TKL)**.** Suppose that, for a fixed $\beta > 1/2$,

$$\nu_j \asymp j^{-2\beta} \qquad (j \geq 1). \tag{D.41}$$

Let $a_- > 0$ denote a fixed lower comparison constant in (D.41), so $\nu_j \geq a_- j^{-2\beta}$ for every $j \geq 1$.

The infima below are over all measurable predictors $\widehat{f}$ using all n inputs and only the training labels. They may use independent randomization and know ρ, K, μ_0, σ_0. Their test predictions need not belong to an RKHS ball. Write $\Pr_f$ for probability over the random split, inputs, noises, and any learner randomization, with signal f.

There are constants $c_{\mathrm{low}}, c_{\mathrm{prob}} > 0$ and an integer $m_0 \geq 2$, depending only on $\beta, a_-, \kappa, \mu_0, \sigma_0$, such that for all integers $u \geq m \geq m_0$,

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{H}_K(\mu_0)} \Pr_f\left\{\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f) \geq c_{\mathrm{low}} m^{-2\beta/(2\beta+1)}\right\} \geq 1 - 2\exp\left(-c_{\mathrm{prob}} m^{1/(2\beta+1)}\right). \tag{D.42}$$

The same lower bound holds in the worst case over fixed input samples:

$$\inf_{\widehat{f}} \sup_{\mathbf{X}_n \in \mathcal{X}^n} \sup_{f \in \mathcal{H}_{\mathbf{X}_n}(\mu_0)} \Pr_{f,\mathbf{X}_n} \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f) \geq c_{\mathrm{low}} m^{-2\beta/(2\beta+1)} \right\} \geq 1 - 2\exp\left(-c_{\mathrm{prob}} m^{1/(2\beta+1)}\right), \tag{D.43}$$

where $\Pr_{f,\mathbf{X}_n}$ averages only over the random split, Gaussian noises, and learner randomization for the prescribed inputs and signal. For $0 < \delta < 1$, both lower-bound probabilities are at least $1 - \delta$ whenever $u \geq m$ and

$$m \geq \max\left\{ m_0, \left(c_{\mathrm{prob}}^{-1} \log(2/\delta)\right)^{2\beta+1} \right\}.$$

Both corresponding minimax expected errors are at least $(c_{\mathrm{low}}/2) m^{-2\beta/(2\beta+1)}$ for all sufficiently large m.

The population model in (D.42) is the model of Theorem 4.4 with two-sided EDR and positive Gaussian noise. Each learner has a signal chosen before the inputs and noises are drawn for which (D.42) holds. The additional statement (D.43) takes a worst case over fixed inputs. It is not a lower bound for every signal or every input configuration, and the random-input upper bound does not assert uniformity over those configurations.

Proof of Theorem D.5. Recall that ρ is the input distribution, μ_0 is the RKHS signal radius, and $\sigma_0^2 > 0$ is the Gaussian noise variance. The learner observes all inputs and the training labels, but no test labels. We first include the inputs in the probability space, then obtain the worst-case fixed-input conclusion by conditioning. The eigenvalue lower bound gives $\kappa > 0$.

Proof outline.

(1) Separated signals. Construct a finite subset of the RKHS ball with exponentially many members and a prescribed separation in $L^2(\rho)$.
(2) Probability of identifying a signal. Bound second moments of Gaussian likelihood ratios. Cauchy–Schwarz then makes the average probability of correct identification exponentially small.
(3) Separation on the test set. Apply a bounded-variable tail bound to every pair of signals. The chosen number of eigenfunctions makes $u \geq m$ sufficient to preserve all pairwise distances with high probability.
(4) Prediction error. Turn any accurate test predictor into a correct identification rule and combine the preceding two probability bounds.
(5) Fixed inputs and expected error. Project the signals onto the finite-sample span without changing their evaluations, then condition on the inputs and derive the stated consequences.

Step 1: Separated signals. We use a small fixed multiple of $m^{1/(2\beta+1)}$ eigenfunctions. This retains the desired error rate while allowing the test set to resolve the finite collection of alternatives. Define

$$s = \frac{1}{2\beta+1}, \qquad \eta = \min\left\{1, \left(\frac{a_-}{128\kappa^2}\right)^s\right\}, \qquad \gamma = \min\left\{a_- \mu_0^2, \frac{a_- \sigma_0^2}{\kappa^2}, \frac{\sigma_0^2}{16(e-1)}\right\}.$$

All three constants are positive, and $0 < \eta \leq 1$. We take $m_0 = \lceil (2/\eta)^{1/s} \rceil$, $c_{\mathrm{low}} = \gamma\eta/16$, and $c_{\mathrm{prob}} = \eta/64$. For $m \geq m_0$, put $q = \lfloor \eta m^s \rfloor$ and $a = \sqrt{\gamma/m}$. Since $\eta m^s \geq 2$, the floor and the choice of η imply

$$\frac{\eta m^s}{2} \leq q \leq \eta m^s \leq m^s, \qquad q^{2\beta+1} \overset{①}{\leq} \eta^{2\beta+1} m \overset{②}{\leq} \frac{a_- m}{128\kappa^2}.$$

Here ① raises $q \leq \eta m^s$ to power $2\beta + 1$, and ② uses the definition of η. In particular, $q \geq 2$.

Choose orthonormal RKHS eigenvectors e_j of T_ρ with eigenvalues ν_j, and put $\phi_j = e_j/\sqrt{\nu_j}$. The integral identity gives

$$\int \phi_j \phi_k \, d\rho \overset{③}{=} \frac{\langle e_j, T_\rho e_k \rangle_{\mathcal{H}_K}}{\sqrt{\nu_j \nu_k}} \overset{④}{=} \mathbb{1}_{\{j=k\}}, \qquad \|\phi_j\|_{\mathcal{H}_K}^2 = \nu_j^{-1}.$$

Step ③ uses the definition of T_ρ and the reproducing property, and ④ uses its orthonormal eigenvectors. For each $\omega \in \{-1,1\}^q$, define $f_\omega = a\sum_{j=1}^q \omega_j \phi_j$. The lower bound in (D.41) ensures

$$\|f_\omega\|_{\mathcal{H}_K}^2 \overset{⑤}{=} a^2 \sum_{j=1}^q \nu_j^{-1} \overset{⑥}{\le} \frac{\gamma q^{2\beta+1}}{m a_-} \overset{⑦}{\le} \frac{\gamma}{a_-} \overset{⑧}{\le} \mu_0^2,$$

$$\sup_{\mathbf{x}\in\mathcal{X}} f_\omega(\mathbf{x})^2 \overset{⑨}{\le} \kappa^2 \|f_\omega\|_{\mathcal{H}_K}^2 \overset{①}{\le} \frac{\kappa^2\gamma}{a_-} \overset{②}{\le} \sigma_0^2.$$

Here ⑤ uses RKHS orthogonality, ⑥ uses $\nu_j^{-1} \le j^{2\beta}/a_-$, ⑦ uses $q \le m^s$, and ⑧ uses $\gamma \le a_-\mu_0^2$. Step ⑨ is the reproducing inequality, ① uses the norm bound, and ② uses $\gamma \le a_-\sigma_0^2/\kappa^2$. Also, $\|f_\omega\|_{L^2(\rho)}^2 = a^2 q$ by the orthonormality of $(\phi_j)_{j=1}^q$.

We now select sufficiently many sign vectors whose signals are separated. Define the Hamming distance $d_H(\omega,\omega') = \sum_{j=1}^q \mathbb{1}_{\{\omega_j \ne \omega'_j\}}$. Starting from the entire sign cube, keep one vector and remove all vectors within Hamming distance $\lfloor q/4 \rfloor$ of it. Repeat until none remain. The retained set $\mathcal{W}$ has pairwise distances greater than $q/4$. Each step removes at most $\sum_{k=0}^{\lfloor q/4\rfloor}\binom{q}{k}$ vectors. For $k \le \lfloor q/4\rfloor$, $3^{q/4-k} \ge 1$, so the binomial theorem yields

$$\sum_{k=0}^{\lfloor q/4\rfloor}\binom{q}{k} \overset{③}{\le} 3^{q/4}\sum_{k=0}^{q}\binom{q}{k}3^{-k} \overset{④}{=} 3^{q/4}(4/3)^q, \qquad |\mathcal{W}| \overset{⑤}{\ge} \frac{2^q}{3^{q/4}(4/3)^q} \overset{⑥}{\ge} \exp(q/8).$$

Step ③ inserts the weights and extends the sum, ④ applies the binomial theorem, and ⑤ uses coverage of all 2^q vectors by the removed sets. Step ⑥ uses $\log 2 - \frac{1}{4}\log 3 - \log(4/3) > 1/8$. The construction depends only on q, not on any observed data. For distinct $\omega,\omega' \in \mathcal{W}$, orthonormality gives

$$\|f_\omega - f_{\omega'}\|_{L^2(\rho)}^2 \overset{⑦}{=} 4a^2 d_H(\omega,\omega') \overset{⑧}{>} \frac{\gamma q}{m}. \tag{D.44}$$

Here ⑦ counts the coefficients that differ, and ⑧ uses $d_H > q/4$ and $a^2 = \gamma/m$.

Step 2: Probability of identifying a signal. We compare each alternative with zero using the learner's full observation. The inputs, independent split, and random seed have the same law under every signal, so only the conditional training-label densities enter the likelihood ratio. Let Q_ω be the law of all inputs, the split, the training labels, and an independent random seed under f_ω. Let Q_0 be their law under the zero signal. Conditional on the inputs and split, the training labels have independent Gaussian densities with common positive variance σ_0^2. Hence Q_ω is absolutely continuous with respect to Q_0, with likelihood ratio

$$L_\omega := \frac{dQ_\omega}{dQ_0} \overset{⑨}{=} \exp\left(\frac{1}{\sigma_0^2}\sum_{i\in\mathcal{M}} y_i f_\omega(\mathbf{x}_i) - \frac{1}{2\sigma_0^2}\sum_{i\in\mathcal{M}} f_\omega(\mathbf{x}_i)^2\right).$$

Step ⑨ divides the Gaussian densities. The input, split, and random-seed laws cancel. Test labels are not observed and therefore do not occur in this ratio. The second moment of L_ω measures how well the observations can distinguish the alternative from zero. Under Q_0, each training label is conditionally $N(0,\sigma_0^2)$, so independence gives

$$\mathbb{E}_{Q_0}\left[L_\omega^2 \mid \mathbf{X}_n, \mathcal{M}\right] \overset{①}{=} \prod_{i\in\mathcal{M}}\left[\exp\left(-\frac{f_\omega(\mathbf{x}_i)^2}{\sigma_0^2}\right)\mathbb{E}_{Q_0}\left[\exp\left(\frac{2y_i f_\omega(\mathbf{x}_i)}{\sigma_0^2}\right) \mid \mathbf{X}_n, \mathcal{M}\right]\right]$$

$$\overset{②}{=} \exp\left(\frac{1}{\sigma_0^2}\sum_{i\in\mathcal{M}} f_\omega(\mathbf{x}_i)^2\right).$$

Steps ①, ② use conditional independence and $\mathbb{E}\left[e^{tY}\right] = e^{\sigma_0^2 t^2/2}$ for $Y \sim N(0,\sigma_0^2)$. The training inputs have law ρ independently for each fixed split. Since $0 \le f_\omega(\mathbf{x})^2/\sigma_0^2 \le 1$, the chord bound $e^t \le 1 + (e-1)t$ on $[0,1]$ gives

$$\mathbb{E}_{Q_0}\left[L_\omega^2\right] \overset{③}{=} \left(\int \exp\left(f_\omega(\mathbf{x})^2/\sigma_0^2\right)\, d\rho(\mathbf{x})\right)^m \overset{④}{\le} \left(1 + \frac{(e-1)a^2 q}{\sigma_0^2}\right)^m$$

$$\overset{⑤}{\leq} \exp\left(\frac{(e-1)\gamma q}{\sigma_0^2}\right) \overset{⑥}{\leq} \exp(q/16). \tag{D.45}$$

Here ③ averages the conditional second moment, ④ uses the exponential bound and $\|f_\omega\|^2_{L^2(\rho)} = a^2 q$, ⑤ uses $1 + t \leq \exp(t)$ and $ma^2 = \gamma$, and ⑥ uses the choice of γ.

Let $\widehat{\omega}$ be any measurable identification rule taking values in $\mathcal{W}$. For $\omega \in \mathcal{W}$, put $A_\omega = \{\widehat{\omega} = \omega\}$. These events partition the observation space. Applying Cauchy–Schwarz first to each likelihood ratio and then to the finite sum gives

$$\begin{aligned}\frac{1}{|\mathcal{W}|}\sum_{\omega\in\mathcal{W}} Q_\omega(A_\omega) &\overset{⑦}{=} \frac{1}{|\mathcal{W}|}\sum_{\omega\in\mathcal{W}} \mathbb{E}_{Q_0}\left[L_\omega \mathbb{I}_{\{A_\omega\}}\right] \\ &\overset{⑧}{\leq} \frac{\exp(q/32)}{|\mathcal{W}|}\sum_{\omega\in\mathcal{W}}\sqrt{Q_0(A_\omega)} \overset{⑨}{\leq} \frac{\exp(q/32)}{\sqrt{|\mathcal{W}|}} \overset{①}{\leq} \exp(-q/32).\end{aligned} \tag{D.46}$$

Step ⑦ is change of measure, and ⑧ applies Cauchy–Schwarz to each $\mathbb{E}_{Q_0}\left[L_\omega \mathbb{I}_{\{A_\omega\}}\right]$, using (D.45). For ⑨, Cauchy–Schwarz gives $\sum_\omega \sqrt{Q_0(A_\omega)} \leq \sqrt{|\mathcal{W}| \sum_\omega Q_0(A_\omega)} = \sqrt{|\mathcal{W}|}$. Step ① uses $|\mathcal{W}| \geq \exp(q/8)$. This bound is uniform over all identification rules based on the full observation, including rules using every test input.

Step 3: Separation on the test set. Only the distances among the finite alternatives need to be preserved. This avoids a uniform comparison of empirical and population norms over the entire RKHS ball. Fix distinct $\omega, \omega' \in \mathcal{W}$, and define $Z(\mathbf{x}) = (f_\omega(\mathbf{x}) - f_{\omega'}(\mathbf{x}))^2$ and $\mu = \int Z\, d\rho$. Equation (D.44) gives $\mu > \gamma q/m$. Since each nonzero difference coefficient has magnitude $2a$, the reproducing inequality and the eigenvalue lower bound give

$$0 \leq Z(\mathbf{x}) \overset{②}{\leq} \kappa^2 \|f_\omega - f_{\omega'}\|^2_{\mathcal{H}_K} \overset{③}{\leq} 4\kappa^2 a^2 \sum_{j=1}^{q} \nu_j^{-1} \overset{④}{\leq} \frac{4\kappa^2 \gamma q^{2\beta+1}}{m a_-} := B_m.$$

Step ② uses bounded evaluation, ③ includes every coefficient, and ④ uses $\nu_j^{-1} \leq j^{2\beta}/a_-$. The upper bound $B_m > 0$ is the same for every pair.

Conditional on the split, the variables $Z(\mathbf{x}_i)$, $i \in \mathcal{U}$, are independent with mean μ and values in $[0, B_m]$. We give the needed lower-tail estimate explicitly. For any $t > 0$, convexity on $[0,1]$ gives $\exp(-tz) \leq 1 - (1 - \exp(-t))z$ for $0 \leq z \leq 1$. Consequently $\mathbb{E}\left[\exp(-tZ/B_m)\right] \leq \exp(-(1-\exp(-t))\mu/B_m)$. Apply Markov's inequality to $\exp(-t\sum_{i\in\mathcal{U}} Z(\mathbf{x}_i)/B_m)$. For any $t > 0$ this gives

$$\Pr\{P_{\mathcal{U}} Z < \mu/2 \mid \mathcal{M}\} \overset{⑤}{\leq} e^{tu\mu/(2B_m)} \left(\mathbb{E}_\rho\left[e^{-tZ/B_m}\right]\right)^u \overset{⑥}{\leq} \exp\left(\frac{u\mu}{B_m}\left(t/2 - 1 + e^{-t}\right)\right).$$

Step ⑤ uses independence of the test inputs and Markov's inequality. Step ⑥ substitutes the exponential-moment bound. Taking $t = \log 2$ yields

$$\begin{aligned}\Pr\{P_{\mathcal{U}} Z < \mu/2 \mid \mathcal{M}\} &\overset{⑦}{\leq} \exp\left(-\frac{1-\log 2}{2}\frac{u\mu}{B_m}\right) \overset{⑧}{\leq} \exp\left(-\frac{u\mu}{8B_m}\right) \\ &\overset{⑨}{\leq} \exp\left(-\frac{u a_-}{32\kappa^2 q^{2\beta}}\right) \overset{①}{\leq} \exp(-4q).\end{aligned}$$

Here ⑦ uses the preceding moment bound, ⑧ uses $(1 - \log 2)/2 > 1/8$, ⑨ uses $\mu > \gamma q/m$ and the value of B_m, and ① uses $q^{2\beta+1} \leq a_- m/(128\kappa^2)$ and $u \geq m$. These estimates do not depend on the particular split.

Let G be the event that $P_{\mathcal{U}}(f_\omega - f_{\omega'})^2 \geq \|f_\omega - f_{\omega'}\|^2_{L^2(\rho)}/2$ for every distinct pair in $\mathcal{W}$. There are at most $|\mathcal{W}|^2 \leq 2^{2q}$ pairs, so the union bound gives

$$\Pr(G^c) \overset{②}{\leq} 2^{2q}\exp(-4q) \overset{③}{\leq} \exp(-2q).$$

Step ② also averages over the split, and ③ uses $\log 2 < 1$. The event involves only the inputs and the random split. Its probability over these variables is therefore the same under every Q_ω. On G, (D.44) implies

$$P_{\mathcal{U}}(f_\omega - f_{\omega'})^2 > \frac{\gamma q}{2m} \qquad (\omega, \omega' \in \mathcal{W},\ \omega \neq \omega'). \tag{D.47}$$

Step 4: Prediction error. Fix an arbitrary admissible predictor $\widehat{f}$. We convert it into an identification rule by comparing its test predictions with those of the finite alternatives. Fix an ordering of $\mathcal{W}$, and let

$$\widehat{\omega} \in \arg\min_{\omega \in \mathcal{W}} \frac{1}{u} \sum_{i \in \mathcal{U}} (\widehat{f}(\mathbf{x}_i) - f_\omega(\mathbf{x}_i))^2,$$

with ties resolved by the fixed ordering. This rule is measurable, since it minimizes finitely many measurable quantities, and it uses no test labels. It is therefore covered by (D.46), even if $\widehat{f}$ does not belong to the RKHS.

Set $r_m = \gamma q/(8m)$. If G holds and $\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_\omega) < r_m$, then $\widehat{\omega} = \omega$. To verify this implication, suppose instead that $\omega' = \widehat{\omega} \neq \omega$. The minimizing property gives $\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_{\omega'}) \leq \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_\omega) < r_m$. The triangle inequality in $\mathbb{R}^u$, with the Euclidean norm divided by $\sqrt{u}$, would imply

$$\sqrt{P_{\mathcal{U}}(f_\omega - f_{\omega'})^2} \overset{④}{\leq} \sqrt{\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_\omega)} + \sqrt{\mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_{\omega'})} \overset{⑤}{<} 2\sqrt{r_m} = \sqrt{\frac{\gamma q}{2m}},$$

contradicting (D.47). Here ④ is the triangle inequality and ⑤ uses the two error bounds. Consequently, without requiring independence between G and the rule,

$$Q_\omega \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_\omega) < r_m \right\} \overset{⑥}{\leq} Q_\omega \{\widehat{\omega} = \omega\} + Q_\omega(G^c).$$

Step ⑥ follows from the proved event inclusion. Average over the finite set $\mathcal{W}$ and use (D.46) and $Q_\omega(G^c) \leq \exp(-2q)$:

$$\frac{1}{|\mathcal{W}|} \sum_{\omega \in \mathcal{W}} Q_\omega \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_\omega) \geq r_m \right\} \overset{⑦}{\geq} 1 - \exp(-q/32) - \exp(-2q).$$

Step ⑦ takes complements of the preceding upper bounds. At least one ω attains or exceeds this average. It can depend on the predictor and sample sizes, but not on the realized observations. Finally, $q \geq \eta m^s/2$ implies

$$r_m \overset{⑧}{\geq} \frac{\gamma\eta}{16} m^{s-1} = c_{\mathrm{low}} m^{-2\beta/(2\beta+1)},$$

$$\exp(-q/32) + \exp(-2q) \overset{⑨}{\leq} 2\exp(-q/32) \overset{①}{\leq} 2\exp\left(-\frac{\eta}{64} m^s\right) = 2\exp\left(-c_{\mathrm{prob}} m^{1/(2\beta+1)}\right).$$

Step ⑧ uses the lower bound on q, ⑨ uses $2q \geq q/32$, and ① uses the same lower bound on q. Since $f_\omega \in \mathcal{H}_K(\mu_0)$ and the predictor was arbitrary, taking its infimum proves (D.42).

Step 5: Fixed inputs and expected error. We retain the same error event when replacing a global signal by its finite-sample representative. Let $\Pi_{\mathbf{X}_n}$ be the RKHS orthogonal projection onto $\mathcal{H}_{\mathbf{X}_n} = \mathrm{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. This span is finite dimensional and closed. For a fixed global signal $f \in \mathcal{H}_K(\mu_0)$, put $f_{\mathbf{X}_n} = \Pi_{\mathbf{X}_n} f$. Orthogonal projection is a contraction, so $f_{\mathbf{X}_n} \in \mathcal{H}_{\mathbf{X}_n}(\mu_0)$. The orthogonal residual $f - f_{\mathbf{X}_n}$ vanishes at every sample input, as in Step 2 of the proof of Theorem 4.4. Conditional on the inputs, this replacement therefore changes neither the distribution of the observed training labels nor any test target.

For the fixed predictor, choose the signal obtained in Step 4, and set $t_m = c_{\mathrm{low}} m^{-2\beta/(2\beta+1)}$. Conditioning on the inputs gives

$$\begin{aligned}
&\sup_{\mathbf{X}_n \in \mathcal{X}^n} \sup_{h \in \mathcal{H}_{\mathbf{X}_n}(\mu_0)} \Pr_{h, \mathbf{X}_n} \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, h) \geq t_m \right\} \\
&\overset{②}{\geq} \int \Pr_{f_{\mathbf{X}_n}, \mathbf{X}_n} \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f_{\mathbf{X}_n}) \geq t_m \right\} d\rho^{\otimes n}(\mathbf{X}_n) \\
&\overset{③}{=} \Pr_f \left\{ \mathrm{MSE}_{\mathcal{U}}(\widehat{f}, f) \geq t_m \right\} \overset{④}{\geq} 1 - 2\exp(-c_{\mathrm{prob}} m^s).
\end{aligned}$$

Here ② bounds an average by the supremum over its admissible inputs and signals, ③ uses the identical evaluations and label laws, and ④ uses Step 4. The integrand can equivalently be expressed

using the original fixed f and its Gaussian label means, which also verifies measurability. Taking the infimum over predictors proves (D.43).

The stated sample-size condition gives $c_{\text{prob}} m^s \geq \log(2/\delta)$, hence probability at least $1-\delta$. For sufficiently large m, this probability is at least $1/2$. Apply $\mathbb{E}[Z] \geq t\Pr\{Z \geq t\}$ with Z the test MSE and $t = t_m$, then take the corresponding suprema and infimum. Equations (D.42) and (D.43) give both expectation bounds without requiring attainment of the suprema. □

D.5 Computing the scales of the subset solutions

The next lemma computes the norms of the subset solutions from the subset Gram matrices and uses them to bound the diameter and residual maximum. Its proof also gives exact formulas for both quantities and verifies that jointly rescaling the kernel and ridge leaves the risk bounds unchanged.

Lemma D.6 (Exact Subset Solution Norms And Bounds For The Kernel Scales)**.** For a relevant subset A, write $p = |A|$, let $\mathbf{K}_A = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j\in A}$, and let $\mathbf{y}_A = (y_i)_{i\in A}$. Let $\rho_{A,1} \geq \cdots \geq \rho_{A,p} \geq 0$ be the eigenvalues of $\mathbf{K}_A/p$, counted with multiplicity, and choose a corresponding Euclidean orthonormal eigenbasis $(\mathbf{v}_{A,j})_{j=1}^{p}$. Recall that R_λ in (D.2) is the largest RKHS norm of a subset solution, and define $Y \geq 0$ by

$$Y^2 := \max_{\substack{A\subseteq[n]\\ |A|\in\{m,u,n\}}} P_A y^2.$$

In sums below, $j : \rho_{A,j} > 0$ means $j \in \{1,\dots,|A|\}$ with $\rho_{A,j} > 0$. If there are no positive eigenvalues, the sum is zero. Then

$$R_\lambda^2 = \max_{\substack{A\subseteq[n]\\ |A|\in\{m,u,n\}}} \frac{1}{|A|} \sum_{j:\rho_{A,j}>0} \frac{\rho_{A,j}}{(\rho_{A,j}+\lambda)^2} \left(\mathbf{v}_{A,j}^\top \mathbf{y}_A\right)^2. \tag{D.48}$$

The diameter D_λ and residual maximum L_λ satisfy

$$D_\lambda \leq 2R_\lambda, \qquad L_\lambda \leq \max_{\substack{i\leq n\\ (\mathbf{K}_\lambda)_{ii}>0}} \left(|y_i| + R_\lambda\sqrt{K(\mathbf{x}_i,\mathbf{x}_i)}\right)^2.$$

The empty maximum is zero under the convention at the beginning of Section D. Also $Y^2 \leq \min\left\{\|\mathbf{y}\|_\infty^2, nP_n y^2/N_{u,m}\right\}$. For $\lambda > 0$ the sharp multiplier bound on $0 < t \leq K_{\max}$ gives

$$R_\lambda^2 \leq \begin{cases} Y^2/(4\lambda), & 0 < \lambda \leq K_{\max}, \\ Y^2 K_{\max}/(K_{\max}+\lambda)^2, & \lambda > K_{\max}. \end{cases}$$

For $\lambda = 0$ and $\mathbf{K} \neq \mathbf{0}$, recall that $K_{\min}^+$ is the smallest positive eigenvalue among all relevant $\mathbf{K}_A/|A|$. Then $R_0^2 \leq Y^2/K_{\min}^+$. If $\mathbf{K} = \mathbf{0}$, every subset solution is zero and the excess risk is zero for every $\lambda \geq 0$. The identity (D.48) retains the label alignment and is at least as sharp as these bounds.

Proof of Lemma D.6 And Invariance Under Kernel Rescaling.

Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot,\mathbf{x}_i) : i \in [n]\}$. Here R_λ is the largest RKHS norm of a subset solution, D_λ is the diameter of the subset solutions, and L_λ is their residual maximum. Also, Y^2 is the largest subset label average $P_A y^2$. These maxima use subsets of sizes in $\{m,u,n\}$, while $K_{\max} = \max_i K(\mathbf{x}_i,\mathbf{x}_i)$ is a coordinate maximum. The eigenpairs $(\rho_{A,j}, \mathbf{v}_{A,j})$ belong to the original subset matrix $\mathbf{K}_A/|A|$, not the projected full-sample matrix. For zero ridge, $K_{\min}^+$ denotes the smallest positive eigenvalue among these relevant subset matrices. Step 2 verifies its positivity when $\mathbf{K} \neq \mathbf{0}$.

Proof outline.

(1) Exact matrix formulas. We apply the representer formula for each subset solution to compute the RKHS norm, prediction vectors, distances, and residuals.
(2) Spectral bounds. We diagonalize each subset matrix and maximize its scalar multiplier on the allowed spectral interval.

(3) Scale bounds. We use the triangle and reproducing inequalities and bound the subset label energy.
(4) Kernel normalization. We verify that joint rescaling of the kernel and ridge preserves the learner, the risk bounds, and the training-label ridge rule.

Step 1: Exact matrix formulas. Finite matrix formulas make the scales of the subset solutions explicit without requiring coordinates in an abstract RKHS basis. We fix a relevant subset A and put $p = |A|$. Here $\mathbf{K}_A$ is its Gram matrix, $\mathbf{y}_A = (y_i)_{i\in A}$, and $\dagger$ denotes the Moore–Penrose inverse, which reciprocates the positive eigenvalues and annihilates the zero eigendirections. Let $\mathbf{c}_{A,\lambda} = (\mathbf{K}_A + \lambda p \mathbf{I}_p)^\dagger \mathbf{y}_A$. The representer formula in Step 1 of the proof of Lemma D.1 gives $\widehat{f}_{A,\lambda} = \sum_{i\in A} K(\cdot, \mathbf{x}_i)(\mathbf{c}_{A,\lambda})_i$ and $\left\|\widehat{f}_{A,\lambda}\right\|_{\mathcal{H}_K}^2 = \mathbf{c}_{A,\lambda}^\top \mathbf{K}_A \mathbf{c}_{A,\lambda}$. For nonempty subsets $A, B \subseteq [n]$, define $\mathbf{K}_{A,B} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i\in A, j\in B}$, with both index sets ordered increasingly. Expanding the finite representer sums gives

$$\left\langle \widehat{f}_{A,\lambda}, \widehat{f}_{B,\lambda} \right\rangle_{\mathcal{H}_K} \overset{①}{=} \sum_{i\in A}\sum_{j\in B} (\mathbf{c}_{A,\lambda})_i (\mathbf{c}_{B,\lambda})_j K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{c}_{A,\lambda}^\top \mathbf{K}_{A,B} \mathbf{c}_{B,\lambda},$$

$$\widehat{f}_{A,\lambda}(\mathbf{x}_i) \overset{②}{=} \sum_{j\in A} K(\mathbf{x}_i, \mathbf{x}_j)(\mathbf{c}_{A,\lambda})_j = \mathbf{K}_{\{i\},A}\mathbf{c}_{A,\lambda}.$$

Here ① uses the representer expansions and kernel inner products. ② evaluates the representer expansion at $\mathbf{x}_i$.

The first identity and $\|f-g\|_{\mathcal{H}_K}^2 = \|f\|_{\mathcal{H}_K}^2 + \|g\|_{\mathcal{H}_K}^2 - 2\langle f, g\rangle_{\mathcal{H}_K}$ give

$$\left\|\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda}\right\|_{\mathcal{H}_K}^2 \overset{③}{=} \mathbf{c}_{A,\lambda}^\top \mathbf{K}_A \mathbf{c}_{A,\lambda} + \mathbf{c}_{B,\lambda}^\top \mathbf{K}_B \mathbf{c}_{B,\lambda} - 2\mathbf{c}_{A,\lambda}^\top \mathbf{K}_{A,B} \mathbf{c}_{B,\lambda}.$$

Here ③ expands the squared distance using the two norm formulas and the cross inner product.

Overlap of A and B introduces no correction because the double sum already includes every required pair of indices. The prediction identity gives the residual $\mathbf{K}_{\{i\},A}\mathbf{c}_{A,\lambda} - y_i$ at coordinate i. By the zero-diagonal characterization in Lemma D.1, this coordinate enters L_λ exactly when these predictions differ for two relevant subsets. As a result, the distance and prediction formulas compute D_λ and L_λ by taking their defining maxima, without first computing an RKHS projection.

Coefficient nonuniqueness causes no ambiguity because $\mathrm{null}(\mathbf{K}_A) = \mathrm{null}(T_A^*)$, as proved in Lemma D.1. Adding a coefficient vector in this null space changes neither the function nor its predictions, inner products, or distances.

Step 2: Spectral bounds. Diagonalization separates label alignment from a scalar spectral multiplier, whose maximum gives simpler norm bounds. Recall that $\mathbf{K}_A \mathbf{v}_{A,j} = p\rho_{A,j}\mathbf{v}_{A,j}$ and the vectors $\mathbf{v}_{A,j}$ form an orthonormal basis of $\mathbb{R}^p$. In a direction with $\rho_{A,j} + \lambda > 0$, the coefficient is $\mathbf{v}_{A,j}^\top \mathbf{c}_{A,\lambda} = (\mathbf{v}_{A,j}^\top \mathbf{y}_A)/[p(\rho_{A,j}+\lambda)]$. If $\rho_{A,j} = \lambda = 0$, that coefficient is zero by the pseudoinverse convention. Inserting the coefficients into the quadratic norm gives

$$\mathbf{c}_{A,\lambda}^\top \mathbf{K}_A \mathbf{c}_{A,\lambda} \overset{④}{=} \sum_{j=1}^{p} p\rho_{A,j} \left(\mathbf{v}_{A,j}^\top \mathbf{c}_{A,\lambda}\right)^2 \overset{⑤}{=} \frac{1}{p} \sum_{j:\rho_{A,j}>0} \frac{\rho_{A,j}}{(\rho_{A,j}+\lambda)^2} \left(\mathbf{v}_{A,j}^\top \mathbf{y}_A\right)^2.$$

Here ④ diagonalizes the quadratic form. ⑤ substitutes the representer coefficients and removes zero eigendirections.

Maximizing this squared norm over relevant A proves (D.48). If $K_{\max} = 0$, positive semidefiniteness implies $|\mathbf{K}_{ij}|^2 \le \mathbf{K}_{ii}\mathbf{K}_{jj} = 0$, so $\mathbf{K} = \mathbf{0}$. All subset solutions, R_λ, D_λ, and L_λ are then zero, proving the asserted bounds without maximizing over an empty spectral interval. Assume henceforth that $K_{\max} > 0$. Parseval's identity and nonnegativity of the eigenvalues give the label-energy and spectral bounds

$$\frac{1}{p}\sum_{j:\rho_{A,j}>0} \left(\mathbf{v}_{A,j}^\top \mathbf{y}_A\right)^2 \overset{⑥}{\le} \frac{1}{p}\sum_{j=1}^{p} \left(\mathbf{v}_{A,j}^\top \mathbf{y}_A\right)^2 \overset{⑦}{=} \frac{\|\mathbf{y}_A\|_2^2}{p} = P_A y^2,$$

$$0 < \rho_{A,j} \overset{⑧}{\le} \sum_{l=1}^{p} \rho_{A,l} \overset{⑨}{=} \frac{1}{p}\sum_{i\in A} K(\mathbf{x}_i, \mathbf{x}_i) \overset{①}{\le} K_{\max}.$$

Steps ⑥, ⑦ use Parseval, and ⑧, ⑨, ① bound a positive eigenvalue by the trace and then the largest kernel diagonal entry.

For $\lambda > 0$, the derivative of $t/(t+\lambda)^2$ is $(\lambda - t)/(t+\lambda)^3$. Its maximum on $(0, K_{\max}]$ is $1/(4\lambda)$ when $\lambda \le K_{\max}$, and $K_{\max}/(K_{\max}+\lambda)^2$ otherwise. Multiplying by $P_A y^2 \le Y^2$ proves the positive-ridge bounds. At zero ridge the multiplier is $1/t$. The finite collection of positive subset eigenvalues is nonempty because it includes the positive spectrum of $\mathbf{K}/n$, so its minimum $K^+_{\min}$ is positive. Hence

$$\left\|\widehat{f}_{A,0}\right\|^2_{\mathcal{H}_K} \overset{②}{=} \frac{1}{p} \sum_{j:\rho_{A,j}>0} \frac{(\mathbf{v}_{A,j}^\top \mathbf{y}_A)^2}{\rho_{A,j}} \overset{③}{\le} \frac{1}{K^+_{\min}} \frac{1}{p} \sum_{j:\rho_{A,j}>0} (\mathbf{v}_{A,j}^\top \mathbf{y}_A)^2 \overset{④}{\le} \frac{P_A y^2}{K^+_{\min}} \overset{⑤}{\le} \frac{Y^2}{K^+_{\min}}.$$

Here ② uses the zero-ridge norm identity. ③ uses $\rho_{A,j} \ge K^+_{\min}$. ④ uses Parseval's bound. ⑤ uses the definition of Y.

Maximizing over A gives $R_0^2 \le Y^2/K^+_{\min}$. Every denominator in the sum is positive. A zero subset Gram matrix contributes an empty sum, so no nonsingularity assumption is needed.

Step 3: Scale bounds.

For relevant A, B, the triangle and reproducing inequalities give

$$\left\|\widehat{f}_{A,\lambda} - \widehat{f}_{B,\lambda}\right\|_{\mathcal{H}_K} \le 2R_\lambda, \qquad \left|\widehat{f}_{A,\lambda}(\mathbf{x}_i) - y_i\right| \le |y_i| + R_\lambda \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)}.$$

Taking the first maximum proves $D_\lambda \le 2R_\lambda$.

This estimate uses the original kernel section because an individual subset solution need not lie in the range of Π_λ. The projected section represents differences between subset solutions, not necessarily the individual predictions. As a result, the matrix order alone does not justify replacing $K(\mathbf{x}_i, \mathbf{x}_i)$ by $(\mathbf{K}_\lambda)_{ii}$ in this residual estimate. Taking the maximum only over coordinates with $(\mathbf{K}_\lambda)_{ii} > 0$ gives the asserted bound on L_λ. Finally, $y_i^2 \le \|\mathbf{y}\|^2_\infty$ gives $P_A y^2 \le \|\mathbf{y}\|^2_\infty$. Nonnegativity gives $|A|\, P_A y^2 = \sum_{i \in A} y_i^2 \le n P_n y^2$. Since $|A| \ge N_{u,m}$, this also yields $P_A y^2 \le n P_n y^2 / N_{u,m}$. Both bounds hold for every relevant A, so their minimum bounds Y^2. The residual estimate also covers an empty set of admissible coordinates, because both residual maxima are then zero by convention.

Step 4: Kernel normalization. A change of kernel units should not improve a bound for an unchanged learner. We check that the corresponding norm and spectral factors cancel. In this step $a > 0$ denotes a kernel-scaling factor. We compare the problem with kernel K and ridge λ to the problem with kernel aK and ridge $a\lambda$.

The functions, norms, and subset solutions. The two finite spans contain exactly the same functions, since each new section is a times its original section. For $f = \sum_i \alpha_i K(\cdot, \mathbf{x}_i) = \sum_i (\alpha_i/a)(aK)(\cdot, \mathbf{x}_i)$ and $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$, the Gram norm formula gives

$$\|f\|^2_{\mathcal{H}_{aK}} = (\boldsymbol{\alpha}/a)^\top (a\mathbf{K})(\boldsymbol{\alpha}/a) = a^{-1} \|f\|^2_{\mathcal{H}_K}.$$

These identities hold for every coefficient representation, including when the Gram matrix is singular. Polarization therefore gives $\langle f, g \rangle_{\mathcal{H}_{aK}} = \frac{1}{a} \langle f, g \rangle_{\mathcal{H}_K} \qquad (f, g \in \mathcal{F})$. All subset solutions and all sections used in the projected Gram matrix belong to this span, so these identities cover every object in the bound.

Consequently $P_A(f - y)^2 + (a\lambda)\|f\|^2_{\mathcal{H}_{aK}} = J_A(f)$ for every relevant A and f in the common span.

The identical objectives preserve the unique positive-ridge subset solution. At zero ridge, scaling all RKHS norms by the same positive factor preserves minimum-norm selection.

The projected geometry and the risk bound. The difference span is unchanged, and scaling the inner product preserves orthogonality. Its projection is therefore unchanged.

The new projected sections are $a\Pi_\lambda K(\cdot, \mathbf{x}_i)$. Their inner products are multiplied by a, so the projected Gram matrix becomes $a\mathbf{K}_\lambda$. Its zero diagonal entries are unchanged. Consequently L_λ is unchanged, D_λ^2 becomes D_λ^2/a, and the eigenvalues become $a\widehat{\lambda}_j$. Both terms in the spectral minimum are unchanged because

$$\frac{a\widehat{\lambda}_j}{a\widehat{\lambda}_j + a\lambda} = \frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \lambda}, \qquad (D_\lambda^2/a)(a\widehat{\lambda}_j) = D_\lambda^2 \widehat{\lambda}_j.$$

The zero-eigenvalue convention is preserved as well. The confidence term is unchanged. The ordinary squared losses and predictions of the subset solutions are unchanged, so both ordinary excess and $b_{\lambda,u}$ are unchanged. The quantity R_0^2 is divided by a, while $K_{\max}$ and $K_{\min}^+$ are multiplied by a. Hence both sides of the ridge restriction (D.20) are invariant. The upper endpoint in (D.23) is multiplied by a, exactly as the ridge parameter, so that condition is preserved. The same cancellation holds in every head–tail bound of Corollary D.4.

At each cutoff, $D_\lambda^2 \sum_{j>Q} \widehat{\lambda}_j$ and the head ratios are unchanged. Hence every head–tail bound is identical before and after minimization.

The training-label ridge rule. For the rule in Theorem 4.3, recall $M_{\mathcal{M}} = \sqrt{K_{\max}} \left\| \widehat{f}_{m,0} \right\|_{\mathcal{H}_K}$. When $\mathbf{K}_{\mathcal{M}} \neq \mathbf{0}$, $\kappa_{\mathcal{M}}$ is the smallest positive eigenvalue of $\mathbf{K}_{\mathcal{M}}/m$. Scaling the kernel by a multiplies $\kappa_{\mathcal{M}}$ by a and leaves $M_{\mathcal{M}}$ unchanged. Formula (4.4) therefore multiplies the ridge by a, preserving the subset solution by the objective identity above. If $\mathbf{K}_{\mathcal{M}} = \mathbf{0}$, it remains zero after scaling, and both subset solutions are zero for every positive ridge choice. As a result, the subset solutions and tolerance bound are invariant in this case as well, without assigning a smallest positive eigenvalue to a zero matrix. □

D.6 Verification of the kernel comparisons

Comparisons and scope. We compare bounds for the same learner, full sample, and confidence. First, take $m = u = n/2$, $\lambda = 0$, $\mathbf{K} = \mathbf{I}_n$, and $y_i = \mathbb{1}_{\{i \in S\}}$ for $S \subseteq [n]$ with $1 \leq k = |S| \leq n/2$. Then $D_0^2 = k$ and $L_0 = 1$. The projected and original matrices have k and n unit eigenvalues, respectively. Their fixed points are $\sqrt{2}\,k/n$ and $\sqrt{2k/n}$, giving bounds of order $(k+x)/n$ and $\sqrt{k/n} + x/n$ with the same scales. The fixed-point ratio $\sqrt{k/n}$ tends to zero when $k/n \to 0$, improving the rate order for fixed x. The actual excess is $|S \cap \mathcal{U}|/u$, with mean k/n, so the projected fixed point tracks the directions contributing to test error.

Next, take a diagonal Gram matrix with ones on S and zeros elsewhere, and labels $y_i = 1$ on S and $y_i = M \geq 1$ elsewhere. All feasible predictors vanish outside S, so the common loss M^2 cancels from ordinary excess. As a result, $L_0 = 1$ and the bound has order $(k+x)/n$, independently of M. Using all coordinates in the residual maximum instead gives order $Mk/n + M^2 x/n$. Finite-envelope STLC can also use this projection and cancellation. These comparisons isolate the deterministic reductions and do not assert uniform dominance after identical preprocessing.

The subset solutions determine the analysis class, while the learner still optimizes over the full span. Convergence requires control of the displayed scales. For an identity kernel with all labels equal to one, zero-ridge test excess is one on every split, as verified below.

Verification Of The Kernel Comparisons. Recall that $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$. All three examples use $\lambda = 0$. As a result, $\widehat{f}_{A,0}$ is the minimum-RKHS-norm least-squares subset solution, D_0 is the diameter of all relevant subset solutions, and L_0 is the residual maximum over coordinates where they can differ. The matrix $\mathbf{K}_0$ is formed by projecting the kernel sections onto the span of these differences between subset solutions. For the projection comparison, we keep D_0, L_0 and the learner fixed. We compare the bound $C(r_{K,0} + L_0 x/N_{u,m})$ in (D.21) with the same expression using the full-kernel spectrum in $r_{K,0}$.

Proof outline.

(1) Subset solutions and scales. Compute the minimum-norm subset solutions, diameter, and residual maximum for the identity kernel.
(2) Projected spectrum. Identify the span of differences between subset solutions as the coordinates with nonzero labels.
(3) Critical radii and excess risk. Solve both fixed-point equations and evaluate the exact test excess.
(4) Common-loss cancellation. Restrict the kernel to those coordinates. Show that changing labels elsewhere preserves the subset solutions and the refined residual maximum.
(5) A small diameter does not imply convergence. Use the identity-kernel subset solutions and kernel normalization to obtain a vanishing diameter while test excess remains one.

Step 1: Subset solutions and scales. The identity kernel makes minimum-norm selection coordinate-wise, allowing the scales to be computed exactly. Let $n = 2m = 2u$, $\mathbf{K} = \mathbf{I}_n$, and $\lambda = 0$. We fix $S \subseteq [n]$ with $|S| = k \in \{1, \dots, n/2\}$, and set $y_i = \mathbb{1}_{\{i \in S\}}$. In this RKHS the evaluation map identifies $\mathcal{F}$ isometrically with Euclidean $\mathbb{R}^n$. Indeed, $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{1}_{\{i=j\}}$, so for $f = \sum_{j=1}^n \alpha_j K(\cdot, \mathbf{x}_j)$ one has

$$f(\mathbf{x}_i) \overset{①}{=} \alpha_i, \qquad \|f\|_{\mathcal{H}_K}^2 \overset{②}{=} \sum_{i=1}^n \alpha_i^2, \qquad J_A(f) \overset{③}{=} \frac{1}{|A|} \sum_{i \in A} (\alpha_i - y_i)^2.$$

Here ① uses the identity Gram matrix. ② uses orthonormality of its kernel sections. ③ inserts the coordinate evaluations into J_A.

Least squares requires $\alpha_i = y_i$ on A, and minimum norm sets every remaining coefficient to zero. Hence $\widehat{f}_{A,0}(\mathbf{x}_i) = \mathbb{1}_{\{i \in S \cap A\}}$, and for relevant A, B,

$$\left\| \widehat{f}_{A,0} - \widehat{f}_{B,0} \right\|_{\mathcal{H}_K}^2 \overset{④}{=} \sum_{i \in S} \left(\mathbb{1}_{\{i \in A\}} - \mathbb{1}_{\{i \in B\}} \right)^2 \overset{⑤}{=} |S \cap (A \setminus B)| + |S \cap (B \setminus A)| \overset{⑥}{\leq} k.$$

Here ④ uses the explicit subset solutions. ⑤ counts the nonzero squared indicator differences. ⑥ bounds the disjoint subsets of S by $|S| = k$.

The full-sample subset solution has evaluation vector $\mathbf{y}$, while a subset of size m disjoint from S has the zero subset solution. Such a subset exists because $n - k \geq m$. Their squared RKHS distance is k, proving $D_0^2 = k$. All residual magnitudes are at most one, and this zero subset solution has a unit residual on a coordinate where the zero subset solution and the full-sample subset solution differ. That coordinate is included in the defining maximum for L_0 by the zero-diagonal characterization in Lemma D.1. Consequently $L_0 = 1$.

Step 2: Projected spectrum. We identify the directions in which predictions of subset solutions can vary, since only those directions enter the projected spectrum. Every difference between subset solutions is supported on S. Conversely, fix $i \in S$ and $j \notin S$, and choose a set $C \subseteq [n] \setminus \{i, j\}$ of size $m - 1$. Such a set exists because $n - 2 = 2m - 2 \geq m - 1$, and $j \notin S$ exists because $k \leq n/2 < n$. The subset solutions on $C \cup \{i\}$ and $C \cup \{j\}$ satisfy

$$\mathbb{1}_{\{l \in S \cap (C \cup \{i\})\}} - \mathbb{1}_{\{l \in S \cap (C \cup \{j\})\}} \overset{⑦}{=} \mathbb{1}_{\{l=i\}}, \qquad l \in [n].$$

Here ⑦ cancels the common subset C and uses $i \in S$, $j \notin S$.

The difference is the ith unit vector. Hence the difference span is exactly the coordinate subspace supported on S, whose projection keeps coordinates in S and zeros the rest. Consequently $(\mathbf{K}_0)_{il} = \mathbb{1}_{\{i \in S\}} \mathbb{1}_{\{l \in S\}} \mathbb{1}_{\{i=l\}}$. The normalized projected matrix has k eigenvalues $1/n$, while $\mathbf{K}/n$ has n such eigenvalues.

Step 3: Critical radii and excess risk. With the geometry computed, we can solve the two fixed-point equations and compare them with the exact test loss. We substitute $D_0^2 = k$, $L_0 = 1$, and the k positive eigenvalues $1/n$ into the definition of $r_{K,0}$. Since $N_{u,m} = n/2$, its positive fixed-point equation is $r^2 \overset{⑧}{=} (2k/n) \min\{r, k/n\}$, whereas the full spectrum gives $r^2 \overset{⑨}{=} 2 \min\{r, k/n\}$. Here ⑧ inserts the k positive projected eigenvalues and the computed scales into the fixed-point equation. Here ⑨ inserts all n positive original eigenvalues with the same scales. We solve both equations by checking their two branches. In the projected equation, a solution with $r \leq k/n$ would satisfy $r^2 = (2k/n)r$ and hence $r = 2k/n > k/n$. In the full-spectrum equation, that branch would give $r^2 = 2r$ and hence $r = 2 > k/n$. Both contradict the branch restriction. On the other branch the equations are $r^2 = 2k^2/n^2$ and $r^2 = 2k/n$, respectively. Their positive roots are $\sqrt{2}\, k/n$ and $\sqrt{2k/n}$. Both satisfy $r \geq k/n$ because $k > 0$ and $k/n \leq 1/2$, so they are the unique positive solutions. With the numerical constant C from Theorem D.3, direct substitution gives the projected bound $C(\sqrt{2}\, k + 2x)/n$ and the full-spectrum bound $C(\sqrt{2kn} + 2x)/n$. Their respective orders are $(k + x)/n$ and $\sqrt{k/n} + x/n$. These compare the same D_0, L_0 and confidence level, so the difference is solely the spectrum used in the majorant.

For the actual split, the training subset solution is zero on $\mathcal{U}$, whereas the test subset solution interpolates every test label. Disjointness of $\mathcal{M}$ and $\mathcal{U}$ gives

$$\mathcal{E}_{\mathcal{U},0}(\widehat{f}_{m,0}) \overset{①}{=} \frac{1}{u} \sum_{i \in \mathcal{U}} \left[(0 - y_i)^2 - (y_i - y_i)^2 \right] \overset{②}{=} \frac{1}{u} \sum_{i \in \mathcal{U}} \mathbb{1}_{\{i \in S\}} \overset{③}{=} \frac{|S \cap \mathcal{U}|}{u}.$$

Here ① uses zero training predictions and exact test interpolation on $\mathcal{U}$. ② uses the binary labels. ③ counts the indices in $S \cap \mathcal{U}$.

Each index belongs to the uniform test subset with probability u/n. Linearity of expectation gives

$$\mathbb{E}_{\mathcal{U}}\left[\mathcal{E}_{\mathcal{U},0}(\widehat{f}_{m,0})\right] \overset{④}{=} \frac{1}{u}\sum_{i\in S}\Pr\left[i \in \mathcal{U}\right] \overset{⑤}{=} \frac{1}{u}\, k\frac{u}{n} = \frac{k}{n}.$$

Here ④ uses linearity of expectation. ⑤ uses inclusion probability u/n for each of the k indices.

This calculation does not require independence of the inclusion indicators. Also $|S \cap \mathcal{U}| \le k$, so the exact excess is at most $2k/n$ for every split. Since the specified C exceeds two, this deterministic bound in particular proves the simpler bound $C(k+x)/n$ with that same C. The common-loss example in Step 4 has the same excess risk, so it has the same deterministic excess bound with the theorem's numerical constant. For the critical radii themselves, the ratio is exactly $\frac{\sqrt{2}\,k/n}{\sqrt{2k/n}} = \sqrt{\frac{k}{n}}$. It tends to zero whenever $k/n \to 0$. For fixed x and $k \ge 1$, the confidence term x/n is at most a fixed multiple of k/n, so the same improvement in order holds for the complete bounds in this regime.

Step 4: Common-loss cancellation. Coordinates where every feasible prediction is fixed can carry large common loss without contributing to excess risk. Keep S, n, m, u fixed, but now let $\mathbf{K}$ be diagonal with ones on S and zeros elsewhere. Its RKHS evaluation space consists exactly of vectors supported on S, with Euclidean norm on those coordinates. We set $y_i = 1$ on S and $y_i = M$ elsewhere, where $M \ge 1$ is finite. For $i \notin S$, the zero kernel diagonal implies $K(\cdot, \mathbf{x}_i) = 0$, so every feasible function vanishes there. The subset objective is

$$\frac{1}{|A|}\sum_{i\in A\cap S}\left(f(\mathbf{x}_i)-1\right)^2 + \frac{M^2\,|A\setminus S|}{|A|}.$$

The second term is independent of f. Least squares and minimum-norm tie-breaking therefore give exactly the evaluation vectors from Step 1. Their norms and pairwise distances also agree. The projected matrix is again diagonal with ones on S, so L_0 takes its maximum only on S and equals one. The unrestricted residual maximum, however, is M^2. Replacing $L_0 = 1$ by M^2 in the fixed-point equation gives $r^2 \overset{⑥}{=} (2M^2k/n)\min\{r, k/n\}$. Here ⑥ replaces only the residual maximum by M^2 in the projected fixed-point equation. On the branch $r \le k/n$, a positive root would satisfy $r = 2M^2k/n > k/n$, a contradiction because $M \ge 1$. On the other branch, $r^2 = 2M^2k^2/n^2$, whose positive root $\sqrt{2}\,Mk/n$ satisfies $r \ge k/n$. The confidence term becomes $2M^2x/n$. As a result, the bound is exactly $C(\sqrt{2}\,Mk + 2M^2x)/n$ with the theorem's same C. The refined bound remains $C(\sqrt{2}\,k + 2x)/n$. The ordinary excess is also unchanged, because the common losses cancel between the two objective values:

$$J_{\mathcal{U}}(\widehat{f}_{m,0}) \overset{⑦}{=} \frac{|S\cap\mathcal{U}|}{u} + \frac{M^2\,|\mathcal{U}\setminus S|}{u}, \qquad J_{\mathcal{U}}(\widehat{f}_{u,0}) \overset{⑧}{=} \frac{M^2\,|\mathcal{U}\setminus S|}{u}.$$

Here ⑦ evaluates the training subset solution on both parts of $\mathcal{U}$. ⑧ uses exact fitting on $S \cap \mathcal{U}$ and forced zero predictions elsewhere.

Their difference is $|S \cap \mathcal{U}|\,/u$, independent of M. This proves the common-loss comparison stated above.

Step 5: A small diameter does not imply convergence. Kernel rescaling separates the numerical RKHS diameter from prediction performance, showing why the remaining scales must be retained. Keep $n = 2m = 2u$, but now set $y_i = 1$ for every i. The minimum-norm calculation in Step 1 applies to any label vector. For $\mathbf{K} = \mathbf{I}_n$, it gives $\widehat{f}_{A,0}(\mathbf{x}_i) = \mathbb{1}_{\{i\in A\}}$ for every nonempty subset A. Next replace the kernel by n^2K. Step 4 of the proof of Lemma D.6 shows that the subset solutions are unchanged and squared RKHS norms are divided by n^2. As a result, for the resulting Gram matrix $\mathbf{K} = n^2\mathbf{I}_n$, an evaluation vector $\mathbf{v}$ has squared norm $\|\mathbf{v}\|_2^2/n^2$. For relevant subsets A, B, the squared distance is therefore

$$\left\|\widehat{f}_{A,0} - \widehat{f}_{B,0}\right\|_{\mathcal{H}_K}^2 \overset{⑨}{=} \frac{|A\setminus B| + |B\setminus A|}{n^2} \overset{①}{\le} \frac{n}{n^2} = \frac{1}{n}.$$

Here ⑨ uses the indicator predictions of the subset solutions and the rescaled RKHS norm. ① bounds the size of the symmetric difference by n.

Taking $A = \mathcal{M}$ and $B = \mathcal{U}$ attains equality, so $D_0^2 = 1/n$. On $\mathcal{U}$, the training subset solution is zero and the test subset solution equals one. Their squared losses are therefore $(0-1)^2 = 1$ and $(1-1)^2 = 0$ at every test coordinate, so $J_{\mathcal{U}}(\widehat{f}_{m,0}) - J_{\mathcal{U}}(\widehat{f}_{u,0}) = 1$ for every split. As a result, the excess is one, while $D_0 = 1/\sqrt{n}$ tends to zero as n increases through even integers. This proves the stated obstruction to deducing convergence from the diameter alone. □

D.7 Recovery of the bounded TKL rate

We compare the bounded-loss special case with Yang (2026, Theorem 5.2) using the original kernel spectrum. The following statement supplies the comparison cited in Section 4.3 and Table 2.

Bounded-loss special case at zero ridge. For fixed $\mu \geq 0$, $M > 0$, let $\mathcal{F}_\mu = \{f \in \mathcal{F} : \|f\|_{\mathcal{H}_K} \leq \mu\}$. For this bounded-ball comparison, allow a general loss $\ell_f(i) = \ell(f(\mathbf{x}_i), y_i)$. Take both empirical minimizers over this ball without ridge, with $\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) = P_{\mathcal{U}}\ell_{\widehat{f}_m} - \inf_{f\in\mathcal{F}_\mu} P_{\mathcal{U}}\ell_f$. Let $f_n^\star$ be the full-sample oracle on the ball. For constants $L \geq 0$ and $B' > 0$, assume $0 \leq \ell_f(i) \leq M$, $a \mapsto \ell(a, y_i)$ is L-Lipschitz on $\{f(\mathbf{x}_i) : f \in \mathcal{F}_\mu\}$ for each i, and $P_n(f - f_n^\star)^2 \leq B' P_n(\ell_f - \ell_{f_n^\star})$ for every $f \in \mathcal{F}_\mu$. For $x \geq 1$, with probability at least $1 - \exp(-x)$ over the random split,

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) \leq C_{\mathrm{bd}} \min_{Q\in\{0,\dots,n\}} \left\{ Q\left(\frac{1}{u} + \frac{1}{m}\right) + \left(\frac{1}{\sqrt{u}} + \frac{1}{\sqrt{m}}\right)\sqrt{\sum_{j>Q} \lambda_j(\mathbf{K}/n)} \right\} + C_{\mathrm{bd}} \frac{x}{N_{u,m}}. \tag{D.49}$$

The constant $C_{\mathrm{bd}} > 0$ depends only on M, L, B', μ. This recovers the rate of Yang (2026, Theorem 5.2) as a bounded-loss special case under the stated Lipschitz and prediction-localization conditions. For squared loss, $L = 2\sqrt{M}$ and $B' = 1$ suffice. If the ball contains all relevant zero-ridge subset solutions, then $D_0 \leq 2\mu$, $L_0 \leq M$, and Theorem 4.3 gives (D.49) via $N_{u,m}^{-1} \leq u^{-1} + m^{-1}$ and $N_{u,m}^{-1/2} \leq u^{-1/2} + m^{-1/2}$, preserving the learner and test oracle. For unbounded squared loss, (4.5) retains this spectral order when D_0, L_0 remain bounded, plus $2\varepsilon^2$ for positive ridge. The proof below establishes these bounded special cases.

Proof of The Bounded-Loss Specialization (D.49). Recall that $\mathcal{F}_\mu$ is the RKHS ball of radius μ in $\mathcal{F} = \operatorname{span}\{K(\cdot, \mathbf{x}_i) : i \in [n]\}$, and $f_n^\star$ is its full-sample oracle. The constant M bounds the nonnegative losses, L is their Lipschitz constant on the permitted prediction intervals, and B' satisfies $P_n(f - f_n^\star)^2 \leq B' P_n(\ell_f - \ell_{f_n^\star})$ on this ball. For the bounded, unnormalized losses, put $g_f = \ell_f - \ell_{f_n^\star}$ and $\mathcal{G}_\star = \{g_f : f \in \mathcal{F}_\mu\}$. The surrogate will be $V_\star(g_f) = BP_n g_f$, with B specified in Step 1, and $\mathcal{G}_\star(r) = \{g_f \in \mathcal{G}_\star : V_\star(g_f) \leq r\}$. Both empirical minimizers and the ordinary excess in (D.49) refer to this same ball.

Proof outline.

(1) Bounded-ball geometry. We establish existence, the Bernstein condition, and $B_\infty(\mathcal{G}_\star) \leq 2M$. We verify the prediction-localization condition for squared loss.
(2) Localized spectral bound. We bound the localized prediction process by splitting its empirical eigenbasis. We transfer this bound to the localized excess-loss class and its centered squares, verifying (4.1) with one common sub-root majorant.
(3) Fixed points and ERM. We solve its fixed-point inequality and insert the result into the parameterized ERM bound (C.1). We compare the resulting split-size dependence with the cited bounded-TKL rate and the zero-ridge corollary.

Step 1: Bounded-ball geometry. The bounded-ball result uses constrained rather than full-span optimization, so we verify the appropriate directional inequality first. The RKHS ball in the finite-dimensional space $\mathcal{F}$ is compact. Coordinate evaluations and the Lipschitz loss are continuous, so the training, test, and full-sample risks attain their minima. We fix any selections. There are finitely many splits, hence these selections are measurable. If $\mu = 0$ or $\mathbf{K} = \mathbf{0}$, the ball consists of the zero function. If $M = 0$, the nonnegative losses are zero for every feasible predictor and every coordinate.

If $L = 0$, each coordinate loss is constant on its permitted prediction interval. In each of these cases all feasible losses agree at every sample point, so the excess is zero for every split. We may therefore assume $\mu > 0$, $\mathbf{K} \neq \mathbf{0}$, and $L > 0$ for the remaining steps. We set $B = \max\{1, L^2 B'\}$ and $\theta = 1$. For $g_f = \ell_f - \ell_{f_n^\star}$, the two comparison assumptions give

$$\mathrm{Var}[g_f] \overset{①}{\leq} P_n g_f^2 \overset{②}{\leq} L^2 P_n (f - f_n^\star)^2 \overset{③}{\leq} L^2 B' P_n g_f \overset{④}{\leq} B P_n g_f.$$

Here ① bounds variance by the second moment. ② uses loss Lipschitzness. ③ uses prediction localization. ④ uses $B \geq L^2 B'$ and $P_n g_f \geq 0$.

As a result, Assumption 1(2) holds. Step 2 verifies the localized TC condition (4.1) for $\mathcal{G}_\star$ and $V_\star$. Every g_f takes values in $[-M, M]$, hence $|g_f(i) - P_n g_f| \leq |g_f(i)| + P_n |g_f| \leq 2M$. Consequently $B_\infty(\mathcal{G}_\star) \leq 2M$.

For the squared-loss assertion, we fix f in the ball and put $d = f - f_n^\star$. Unlike the full-span minimizer in Lemma D.1, the ball-constrained minimizer may lie on the boundary. Its first-order condition therefore gives an inequality, not a normal-equation equality.

Convexity of the ball makes $f_n^\star + td$ feasible for $0 \leq t \leq 1$.

For $0 < t \leq 1$, feasibility and optimality give

$$0 \overset{⑤}{\leq} \frac{P_n(f_n^\star + td - y)^2 - P_n(f_n^\star - y)^2}{t} \overset{⑥}{=} 2P_n[(f_n^\star - y)d] + tP_n d^2.$$

Here ⑤ uses feasibility and oracle optimality. ⑥ expands the quadratic difference and divides by $t > 0$.

Taking $t \downarrow 0$ gives $2P_n[(f_n^\star - y)d] \geq 0$, including at a boundary oracle. At $t = 1$ the expansion gives $P_n g_f = P_n d^2 + 2P_n[(f_n^\star - y)d] \overset{⑦}{\geq} P_n d^2$. Here ⑦ uses the nonnegative directional derivative at the constrained oracle. This proves prediction localization with $B' = 1$. The permitted predictions at coordinate i form a compact interval, as the image of the compact convex ball under evaluation. Bounded squared loss gives $|a - y_i|, |b - y_i| \leq \sqrt{M}$ for any two predictions in this interval. Factoring the squared-loss difference gives $\left|(a - y_i)^2 - (b - y_i)^2\right| \leq 2\sqrt{M}\,|a - b|$.

As a result, squared loss is $2\sqrt{M}$-Lipschitz on this interval. Step 2 translates the interval by the oracle prediction before extending the loss-difference map, so that the extension vanishes at zero as required by contraction.

Step 2: Localized spectral bound. Prediction localization controls the spectral head and the ball radius controls the tail. Contraction then transfers this prediction bound to both loss processes. Fix $r \geq 0$ and $p \in \{m, u\}$. We use $\mathcal{G}_\star(r)$ with the B and $\theta = 1$ from Step 1. The surrogate $V_\star(g_f) = BP_n g_f$ and prediction localization give

$$g_f \in \mathcal{G}_\star(r) \quad \Longrightarrow \quad P_n(f - f_n^\star)^2 \overset{⑧}{\leq} B'r/B, \qquad \|f - f_n^\star\|_{\mathcal{H}_K} \overset{⑨}{\leq} 2\mu.$$

Here ⑧ uses prediction localization and $BP_n g_f \leq r$. ⑨ uses the triangle inequality on the ball.

The localized ball is compact by continuity of the full-sample loss, and contains the oracle. The induced prediction and loss-difference classes are therefore deterministic and contain zero. At $r = 0$, prediction localization forces all sample differences to vanish, hence both localized TC quantities are zero.

We write $e_j = \lambda_j(\mathbf{K}/n)$ in this proof, and let $d_0 = \mathrm{rank}(\mathbf{K})$. For $j \leq d_0$, we construct basis functions $\phi_j \in \mathcal{F}$ from the eigenvectors of $\mathbf{K}/n$ with positive eigenvalues. These empirical eigenfunctions are normalized so that $P_n(\phi_j \phi_k) = \mathbb{1}_{\{j=k\}}$ and $\langle \phi_j, \phi_k \rangle_{\mathcal{H}_K} = \mathbb{1}_{\{j=k\}}/e_j$, as in Step 3 of the proof of Lemma D.1. Explicitly, choose Euclidean orthonormal eigenvectors $\mathbf{v}_j$ satisfying $\mathbf{K}\mathbf{v}_j = ne_j\mathbf{v}_j$ for $1 \leq j \leq d_0$, and put $\phi_j = (\sqrt{n}\, e_j)^{-1} \sum_{i=1}^n (\mathbf{v}_j)_i K(\cdot, \mathbf{x}_i)$.

The same Gram calculation as in Step 3 of Lemma D.1 gives $\phi_j(\mathbf{x}_i) = \sqrt{n}(\mathbf{v}_j)_i$ and the two stated orthogonality relations. Since $d_0 = \dim\mathcal{F}$, these functions form a basis.

An expansion $d = \sum_{j=1}^{d_0} a_j \phi_j$ therefore satisfies $P_n d^2 = \sum_{j=1}^{d_0} a_j^2$ and $\|d\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{d_0} a_j^2/e_j$. These coefficient sums contain only positive eigendirections.

Let $J_1, \dots, J_p$ be independent uniform indices in $[n]$ and $\varepsilon_1, \dots, \varepsilon_p$ independent Rademacher variables, independent also of the sampled indices. For $j \le d_0$, put $Z_j = p^{-1} \sum_{k=1}^{p} \varepsilon_k \phi_j(\mathbf{x}_{J_k})$.

Independence and empirical orthonormality give $\mathbb{E}_{J,\varepsilon}\left[Z_j^2\right] = 1/p$, as in Step 3 of Lemma D.1. Set $Z_j = 0$ for $j > d_0$. For a cutoff $0 \le Q \le n$, weighted Cauchy–Schwarz gives, uniformly over $P_n d^2 \le B'r/B$ and $\|d\|_{\mathcal{H}_K} \le 2\mu$,

$$\left|\sum_j a_j Z_j\right| \le \sqrt{B'r/B}\sqrt{\sum_{j\le Q} Z_j^2} + 2\mu\sqrt{\sum_{j>Q} e_j Z_j^2}.$$

The first coefficient bound uses prediction localization, and the second uses the RKHS norm. Taking the supremum and then expectation gives, by Jensen's inequality and the second moments,

$$\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{d\in\mathcal{F}\\ P_n d^2 \le B'r/B,\ \|d\|_{\mathcal{H}_K}\le 2\mu}} \left|\frac{1}{p}\sum_{k=1}^{p} \varepsilon_k d(\mathbf{x}_{J_k})\right|\right] \overset{①}{\le} \sqrt{\frac{B'Qr}{Bp}} + 2\mu\sqrt{\frac{1}{p}\sum_{j>Q} e_j}. \tag{D.50}$$

Step ① uses $\mathbb{E}\left[\sqrt{\sum_{j\le Q} Z_j^2}\right] \le \sqrt{Q/p}$ and $\mathbb{E}\left[\sqrt{\sum_{j>Q} e_j Z_j^2}\right] \le \sqrt{p^{-1}\sum_{j>Q} e_j}$. No independence across eigendirections is required. Zero padding and empty sums cover singular matrices and both endpoint cutoffs.

For each coordinate i, write $[\alpha_i, \beta_i] = \{f(\mathbf{x}_i) - f_n^\star(\mathbf{x}_i) : f \in \mathcal{F}_\mu\}$. This is a compact interval because the ball is compact and convex and evaluation is continuous and linear. Also $\alpha_i \le 0 \le \beta_i$, since $f_n^\star$ belongs to the ball. On this interval the map $z \mapsto \ell(f_n^\star(\mathbf{x}_i) + z, y_i) - \ell(f_n^\star(\mathbf{x}_i), y_i)$ is L-Lipschitz and vanishes at zero. An extension to all of $\mathbb{R}$ is

$$z \longmapsto \ell\left(f_n^\star(\mathbf{x}_i) + \min\left\{\beta_i, \max\left\{\alpha_i, z\right\}\right\}, y_i\right) - \ell\left(f_n^\star(\mathbf{x}_i), y_i\right).$$

Clipping is 1-Lipschitz and fixes $[\alpha_i, \beta_i]$, including zero. The extension therefore preserves the loss differences and their Lipschitz constant, as in Lemma D.1. Put $\mathcal{H} = \mathcal{G}_\star(r)$. Theorem A.2 applies directly to this loss class through its localized prediction differences, with Lipschitz constant L. We first contract on the localized ball and then enlarge the resulting linear prediction class to the two-constraint set in (D.50). This gives

$$\begin{aligned}\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}) &\overset{②}{\le} L\mathbb{E}_{J,\varepsilon}\left[\sup_{\substack{f\in\mathcal{F}_\mu\\ g_f\in\mathcal{G}_\star(r)}} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (f - f_n^\star)(\mathbf{x}_{J_k})\right] \\ &\overset{③}{\le} L\left[\sqrt{\frac{B'}{B}}\sqrt{\frac{Qr}{p}} + 2\mu\sqrt{\frac{1}{p}\sum_{j>Q} e_j}\right].\end{aligned}$$

Here ② contracts the loss on the localized ball. ③ enlarges to the prediction constraint set and uses (D.50).

The prediction constraint set is symmetric, so its linear supremum equals the absolute supremum in (D.50). The bound holds at every cutoff.

Apply Theorem A.3 directly to $\mathcal{H}$:

$$\max_{\eta\in\{+,-\}} \mathfrak{R}_p^{\eta}(\mathcal{H}) \overset{④}{\le} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}).$$

Here ④ applies Theorem A.3 with the original class.

This is a fixed compact class, so all suprema are finite and the cited comparison applies with the same independent uniform indices. For the second level, each $h - P_n h$, with $h \in \mathcal{H}$, takes values in $[-2M, 2M]$, where squaring is $4M$-Lipschitz and vanishes at zero. Apply Theorem A.3 to $\mathcal{H}^{[2]}$,

then use Theorem A.2 and Lemma A.1. The complete chain is

$$\mathfrak{R}_p^+(\mathcal{H}^{[2]}) \overset{⑤}{\le} 2\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}^{[2]}) = 2\mathbb{E}_{J,\varepsilon}\left[\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (h(J_k) - P_n h)^2\right]$$
$$\overset{⑥}{\le} 8M\mathbb{E}_{J,\varepsilon}\left[\sup_{h\in\mathcal{H}} \frac{1}{p}\sum_{k=1}^{p} \varepsilon_k (h(J_k) - P_n h)\right] \overset{⑦}{\le} 16M\mathfrak{R}_p^{(\mathrm{ind})}(\mathcal{H}).$$

Here ⑤ applies Theorem A.3 to the centered-square class. ⑥ contracts squaring on $[-2M, 2M]$. ⑦ applies Lemma A.1.

Clipping to $[-2M, 2M]$ makes the square-map extension globally Lipschitz. Including loss contraction, the two TC bounds have common multiplier $L\max\{2, 16M\} \le 8L\max\{1, 16M\}$. We take $p = N_{u,m}$ and set $C_1 := \max\left\{1, 8L\max\{1, 16M\}\max\left\{\sqrt{B'/B}, 2\right\}\right\}$.

For each cutoff, the coefficient comparison is

$$8L\max\{1, 16M\}\left[\sqrt{B'/B}\sqrt{\frac{Qr}{N_{u,m}}} + 2\mu\sqrt{\frac{1}{N_{u,m}}\sum_{j>Q} e_j}\right]$$
$$\overset{⑧}{\le} C_1\left[\sqrt{\frac{Qr}{N_{u,m}}} + \mu\sqrt{\frac{1}{N_{u,m}}\sum_{j>Q} e_j}\right].$$

Here ⑧ uses the definition of C_1 to bound the head and tail coefficients separately.

Minimizing over cutoffs verifies (4.1) with

$$\psi(r) = C_1 \min_{Q\in\{0,\dots,n\}}\left\{\sqrt{\frac{Qr}{N_{u,m}}} + \mu\sqrt{\frac{1}{N_{u,m}}\sum_{j>Q} e_j}\right\}, \qquad r \ge 0.$$

Each candidate is continuous and nondecreasing, so the finite minimum is continuous and nondecreasing. For $r > 0$, division by the common positive factor $\sqrt{r}$ gives

$$\frac{\psi(r)}{\sqrt{r}} \overset{⑨}{=} C_1 \min_{Q\in\{0,\dots,n\}}\left\{\sqrt{\frac{Q}{N_{u,m}}} + \mu\sqrt{\frac{1}{N_{u,m}r}\sum_{j>Q} e_j}\right\}.$$

Here ⑨ divides every candidate by the common positive factor $\sqrt{r}$.

A finite minimum of nonincreasing functions is nonincreasing. Also $\psi(0) = 0$ by $Q = d_0$, while every candidate is positive for $r > 0$ because $\mu, d_0 > 0$. Hence ψ is nonzero and sub-root. Lemma C.1 gives its finite positive fixed point.

Step 3: Fixed points and ERM. The common majorant is explicit at every cutoff, allowing half of its fixed point to be absorbed before the ERM substitution. We recall that $B = \max\{1, L^2B'\}$ and $\theta = 1$ were verified in Step 1, and that C_1 is the common TC multiplier from Step 2. We use the UTLC constant $c_1 = 64(4 + c_0)^2$. Here $R_\star$ is the UTLC remainder for the unnormalized oracle excess class, and Q_t is its scalar absorption bound. No division by the full-span residual maximum L_λ is used. Choose $\psi_\star = \psi$ and write r_{exc} for its fixed point. For every fixed cutoff Q, the definition of ψ gives

$$r_{\mathrm{exc}} = \psi(r_{\mathrm{exc}}) \overset{①}{\le} C_1\sqrt{Qr_{\mathrm{exc}}/N_{u,m}} + C_1\mu\sqrt{N_{u,m}^{-1}\sum_{j>Q} e_j}.$$

Here ① evaluates the minimum defining ψ at the chosen cutoff.

The first term is at most $r_{\mathrm{exc}}/2 + C_1^2Q/(2N_{u,m})$ by $2ab \le a^2 + b^2$. Absorbing $r_{\mathrm{exc}}/2$ and minimizing over cutoffs gives

$$r_{\mathrm{exc}} \overset{②}{\le} \min_{Q\in\{0,\dots,n\}}\left\{\frac{C_1^2Q}{N_{u,m}} + 2C_1\mu\sqrt{\frac{1}{N_{u,m}}\sum_{j>Q} e_j}\right\}.$$

Here ② absorbs half the fixed point by $2ab \le a^2 + b^2$ and minimizes over cutoffs.

All cutoff bounds hold deterministically, so minimization has no confidence cost. Apply the parameterized ERM bound (C.1) from Section C.2 to $\mathcal{F}_\mu$ with $a = 2B \ge 2$ and $x_m = x_u = s = x + \log 2$. The linear absorption calculation (C.5) gives excess at most $3R_\star(2B, s)$, with failure probability at most $2\exp(-s) = \exp(-x)$ over the random split. By the envelope bound in Step 1, $R_\star(2B, s) \le c_1[2Br_{\mathrm{exc}} + (2B + 2M)s/N_{u,m}]$.

Finally, $N_{u,m}^{-1} \le u^{-1} + m^{-1} \le 2N_{u,m}^{-1}$ and $N_{u,m}^{-1/2} \le u^{-1/2} + m^{-1/2} \le 2N_{u,m}^{-1/2}$, because one of u, m equals $N_{u,m}$ and the other is no smaller. Insert the fixed-point bound above and use $s \le (1 + \log 2)x$. Before replacing the split-size factors, the resulting inequality is

$$\mathcal{E}_{\mathcal{U}}(\widehat{f}_m) \overset{③}{\le} 3c_1 \min_{Q\in\{0,\dots,n\}} \left\{ \frac{2BC_1^2 Q}{N_{u,m}} + 4BC_1\mu\sqrt{\frac{1}{N_{u,m}}\sum_{j>Q} e_j} \right\} + 3c_1(2B + 2M)(1 + \log 2)\frac{x}{N_{u,m}}.$$

Here ③ inserts the fixed-point estimate in the ERM remainder and uses $x + \log 2 \le (1 + \log 2)x$.

The split-size replacements enlarge each cutoff expression. Choosing $C_{\mathrm{bd}} = 3c_1 \max\{2BC_1^2, 4BC_1\mu, (2B + 2M)(1 + \log 2)\}$ therefore proves (D.49). By the definitions of B and C_1, this constant depends only on M, L, B', μ. This recovers the spectral expression of Yang (2026, Theorem 5.2) as the bounded-loss special case under its Lipschitz and prediction-localization assumptions, without asserting equality of optimized constants. For unbounded squared loss, the same spectral order follows from (4.5) when D_0, L_0 stay bounded, up to its positive-ridge tolerance term.

For squared loss, the zero-ridge conclusion of Theorem 4.3 also gives a direct comparison if $\widehat{f}_{A,0} \in \mathcal{F}_\mu$ for every $A \subseteq [n]$ with $|A| \in \{m, u, n\}$. Under this additional condition, the triangle inequality gives $D_0 \le 2\mu$, and the loss bound on the ball gives $L_0 \le M$. Recall $\widehat{\lambda}_j^{(0)} = \lambda_j(\mathbf{K}_0/n)$. The matrix order $\mathbf{K}_0 \preceq \mathbf{K}$ and the eigenvalue min–max principle give $\widehat{\lambda}_j^{(0)} \le e_j$. At every cutoff, these bounds yield

$$\frac{L_0}{N_{u,m}}\left(x + \sum_{j\le Q} \mathbb{1}_{\{\widehat{\lambda}_j^{(0)} > 0\}}\right) + D_0\sqrt{\frac{L_0}{N_{u,m}}\sum_{j>Q}\widehat{\lambda}_j^{(0)}} \overset{④}{\le} \frac{M(x + Q)}{N_{u,m}} + 2\mu\sqrt{\frac{M}{N_{u,m}}\sum_{j>Q} e_j}.$$

Here ④ uses $L_0 \le M$, $D_0 \le 2\mu$, eigenvalue domination, and at most Q positive head eigenvalues.

Minimizing over Q and replacing the split-size factors gives the same spectral form, with constants depending on M, μ. Since the unrestricted training and test subset solutions belong to the ball, they also attain its minima. The learner and ordinary excess risk are therefore unchanged. Steps 1–3 prove bounded-ball recovery without this additional containment condition. □