Supervised Kernel Thinning

Albert Gong **Kyuseong Choi** **Raaz Dwivedi**

Cornell Tech, Cornell University

agong,kc728,dwivedi@cornell.edu

Abstract

The kernel thinning algorithm of [10] provides a better-than-i.i.d. compression of a generic set of points. By generating high-fidelity coresets of size significantly smaller than the input points, KT is known to speed up unsupervised tasks like Monte Carlo integration, uncertainty quantification, and non-parametric hypothesis testing, with minimal loss in statistical accuracy. In this work, we generalize the KT algorithm to speed up supervised learning problems involving kernel methods. Specifically, we combine two classical algorithms—Nadaraya-Watson (NW) regression or kernel smoothing, and kernel ridge regression (KRR)—with KT to provide a *quadratic* speed-up in both training and inference times. We show how distribution compression with KT in each setting reduces to constructing an appropriate kernel, and introduce the Kernel-Thinned NW and Kernel-Thinned KRR estimators. We prove that KT-based regression estimators enjoy significantly superior computational efficiency over the full-data estimators and improved statistical efficiency over i.i.d. subsampling of the training data. En route, we also provide a novel multiplicative error guarantee for compressing with KT. We validate our design choices with both simulations and real data experiments.

1 Introduction

In supervised learning, the goal of coreset methods is to find a representative set of points on which to perform model training and inference. On the other hand, coreset methods in *unsupervised* learning have the goal of finding a representative set of points, which can then be utilized for a broad class of downstream tasks—from integration [10, 9] to non-parametric hypothesis testing [8]. This work aims to bridge these two research threads.

Leveraging recent advancements from compression in the unsupervised setting, we tackle the problem of non-parametric regression (formally defined in Sec. 2). Given a dataset of n i.i.d. samples, $(x_i, y_i)_{i=1}^n$, we want to learn a function f such that $f(x_i) \approx y_i$. The set of allowable functions is determined by the kernel function, which is a powerful building block for capturing complex, non-linear relationships. Due to its powerful performance in practice and closed-form analysis, non-parametric regression methods based on kernels (a.k.a "kernel methods") have become a popular choice for a wide range of supervised learning tasks [13, 19, 23].

There are two popular approaches to non-parametric kernel regression. First, perhaps a more classical approach, is kernel smoothing, also referred to as Nadaraya-Watson (NW) regression. The NW estimator at a point x is effectively a smoothing of labels y_i such that x_i is close to x. These weights are computed using the kernel function (see Sec. 2 for formal definitions). Importantly, the NW estimator takes $\Theta(n)$ pre-processing time (to simply store the data) and $\Theta(n)$ inference time for each test point x (n kernel evaluations and n simple operations).

Another popular approach is kernel ridge regression (KRR), which solves a non-parametric least squares subject to the regression function lying in the reproducing kernel Hilbert space (RKHS) of a

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

specified reproducing kernel function. Remarkably, KRR admits a closed-form solution via inverting the associated kernel matrix, and takes $\mathcal{O}(n^3)$ training time and $\Theta(n)$ inference time for each test point x.

Our goal is to overcome the computational bottlenecks of kernel methods, while retaining their favorable statistical properties. Previous attempts at using coreset methods include the work of Boutsidis et al. [4], Zheng and Phillips [30], Phillips [17], which depend on a projection type compression, having similar spirit to the celebrated Johnson–Lindenstrauss lemma, a metric preserving projection result. So accuracy and running depend unfavorably on the desired statistical error rate. Kpotufe [14] propose an algorithm to reduce the query time of the NW estimator to $\mathcal{O}(\log n)$, but the algorithm requires super-linear preprocessing time.

Other lines of work exploit the structure of kernels more directly, especially in the KRR literature. A slew of techniques from numerical analysis have been developed, including work on Nyström subsampling by El Alaoui and Mahoney [11], Avron et al. [1], Díaz et al. [7]. Camoriano et al. [5] and Rudi et al. [21] combine early stopping with Nyström subsampling. Though more distant from our approach, we also note the approach of Rahimi and Recht [20] using random features, Zhang et al. [29] using Divide-and-Conquer, and Tu et al. [27] using block coordinate descent.

Our contributions. In this work, we show how coreset methods can be used to speed up both training and inference in non-parametric regression for a large class of function classes/kernels. At the heart of these algorithms is a general procedure called kernel thinning [10, 9], which provides a worst-case bound on integration error (suited for problems in the original context of unsupervised learning and MCMC simulations). In Sec. 3, we introduce a meta-algorithm that recovers our two thinned non-parametric regression methods each based on NW and KRR. We introduce the *kernel-thinned Nadaraya-Watson estimator* (KT-NW) and the *kernel-thinned kernel ridge regression* estimator (KT-KRR).

We show that KT-NW requires $\mathcal{O}(n \log^3 n)$ time during training and $\mathcal{O}(\sqrt{n})$ time at inference, while achieving a mean square error (MSE) rate of $n^{-\frac{\beta}{\beta+d}}$ (Thm. 1)—a strict improvement over uniform subsampling of the original input points. We show that KT-KRR requires $\mathcal{O}(n^{3/2})$ time during training and and $\mathcal{O}(\sqrt{n})$ time during inference, while achieving an near-minimax optimal rate of $\frac{m \log n}{n}$ when the kernel has finite dimension (Thm. 2). We show how our KT-KRR guarantees can also be extended to the infinite-dimension setting (Thm. 3). In Sec. 5, we apply our proposed methods to both simulated and real-world data. In line with our theory, KT-NW and KT-KRR outperform standard thinning baselines in terms of accuracy while retaining favorable runtimes.

2 Problem setup

We now formally describe the non-parametric regression problem. Let $x_1, \ldots, x_n$ be i.i.d. samples from the data distribution $\mathbb{P}$ over the domain $\mathcal{X} \subset \mathbb{R}^d$ and $w_1, \ldots, w_n$ be i.i.d. samples from $\mathcal{N}(0, 1)$. Then define the response variables $y_1, \ldots, y_n$ by the follow data generating process:

$$y_i \triangleq f^\star(x_i) + v_i \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{1}$$

where $f^\star : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$ is the *regression function* and $v_i \triangleq \sigma w_i$ for some noise level $\sigma > 0$. Our task is to build an estimate for $f^\star$ given the n observed points, denoted by

$$\mathcal{S}_{\text{in}} \triangleq ((x_1, y_1), \ldots, (x_n, y_n)).$$

Nadaraya-Watson (NW) estimator. A classical approach to estimate the function $f^\star$ is kernel smoothing, where one estimates the function value at a point z using a weighted average of the observed outcomes. The weight for outcome y_i depends on how close x_i is to the point z; let $\kappa : \mathbb{R}^d \to \mathbb{R}$ denote this weighting function such that the weight for x_i is proportional to $\kappa(\|x_i - z\|/h)$ for some bandwidth parameter $h > 0$. Let $\mathbf{k} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ denote a shift-invariant kernel defined as

$$\mathbf{k}(x_1, x_2) = \kappa(\|x_1 - x_2\|/h). \tag{2}$$

Then this smoothing estimator, also known as Nadaraya-Watson (NW) estimator, can be expressed as

$$\widehat{f}(\cdot) \triangleq \frac{\sum_{(x,y) \in \mathcal{S}_{\text{in}}} \mathbf{k}(\cdot, x) y}{\sum_{x \in \mathcal{S}_{\text{in}}} \mathbf{k}(\cdot, x)} \tag{3}$$

whenever the denominator in the above display is non-zero. In the case the denominator in (3) is zero, we can make a default choice, which for simplicity here we choose as zero. We refer to the estimator (3) as FULL-NW estimator hereafter. One can easily note that FULL-NW requires $\mathcal{O}(n)$ storage for the input points and $\mathcal{O}(n)$ kernel queries for inference at each point.

Kernel ridge regression (KRR) estimator. Another popular approach to estimate $f^\star$ is that of non-parametric (regularized) least squares. The solution in this approach, often called as the kernel ridge regression (KRR), is obtained by solving a least squares objective where the fitted function is posited to lie in the RKHS $\mathcal{H}$ of a reproducing kernel $\mathbf{k}$, and a regularization term is added to the objective to avoid overfitting.[1] Overall, the KRR estimate is the solution to the following regularized least-squares objective, where $\lambda > 0$ denotes a regularization hyperparameter:

$$\min_{f \in \mathcal{H}} L_{\mathcal{S}_{\text{in}}} + \lambda \|f\|_{\mathbf{k}}^2, \quad \text{where} \quad L_{\mathcal{S}_{\text{in}}} \triangleq \frac{1}{n} \sum_{(x,y) \in \mathcal{S}_{\text{in}}} (f(x) - y)^2. \tag{4}$$

Like NW, an advantage of KRR is the existence of a closed-form solution

$$\widehat{f}_{\text{full},\lambda}(\cdot) \triangleq \sum_{i=1}^{n} \alpha_i \mathbf{k}(\cdot, x_i) \quad \text{where} \tag{5}$$

$$\boldsymbol{\alpha} \triangleq (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad \mathbf{K} \triangleq [\mathbf{k}(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n}. \tag{6}$$

Notably, the estimate $\widehat{f}_{\text{full},\lambda}$, which we refer to as the FULL-KRR estimator, can also be seen as yet another instance of weighted average of the observed outcomes. Notably, NW estimator imposes that the weights across the points sum to 1 (and are also non-negative whenever $\mathbf{k}$ is), KRR allows for generic weights that need not be positive (even when $\mathbf{k}$ is) and need not sum to 1. We note that naïvely solving $\widehat{f}_{\text{full},\lambda}$ requires $\mathcal{O}(n^2)$ kernel evaluations to compute the kernel matrix, $\mathcal{O}(n^3)$ to compute a matrix inverse, and $\mathcal{O}(n)$ kernel queries for inference at each point. One of our primary goals in this work is to tackle this high computational cost of FULL-KRR.

3 Speeding up non-parametric regression

We begin with a general approach to speed up regression by thinning the input datasets. While computationally superior, a generic approach suffers from a loss of statistical accuracy motivating the need for a strategic thinning approach. To that end, we briefly review kernel thinning and finally introduced our supervised kernel thinning approach.

3.1 Thinned regression estimators: Computational and statistical tradeoffs

Our generic approach comprises two main steps. First, we compress the input data by choosing a coreset $\mathcal{S}_{\text{out}} \subset \mathcal{S}_{\text{in}}$ of size $n_{\text{out}} \triangleq \|\mathcal{S}_{\text{out}}\|$. Second, we apply our off-the-shelf non-parametric regression methods from Sec. 2 to the compressed data. By setting $n_{\text{out}} \ll n$, we can obtain notable speed-ups over the FULL versions of NW and KRR.

Before we introduce the thinned versions of NW and KRR, let us define the following notation. Given an input sequence $\mathcal{S}_{\text{in}}$ and output sequence $\mathcal{S}_{\text{out}}$, define the empirical probability measures

$$\mathbb{P}_{\text{in}} \triangleq \frac{1}{n} \sum_{(x,y) \in \mathcal{S}_{\text{in}}} \delta_{(x,y)} \quad \text{and} \quad \mathbb{Q}_{\text{out}} \triangleq \frac{1}{n_{\text{out}}} \sum_{(x,y) \in \mathcal{S}_{\text{out}}} \delta_{(x,y)}. \tag{7}$$

Thinned NW estimator. The thinned NW estimator is the analog of Full-NW except that $\mathcal{S}_{\text{in}}$ is replaced by $\mathcal{S}_{\text{out}}$ in (3) so that the *thinned-NW estimator* is given by

$$\widehat{f}_{\mathcal{S}_{\text{out}}}(\cdot) \triangleq \frac{\sum_{(x,y) \in \mathcal{S}_{\text{out}}} \mathbf{k}(\cdot, x) y}{\sum_{x \in \mathcal{S}_{\text{out}}} \mathbf{k}(\cdot, x)} = \frac{\mathbb{Q}_{\text{out}}(y\mathbf{k})}{\mathbb{Q}_{\text{out}}\mathbf{k}} \tag{8}$$

[1] We note that while KRR approach (15) does require $\mathbf{k}$ to be reproducing, the NW approach (3) in full generality is valid even when $\mathbf{k}$ is a not a valid reproducing kernel.

whenever the denominator in the display is not zero; and 0 otherwise. When compared to the FULL-NW estimator, we can easily deduce the computational advantage of this estimator: more efficient $\mathcal{O}(n_{\text{out}})$ storage as well as the faster $\mathcal{O}(n_{\text{out}})$ computation for inference at each point.

Thinned KRR estimator. Similarly, we can define the *thinned KRR estimator* as

$$\widehat{f}_{\mathcal{S}_{\text{out}},\lambda'}(\cdot) = \sum_{i=1}^{n_{\text{out}}} \alpha'_i \mathbf{k}(\cdot, x'_i), \quad \text{where} \tag{9}$$

$$\boldsymbol{\alpha}' \triangleq (\mathbf{K}' + n_{\text{out}}\lambda' \mathbf{I}_{n_{\text{out}}})^{-1} \begin{bmatrix} y'_1 \\ \vdots \\ y'_{n_{\text{out}}} \end{bmatrix} \in \mathbb{R}^{n_{\text{out}}} \quad \text{and} \quad \mathbf{K}' \triangleq [\mathbf{k}(x'_i, x'_j)]_{i,j=1}^{n_{\text{out}}} \in \mathbb{R}^{n_{\text{out}} \times n_{\text{out}}}$$

given some regularization parameter $\lambda' > 0$. When compared to FULL-KRR, $\widehat{f}_{\mathcal{S}_{\text{out}},\lambda'}$ has training time $\mathcal{O}(n_{\text{out}}^3)$ and prediction time $\mathcal{O}(n_{\text{out}})$.

A baseline approach is standard thinning, whereby we let $\mathcal{S}_{\text{out}}$ be an i.i.d. sample of $n_{\text{out}} = \sqrt{n}$ points from $\mathcal{S}_{\text{in}}$. For NW, let us call the resulting $\widehat{f}_{\mathcal{S}_{\text{out}}}$ (8) the standard-thinned Nadaraya-Watson (ST-NW) estimator. When $n_{\text{out}} = \sqrt{n}$, ST-NW achieves an excess risk rate of $\mathcal{O}(n^{-\frac{\beta}{2\beta+d}})$ compared to the FULL-NW rate of $\mathcal{O}(n^{-\frac{2\beta}{2\beta+d}})$. For KRR, let us call the resulting $\widehat{f}_{\mathcal{S}_{\text{out}},\lambda'}$ (9) the standard-thinned KRR (ST-KRR) estimator. When $n_{\text{out}} = \sqrt{n}$, ST-KRR achieves an excess risk rate of $\mathcal{O}(\frac{m}{n_{\text{out}}})$ compared to the FULL-KRR rate of $\mathcal{O}(\frac{m}{n})$. Our goal is to provide good computational benefits without trading off statistical error. Moreover, we may be able to do better by leveraging the underlying geometry of the input points and summarize of the input distribution more succinctly than i.i.d. sampling.

3.2 Background on kernel thinning

A subroutine central to our approach is kernel thinning (KT) from Dwivedi and Mackey [10, Alg. 1]. We use a variant called KT-COMPRESS++ from Shetty et al. [22, Ex. 6] (see full details in App. A), which provides similar approximation quality as the original KT algorithm of Dwivedi and Mackey [10, Alg. 1], while reducing the runtime from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log^3 n)$.[2] Given an input kernel $\mathbf{k}_{\text{ALG}}$ and input points $\mathcal{S}_{\text{in}}$, KT-COMPRESS++ outputs a coreset $\mathcal{S}_{\text{KT}} \subset \mathcal{S}_{\text{in}}$ with size $n_{\text{out}} \triangleq \sqrt{n} \ll n$. In this work, we leverage two guarantees of KT-COMPRESS++. Informally, $\mathcal{S}_{\text{KT}}$ satisfies (with high probability):

$$(L^\infty \text{ bound}) \quad \|(\mathbb{P}_{\text{in}} - \mathbb{Q}_{\text{out}})\mathbf{k}_{\text{ALG}}\|_\infty \leq C_1 \frac{\sqrt{d} \log n_{\text{out}}}{n_{\text{out}}} \tag{10}$$

$$(\text{MMD bound}) \quad \sup_{\|h\|_{\mathbf{k}_{\text{ALG}}} \leq 1} |(\mathbb{P}_{\text{in}} - \mathbb{Q}_{\text{out}})h| \leq C_2 \frac{\sqrt{\log n_{\text{out}} \cdot \log \mathcal{N}_{\mathbf{k}_{\text{ALG}}}(\mathcal{B}_2(\mathfrak{R}_{\text{in}}), 1/n_{\text{out}})}}{n_{\text{out}}}, \tag{11}$$

where $C_1, C_2 > 0$ are constants that depend on the properties of the input kernel $\mathbf{k}_{\text{ALG}}$ and the chosen failure probability of KT-COMPRESS++, $\mathfrak{R}_{\text{in}}$ characterizes the radius of $\{x_i\}_{i=1}^n$, and $\mathcal{N}_{\mathbf{k}_{\text{ALG}}}(\mathcal{B}_2(\mathfrak{R}_{\text{in}}), 1/n_{\text{out}})$ denotes the kernel covering number of $\mathcal{H}(\mathbf{k}_{\text{ALG}})$ over the ball $\mathcal{B}_2(\mathfrak{R}_{\text{in}}) \subset \mathbb{R}^d$ at a specified tolerance (see Sec. 4.2 for formal definitions).

At its highest level, KT provides good approximation of function averages. The bound (10) (formally stated in Lem. 1) controls the worst-case point-wise error, and is near-minimax optimal by Phillips and Tai [18, Thm. 3.1]. In the sequel, we leverage this type of result to derive generalization bounds for the kernel smoothing problem. The bound (11) (formally stated in Lem. 2) controls the integration error of functions in $\mathcal{H}(\mathbf{k}_{\text{ALG}})$ and is near-minimax optimal by Tolstikhin et al. [25, Thm. 1, 6]. In the sequel, we leverage this type of result to derive generalization bounds for the KRR problem.

[2] In the sequel, we use "KT" and "KT-COMPRESS++" interchangeably since the underlying algorithm (kernel halving [10, Alg. 1a]) and associated approximation guarantees are the same up to small constant factors.

3.3 Supervised kernel thinning

We show how the approximation results from kernel thinning can be extended to the regression setting. We construct two meta-kernels, the Nadaraya-Watson meta-kernel $\mathbf{k}_{\mathrm{NW}}$ and the ridge-regression meta-kernel $\mathbf{k}_{\mathrm{RR}}$, which take in a *base kernel* $\mathbf{k}$ (defined over $\mathcal{X}$ only) and return a new kernel (defined over $\mathcal{X} \times \mathcal{Y}$). When running KT, we set this new kernel as $\mathbf{k}_{\mathrm{ALG}}$.

3.3.1 Kernel-thinned Nadaraya-Watson regression (KT-NW)

A tempting choice of kernel for KT-NW is the kernel $\mathbf{k}$ itself. That is, we can thin the input points using the kernel

$$\mathbf{k}_{\mathrm{ALG}}((x_1, y_1), (x_2, y_2)) \triangleq \mathbf{k}(x_1, x_2). \tag{12}$$

This choice is sub-optimal since it ignores any information in the response variable y. For our supervised learning set-up, perhaps another intuitive choice would be to use KT with

$$\mathbf{k}_{\mathrm{ALG}}((x_1, y_1), (x_2, y_2)) \triangleq \mathbf{k}(x_1 \oplus y_1, x_2 \oplus y_2), \tag{13}$$

where $\oplus$ denotes vector concatenation (so $x_1 \oplus y_1, x_2 \oplus y_2 \in \mathbb{R}^{d+1}$). While this helps improve performance, there remains a better option as we illustrate next.

In fact, a simple but critical observation immediately reveals a superior choice of the kernel to be used in KT for NW estimator. We can directly observe that the NW estimator is a ratio of the averages of two functions:

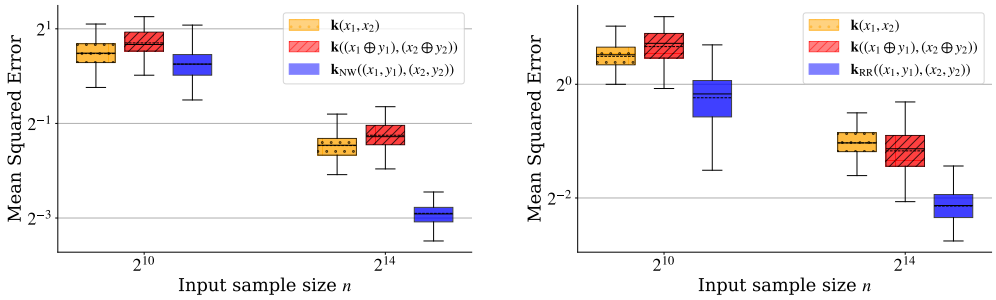
$$f_{\mathrm{numer}}(x, y)(\cdot) \triangleq \mathbf{k}(x, \cdot)\langle y, 1\rangle_{\mathbb{R}}$$
$$\text{and} \quad f_{\mathrm{denom}}(x, y)(\cdot) \triangleq \mathbf{k}(x, \cdot),$$

over the empirical distribution $\mathbb{P}_{\mathrm{in}}$ (7). Recall that KT provides a good approximation of sample means of functions in an RKHS, so it suffices to specify a "correct" choice of the RKHS (or equivalently the "correct" choice of the reproducing kernel). We can verify that f_{denom} lies in the RKHS associated with kernel $\mathbf{k}(x_1, x_2)$ and f_{numer} lies in the RKHS associated with kernel $\mathbf{k}(x_1, x_2) \cdot y_1 y_2$. This motivates our definition for the Nadaraya-Watson kernel:

$$\mathbf{k}_{\mathrm{NW}}((x_1, y_1), (x_2, y_2)) \triangleq \mathbf{k}(x_1, x_2) + \mathbf{k}(x_1, x_2) \cdot y_1 y_2 \tag{14}$$

since then we do have $f_{\mathrm{denom}}, f_{\mathrm{numer}} \in \mathcal{H}(\mathbf{k}_{\mathrm{NW}})$. Intuitively, thinning with $\mathbf{k}_{\mathrm{RR}}$ should simultaneously provide good approximation of averages of f_{denom} and f_{numer} over $\mathbb{P}_{\mathrm{in}}$ (see the formal argument in Sec. 4.1). When $\mathcal{S}_{\mathrm{out}} = \text{KT-COMPRESS++}(\mathcal{S}_{\mathrm{in}}, \mathbf{k}_{\mathrm{NW}}, \delta)$, we call the resulting solution to (8) the kernel-thinned Nadaraya-Watson (KT-NW) estimator, denoted by $\widehat{f}_{\mathrm{KT}}$.

As we show in Fig. 1(a), this theoretically principled choice does provide practical benefits in MSE performance across sample sizes.



(a) NW ablation with Wendland(0) base kernel (b) KRR ablation with Gaussian base kernel

Figure 1: **MSE vs choice of kernels**. For exact settings and further discussion see Sec. 5.1.

3.3.2 Kernel-thinned kernel ridge regression (KT-KRR)

While with NW estimator, the closed-form expression was a ratio of averages, the KRR estimate (5) can not be expressed as an easy function of averages. However, notice that $L_{\mathcal{S}_{\text{in}}}$ in (4) is an average of the function $\ell_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined as

$$\ell_f(x, y) \triangleq f^2(x) - 2f(x)y + y^2 \quad \text{for} \quad f \in \mathcal{H}(\mathbf{k}).$$

Thus, there may be hope of deriving a KT-powered KRR estimator by thinning $L_{\mathcal{S}_{\text{in}}}$ with the appropriate kernel. Assuming $f \in \mathcal{H}(\mathbf{k})$, we can verify that f^2 lies in the RKHS associated with kernel $\mathbf{k}^2(x_1, x_2)$ and that $-2f(x)y$ lies in the RKHS associated with kernel $\mathbf{k}(x_1, x_2) \cdot y_1 y_2$. We now define the ridge regression kernel by

$$\mathbf{k}_{\text{RR}}((x_1, y_1), (x_2, y_2)) \triangleq \mathbf{k}^2(x_1, x_2) + \mathbf{k}(x_1, x_2) \cdot y_1 y_2 \tag{15}$$

and we can verify that $f^2(x) - 2f(x)y$ lies in the RKHS $\mathcal{H}(\mathbf{k}_{\text{RR}})$.[3] When $\mathcal{S}_{\text{out}} \triangleq$ KT-COMPRESS++$(\mathcal{S}_{\text{in}}, \mathbf{k}_{\text{RR}}, \delta)$, we call the resulting solution to (9) the kernel-thinned KRR (KT-KRR) estimator with regularization parameter $\lambda' > 0$, denoted $\widehat{f}_{\text{KT},\lambda'}$. We note that the kernel $\mathbf{k}_{\text{RR}}$ also appears in [12, Lem. 4], except our subsequent analysis comes with generalization bounds for the KT-KRR estimator. Like for NW, in Fig. 1(b) we do a comparison for KRR-MSE across many kernel choices and conclude that the choice (15) is indeed a superior choice compared to the base kernel $\mathbf{k}$ and the concatenated kernel (13).

4 Main results

We derive generalization bounds of our two proposed estimators. In particular, we bound the mean squared error (MSE) defined by $\|f - f^\star\|_2^2 = \mathbb{E}_X\left[(f(X) - f^\star(X))^2\right]$. Our first assumption is that of a well-behaved density on the covariate space. This assumption mainly simplifies our analysis of Nadaraya-Watson and kernel ridge regression, but can in principle be relaxed.

Assumption 1 (Compact support)**.** *Suppose $\mathcal{X} \subset \mathcal{B}_2(\mathfrak{R}_{\text{in}}) \subset \mathbb{R}^d$. Thus, the points $x_1, \ldots, x_n$ are drawn from a distribution with density p that satisfies $0 < p_{\min} \leq p(x) \leq p_{\max}$ for all $x \in \mathcal{X}$.*

4.1 KT-NW

For the analysis of the NW estimator, we define function complexity in terms of Hölder smoothness following prior work [26].

Definition 1. *For $L > 0$ and $\beta \in (0, 2]$, a function $f : \mathcal{X} \to \mathbb{R}$ is (β, L)-Hölder if for all $x_1, x_2 \in \mathcal{X}$,*

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|^\beta \quad \text{if} \quad \beta \in (0, 1] \text{ and}$$
$$|f(x_1) - f(x_2) - \langle \nabla f(x_1), x_2 - x_1 \rangle| \leq L\|x_1 - x_2\|^\beta \quad \text{if} \quad \beta \in (1, 2],$$

where f is assumed to be continuously differentiable for $\beta \in (1, 2]$ but not for $\beta \in (0, 1]$.

Our next assumption is that on the kernel: We require $\mathbf{k}$ to be reproducing kernel to allow for valid analysis for the KT-NW estimator. While typically the NW estimator does not require a reproducing kernel several popular choices in practice, like Gaussian kernel, Laplace kernel, Matérn, Wendland, Sinc, and B-spline kernels, do satisfy this assumption.

Assumption 2 (Shift-invariant kernel)**.** *$\mathbf{k}$ is a shift-invariant (2), reproducing kernel $\mathbf{k}(x_1, x_2) = \kappa(\|x_1 - x_2\|/h)$ such that $h > 0$ and $\kappa : \mathbb{R} \to \mathbb{R}$ is bounded, L_κ-Lipschitz, square-integrable, and decreasing with rate satisfying*

$$\kappa^\dagger(1/n) \cdot h^{-1} = \mathcal{O}(n^\alpha), \quad \text{where} \quad \kappa^\dagger(u) \triangleq \sup\{r : \kappa(r) \geq u\} \quad \text{and} \quad \alpha > 0. \tag{16}$$

Note that this assumption encomposses reproducing kernels with sub-Gaussian, sub-exponential, and poly tails. We now present our main result for the KT-NW estimator.

[3] One might expect the ridge regression kernel to include a term that accounts for y^2. However, the generalization bounds turn out to be essentially the same regardless of whether we include this term when defining $\mathbf{k}_{\text{RR}}$.

Theorem 1 (KT-NW). *Suppose Assum. 1 and 2 hold. Suppose either $f^\star \in \Sigma(\beta, L_f)$ with $\beta \in (0,1]$ or $f^\star \in \Sigma(\beta, L_f)$ for $\beta \in (1,2]$ and $p \in \Sigma(\beta-1, L_p)$, $L_p > 0$ and $2\beta > d$. Then for any fixed $\delta \in (0,1]$, the* KT-NW *estimator (8) with bandwidth $h = n^{-\frac{1}{2\beta+2d}}$ satisfies*

$$\|\widehat{f}_{\mathrm{KT}} - f^\star\|_2^2 \le C n^{-\frac{\beta}{\beta+d}} \log^2 n, \tag{17}$$

with probability at least $1-\delta$, for some positive constant C that does not depend on n.

See App. B for the proof.

Remark 1. *When κ from Assum. 2 has compact support, the condition $2\beta < d$ is not necessary.*

Tsybakov and Tsybakov [26], Belkin et al. [3] show that FULL-NW achieves a rate of $\mathcal{O}(n^{-\frac{2\beta}{2\beta+d}})$, which is minimax optimal for the (β, L)-Hölder function class. Compared to the ST-NW rate of $n^{-\frac{\beta}{2\beta+d}}$, KT-NW achieves strictly better rates for all $\beta > 0$ and $d > 0$, while retaining ST-NW's fast query time of $\mathcal{O}(\sqrt{n})$. Note that our method KT-NW has a training time of $\mathcal{O}(n \log^3 n)$, which is not much more than simply storing the input points.

4.2 KT-KRR

We present our main result for KT-KRR using finite-rank kernels. This class of RKHS includes linear functions, polynomial function classes .

Theorem 2 (KT-KRR for finite-dimensional RKHS). *Assume $f^\star \in \mathcal{H}(\mathbf{k})$, Assum. 1 is satisfied, and $\mathbf{k}$ has rank $m \in \mathbb{N}$. Let $\widehat{f}_{\mathrm{KT},\lambda'}$ denote the* KT-KRR *estimator with regularization parameter $\lambda' = \mathcal{O}(\frac{m \log n_{\mathrm{out}}}{n \wedge n_{\mathrm{out}}^2})$. Then with probability at least $1 - 2\delta - 2e^{-\frac{\|f^\star\|_{\mathbf{k}}^2}{c_1(\|f^\star\|_{\mathbf{k}}^2 + \sigma^2)}}$, the following holds:*

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \le \frac{Cm \cdot \log n_{\mathrm{out}}}{\min(n, n_{\mathrm{out}}^2)} [\|f^\star\|_{\mathbf{k}} + 1]^2 \tag{18}$$

for some constant C that does not depend on n or n_{out}.

See App. C for the proof. Under the same assumptions, Wainwright [28, Ex. 13.19] showed that the Full-KRR estimator $\widehat{f}_{\mathrm{full},\lambda}$ achieves the minimax optimal rate of $O(m/n)$ in $O(n^3)$ runtime. When $n_{\mathrm{out}} = \sqrt{n} \log^c n$, the KT-KRR error rates from Thm. 2 match this minimax rate in $\widetilde{O}(n^{3/2})$ time, a (near) quadratic improvement over the Full-KRR. On the other hand, standard thinning-KRR with similar-sized output achieves a quadratically poor MSE of order $\frac{m}{\sqrt{n}}$.

Our method and theory also extend to the setting of infinte-dimensional kernels. To formalize this, we first introduce the notion of kernel covering number.

Definition 2 (Covering number). *For a kernel $\mathbf{k} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with $\mathcal{B}_{\mathbf{k}} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}$, a set $\mathcal{A} \subset \mathcal{Z}$ and $\epsilon > 0$, the covering number $\mathcal{N}_{\mathbf{k}}(\mathcal{A}, \epsilon)$ is the minimum cardinality of all sets $\mathcal{C} \subset \mathcal{B}_{\mathbf{k}}$ satisfying $\mathcal{B}_{\mathbf{k}} \subset \bigcup_{h \in \mathcal{C}} \{g \in \mathcal{B}_{\mathbf{k}} : \sup_{x \in \mathcal{A}} |h(x) - g(x)| \le \epsilon\}$.*

We consider two general classes of kernels.

Assumption 3. *For some $\mathfrak{C}_d > 0$, all $r > 0$ and $\epsilon \in (0,1)$, and $\mathcal{B}_2(r) = \{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, a kernel $\mathbf{k}$ is*

$\textsc{LogGrowth}(\alpha, \beta)$ *when* $\log \mathcal{N}_{\mathbf{k}}(\mathcal{B}_2(r), \epsilon) \le \mathfrak{C}_d \log(e/\epsilon)^\alpha (r+1)^\beta$ *with* $\alpha, \beta > 0$ *and*

$\textsc{PolyGrowth}(\alpha, \beta)$ *when* $\log \mathcal{N}_{\mathbf{k}}(\mathcal{B}_2(r), \epsilon) \le \mathfrak{C}_d (1/\epsilon)^\alpha (r+1)^\beta$ *with* $\alpha < 2$.

We highlight that the definitions above cover several popular kernels: LOGGROWTH kernels include finite-rank kernels and analytic kernels, like Gaussian, inverse multiquadratic (IMQ), and sinc [9, Prop. 2], while POLYGROWTH kernels includes finitely-many continuously differentiable kernels, like Matérn and B-spline [9, Prop. 3]. For clarity, here we present our guarantee for LOGGROWTH kernels and defer the other case to App. E.

Theorem 3 (KT-KRR guarantee for infinite-dimensional RKHS). *Suppose Assum. 1 is satisfied and $\mathbf{k}$ is* $\textsc{LogGrowth}(\alpha, \beta)$ *(Assum. 3). Then $\widehat{f}_{\mathrm{KT},\lambda'}$ with $\lambda' = \mathcal{O}(1/n_{\mathrm{out}})$ satisfies the following*

bound with probability at least $1 - 2\delta - 2e^{-\frac{\|f^\star\|_{\mathbf{k}}^2 \log^\alpha n}{c_1(\|f^\star\|_{\mathbf{k}}^2+\sigma^2)}}$*:*

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \leq C\Big(\frac{\log^\alpha n}{n} + \frac{\sqrt{\log^\alpha n_{\mathrm{out}}}}{n_{\mathrm{out}}}\Big) \cdot [\|f^\star\|_{\mathbf{k}} + 1]^2. \tag{19}$$

for some constant C *that does not depend on* n *or* n_{out}*.*

See App. E for the proof. When $n_{\mathrm{out}} = \sqrt{n}$, ST-KRR achieves an excess risk rate of $n^{-1/2}\log^\alpha n$ for $\mathbf{k}$ satisfying LogGrowth(α, β). While KT-KRR does not achieve a strictly better excess risk rate bound over ST-KRR, we see that in practice, KT-KRR still obtains an empirical advantage. Obtaining a sharper error rate for the infinite-dimensional kernel setting is an exciting venue for future work.

5 Experimental results

We now present experiments on simulated and real-world data. On real-world data, we compare our KT-KRR estimator with several state-of-the-art KRR methods, including Nyström subsampling-based methods and KRR pre-conditioning methods. All our experiments were run on a machine with 8 CPU cores and 100 GB RAM. Our code can be found at https://github.com/ag2435/npr.

5.1 Simulation studies

We begin with some simulation experiments. For simplicity, let $\mathcal{X} = \mathbb{R}$ and $\mathbb{P} = \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ so that $\mathrm{Var}[X] = 1$. We set

$$f^\star(x) = 8\sin(8\pi x)\exp(x) \quad \text{and} \quad \sigma = 1 \tag{20}$$

and follow (1) to generate $\{y_i\}_{i=1}^n$ (see Fig. 2). We let the input sample size n vary between $2^8, 2^{10}, 2^{12}, 2^{14}$ and always set the output coreset size to be $n_{\mathrm{out}} = \sqrt{n}$. For NW, we use the Wendland(0) kernel defined by

$$\mathbf{k}(x_1, x_2) \triangleq (1 - \tfrac{\|x_1 - x_2\|_2}{h})_+ \quad \text{for} \quad h > 0. \tag{21}$$

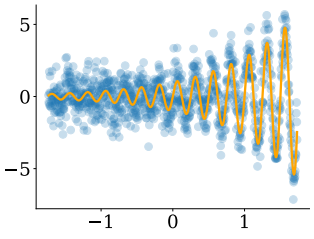


Figure 2: **Simulated data.**

For KRR, we use the Gaussian kernel defined by

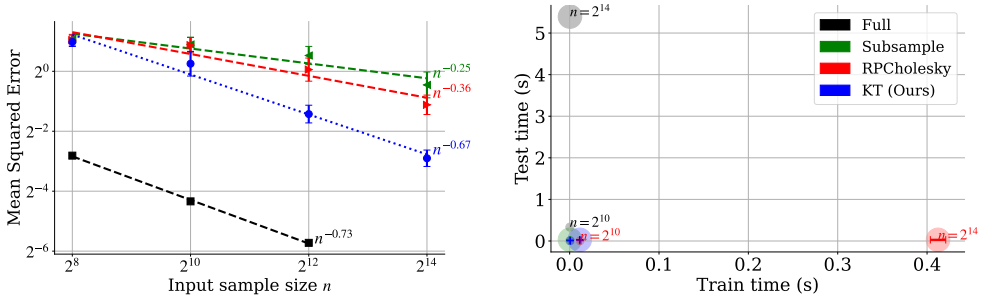
$$\mathbf{k}(x_1, x_2) \triangleq \exp(-\tfrac{\|x_1 - x_2\|_2^2}{2h^2}) \quad \text{for} \quad h > 0. \tag{22}$$

We select the bandwidth h and regularization parameter λ' (for KRR) using grid search. Specifically, we use a held-out validation set of size 10^4 and run each parameter configuration 100 times to estimate the validation MSE since KT-KRR and ST-KRR are random.

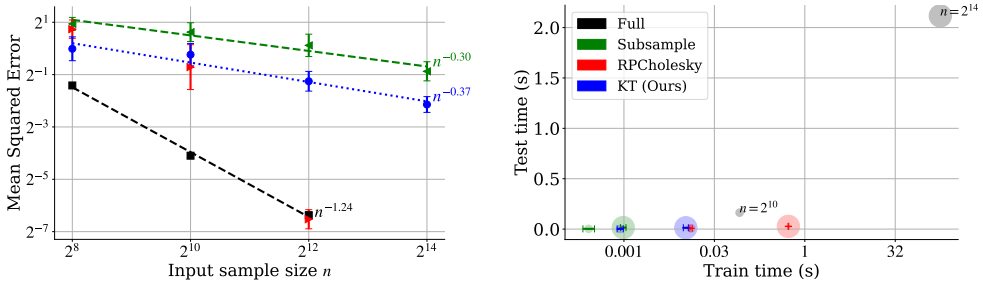
Ablation study. In Fig. 1, we compare thinning with our proposed meta-kernel $\mathbf{k}_{\mathrm{ALG}} = \mathbf{k}_{\mathrm{NW}}$ to thinning with the baseline meta-kernels (12) and (13). For our particular regression function (20), thinning with (12) outperforms thinning with (13). We hypothesize that the latter kernel is not robust to the scaling of the response variables. By inspecting (21), we see that $\|(x_1 \oplus y_1) - (x_2 \oplus y_2)\|_2$ is heavily determined by the y_i values when they are large compared to the values of x_i—as is the case on the right side of Fig. 2 (when $X > 0$). Since $\mathbb{P}$ is a uniform distribution, thinning with (12) evenly subsamples points along the input domain $\mathcal{X}$, even though accuractely learning the left side of Fig. 2 (when $X < 0$) is not needed for effective prediction since it is primarily noise. Validating our theory from Thm. 1, the best performance is obtained when thinning with $\mathbf{k}_{\mathrm{NW}}$ (14), which avoids evenly subsampling points along the input domain and correctly exploits the dependence between X and Y.

In Fig. 1, we perform a similar ablation for KRR. Again we observe that thinning with $\mathbf{k}_{\mathrm{ALG}}((x_1, y_1), (x_2, y_2)) = \mathbf{k}(x_1, x_2)$ outperforms thinning with $\mathbf{k}_{\mathrm{ALG}}((x_1, y_1), (x_2, y_2)) = \mathbf{k}(x_1 \oplus y_1, x_2 \oplus y_2)$, while thinning with $\mathbf{k}_{\mathrm{ALG}} = \mathbf{k}_{\mathrm{RR}}$ achieves the best performance.

Comparison with Full, ST, RPCholesky. In Fig. 3(a), we compare the MSE of KT-NW to Full-NW, ST-NW (a.k.a "Subsample"), and RPCholesky-NW across four values of n. This last method uses the pivot points from RPCholesky as the output coreset $\mathcal{S}_{\mathrm{out}}$. At all n we evaluated, KT-NW achieves lower MSE than ST-NW and RPCholesky-NW. Full-NW achieves the lowest



(a) Nadaraya-Watson estimator with Wendland(0) kernel (21).



(b) Kernel ridge regression estimator with Gaussian kernel (22).

Figure 3: **MSE and runtime comparison on simulated data.** Each point plots the mean and standard deviation across 100 trials (after parameter grid search).

MSE across the board, but it suffers from significantly worse run times, especially at test time. Owing to its $\mathcal{O}(n \log^3 n)$ runtime, KT-NW is significantly faster than RPCHOLESKY-NW for training and nearly matches ST-NW in both training and testing.

In Fig. 3(b), we compare the MSE of KT-KRR to FULL-KRR, ST-KRR (a.k.a "Subsample"), and the RPCHOLESKY-KRR method from Chen et al. [6, Sec. 4.2.2], which uses RPCHOLESKY to select landmark points for the restricted KRR problem. We observe that KT-KRR achieves lower MSE than ST-KRR, but higher MSE than RPCHOLESKY-KRR and FULL-KRR. In Fig. 3(b), we also observe that KT-KRR is orders of magnitude faster than FULL-KRR across the board, with runtime comparable to ST-KRR and RPCHOLESKY-KRR in both training and testing. We hypothesize that RPCHOLESKY—while it provides a good low-rank approximation of the kernel matrix—is not designed to preserve averages.

5.2 Real data experiments

We now move on to our experiments on real-world data using two popular datasets: the California Housing regression dataset from Pace and Barry [16] (https://scikit-learn.org/1.5/datasets/real_world.html#california-housing-dataset; BSD-3-Clause license) and the SUSY binary classification dataset from Baldi et al. [2] (https://archive.ics.uci.edu/dataset/279/susy; CC-BY-4.0 license).

California Housing dataset ($d = 8, N = 2 \times 10^4$)**.** Tab. 1(a) compares the test MSE, train times, and test times. We normalize the input features by subtracting the mean and dividing by the standard deviation and use a 80-20 train-test split. For all methods, we use the Gaussian kernel (22) with bandwidth $h = 10$. We use $\lambda = \lambda' = 10^{-3}$ for FULL-KRR, ST-KRR, and KT-KRR and $\lambda = 10^{-5}$ for RPCHOLESKY-KRR. On this dataset, KT-KRR lies between ST-KRR and RPCHOLESKY-KRR in terms of test MSE. When $n_{\text{out}} = \sqrt{n}$, RPCHOLESKY pivot selection takes $\mathcal{O}(n^2)$ time by Chen et al. [6, Alg. 2], compared to KT-COMPRESS++, which compresses the input points in only $\mathcal{O}(n \log^3 n)$ time. This difference in big-O runtime is reflected in our empirical results, where we see KT-KRR take 0.0153s versus 0.3237s for RPCHOLESKY-KRR.

Method	MSE (%)	Training time (s)	Prediction time (s)
Full	0.4137	11.1095	0.7024
ST-KRR	0.5736 ± 0.0018	0.0018 ± 0.0005	0.0092 ± 0.0006
RPCHOLESKY	0.3503 ± 0.0001	0.3237 ± 0.0094	0.0060 ± 0.0008
KT-KRR (Ours)	0.5580 ± 0.0015	0.0153 ± 0.0013	0.0083 ± 0.0003

(a) California Housing regression dataset.

Method	Test Error (%)	Training Time (s)
RPCholesky	19.99 ± 0.00	3.46 ± 0.03
FALKON	19.99 ± 0.00	5.06 ± 0.02
CG	20.35 ± 0.00	6.16 ± 0.03
ST-KRR	22.71 ± 0.30	0.09 ± 0.00
KT-KRR (Ours)	22.00 ± 0.21	1.79 ± 0.00

(b) SUSY dataset.

Table 1: **Accuracy and runtime comparison on real-world data.** Each cell represents mean ± standard error across 100 trials.

SUSY dataset ($d = 18, N = 5 \times 10^6$)**.** Tab. 1(b) compares our proposed method KT-KRR (with $h = 10, \lambda' = 10^{-1}$) to several large-scale kernel methods, namely RPCholesky preconditioning [7], FALKON [21], and Conjugate Gradient (all with $h = 10, \lambda = 10^{-3}$) in terms of test classification error and training times. For the baseline methods, we use the Matlab implementation provided by Díaz et al. [7] (https://github.com/eepperly/Robust-randomized-preconditioning-for-kernel-ridge-regression). In our experiment, we use 4×10^6 points for training and the remaining 10^6 points for testing. As is common practice for classification tasks, we use the Laplace kernel defined by $\mathbf{k}(x_1, x_2) \triangleq \exp(-\|x_1 - x_2\|_2/h)$. All parameters are chosen with cross-validation.

We observe that KT-KRR achieves test MSE between ST-KRR and RPCHOLESKY preconditioning with training time almost half that of RPCHOLESKY preconditioning. Notably, our Cython implementation of KT-COMPRESS++ thinned the four million training samples in only 1.7 seconds on a single CPU core—with further speed-ups to be gained from parallelizing on a GPU in the future.

6 Conclusions

In this work, we introduce a meta-algorithm for speeding up two estimators from non-parametric regression, namely the Nadaraya-Watson and Kernel Ridge Regression estimators. Our method inherits the favorable computational efficiency of the underlying Kernel Thinning algorithm and stands to benefit from further advancements in unsupervised learning compression methods.

The KT guarantees provided in this work apply only when $f^\star \in \mathcal{H}(\mathbf{k})$ for some base kernel $\mathbf{k}$. In practice, choosing a good kernel $\mathbf{k}$ is indeed a challenge common to all prior work. Our framework is friendly to recent developments in kernel selection to handle this problem: Dwivedi and Mackey [9, Cor. 1] provide integration-error guarantees for KT when $f^\star \notin \mathcal{H}(\mathbf{k})$. Moreover, there are recent results on finding the best kernel (e.g., for hypothesis testing [8, Sec. 4.2]). Radhakrishnan et al. [19] introduce the Recursive Feature Machine, which uses a parameterized kernel $\mathbf{k}_M(x_1, x_2) \triangleq \exp(-(x_1 - x_2)^\top M(x_1 - x_2)/(2h^2))$, and propose an efficient method to learn the matrix parameter M via the average gradient outer product estimator. An exciting future direction would be to combine these parameterized (or "learned") kernels with our proposed KT methods for non-parametric regression.

7 Acknowledgements

AG is supported with funding from the NewYork-Presbyterian Hospital.

References

[1] Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4): 1116–1138, 2017. (Cited on page 2.)

[2] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014. (Cited on page 9.)

[3] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019. (Cited on pages 7 and 21.)

[4] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE transactions on information theory*, 59(10):6880–6892, 2013. (Cited on page 2.)

[5] Raffaello Camoriano, Tomás Angles, Alessandro Rudi, and Lorenzo Rosasco. Nytro: When subsampling meets early stopping. In *Artificial Intelligence and Statistics*, pages 1403–1411. PMLR, 2016. (Cited on page 2.)

[6] Yifan Chen, Ethan N Epperly, Joel A Tropp, and Robert J Webber. Randomly pivoted cholesky: Practical approximation of a kernel matrix with few entry evaluations. *arXiv preprint arXiv:2207.06503*, 2022. (Cited on page 9.)

[7] Mateo Díaz, Ethan N Epperly, Zachary Frangella, Joel A Tropp, and Robert J Webber. Robust, randomized preconditioning for kernel ridge regression. *arXiv preprint arXiv:2304.12465*, 2023. (Cited on pages 2 and 10.)

[8] Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. *arXiv preprint arXiv:2301.05974*, 2023. (Cited on pages 1 and 10.)

[9] Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022. (Cited on pages 1, 2, 7, 10, 14, and 15.)

[10] Raaz Dwivedi and Lester Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25 (152):1–77, 2024. (Cited on pages 1, 2, 4, 13, 14, 18, and 28.)

[11] Ahmed El Alaoui and Michael W Mahoney. Fast randomized kernel methods with statistical guarantees. *stat*, 1050:2, 2014. (Cited on page 2.)

[12] Steffen Grünewälder. Compressed empirical measures (in finite dimensions). *arXiv preprint arXiv:2204.08847*, 2022. (Cited on pages 6 and 41.)

[13] Ming-Yueh Huang and Shu Yang. Robust inference of conditional average treatment effects using dimension reduction. *Statistica Sinica*, 32(Suppl):547, 2022. (Cited on page 1.)

[14] Samory Kpotufe. Fast, smooth and adaptive regression in metric spaces. *Advances in Neural Information Processing Systems*, 22, 2009. (Cited on page 2.)

[15] Lingxiao Li, Raaz Dwivedi, and Lester Mackey. Debiased distribution compression. *arXiv preprint arXiv:2404.12290*, 2024. (Cited on pages 32, 33, and 34.)

[16] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. (Cited on page 9.)

[17] Jeff M Phillips. Coresets and sketches. In *Handbook of discrete and computational geometry*, pages 1269–1288. Chapman and Hall/CRC, 2017. (Cited on page 2.)

[18] Jeff M Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2718–2727. SIAM, 2018. (Cited on page 4.)

[19] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022. (Cited on pages 1 and 10.)

[20] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. (Cited on page 2.)

[21] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. *Advances in neural information processing systems*, 30, 2017. (Cited on pages 2 and 10.)

[22] Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022. (Cited on pages 4, 13, 14, and 15.)

[23] Rahul Singh, Liyuan Xu, and Arthur Gretton. Sequential kernel embedding for mediated and time-varying dose response curves. *arXiv preprint arXiv:2111.03950*, 2021. (Cited on page 1.)

[24] Ingo Steinwart and Simon Fischer. A closer look at covering number bounds for gaussian kernels. *Journal of Complexity*, 62:101513, 2021. (Cited on page 24.)

[25] Ilya Tolstikhin, Bharath K Sriperumbudur, Krikamol Mu, et al. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017. (Cited on page 4.)

[26] Alexandre B Tsybakov and Alexandre B Tsybakov. Nonparametric estimators. *Introduction to Nonparametric Estimation*, pages 1–76, 2009. (Cited on pages 6 and 7.)

[27] Stephen Tu, Rebecca Roelofs, Shivaram Venkataraman, and Benjamin Recht. Large scale kernel learning using block coordinate descent. *arXiv preprint arXiv:1602.05310*, 2016. (Cited on page 2.)

[28] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. (Cited on pages 7, 21, 23, 28, 29, 31, 32, 36, 37, 39, 40, and 41.)

[29] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015. (Cited on page 2.)

[30] Yan Zheng and Jeff M Phillips. Coresets for kernel regression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654, 2017. (Cited on page 2.)

A Background on KT-COMPRESS++

This section details the KT-COMPRESS++ algorithm of Shetty et al. [22, Ex. 6]. In a nutshell, KT-COMPRESS (Alg. 2) takes as input a point sequence of size n, a compression level $\mathfrak{g}$, a (reproducing) kernel functions $\mathbf{k}_{\mathrm{ALG}}$, and a failure probability δ. KT-COMPRESS++ first runs the KT-COMPRESS($\mathfrak{g}$) algorithm of Shetty et al. [22, Ex. 4] to produce an intermediate coreset of size $2^{\mathfrak{g}}\sqrt{n}$. Next, the KT algorithm is run on the intermediate coreset to produce a final output of size $\sqrt{n}$.

KT-COMPRESS proceeds by calling the recursive procedure COMPRESS, which uses KT with kernels $\mathbf{k}_{\mathrm{ALG}}$ as an intermediate halving algorithm. The KT algorithm itself consists of two subroutines: (1) KT-SPLIT (Alg. 3a), which splits a given input point sequence into two equal halves with small approximation error in the $\mathbf{k}_{\mathrm{ALG}}$ reproducing kernel Hilbert space and (2) KT-SWAP (Alg. 3b), which selects the best approximation amongst the KT-SPLIT coresets and a baseline coreset (that simply selects every other point in the sequence) and then iteratively refines the selected coreset by swapping out each element in turn for the non-coreset point that most improves $\mathrm{MMD}_{\mathbf{k}_{\mathrm{ALG}}}$ error. As in Shetty et al. [22, Rem. 3], we symmetrize the output of KT by returning either the KT coreset or its complement with equal probability.

Following Shetty et al. [22, Ex. 6], we always default to $\mathfrak{g} = \lceil \log_2 \log n + 3.1 \rceil$ so that KT-COMPRESS++ has an overall runtime of $\mathcal{O}(n \log^3 n)$. For the sake of simplicity, we drop any dependence on $\mathfrak{g}$ in the main paper.

Algorithm 1: KT-COMPRESS++ – Identify coreset of size $\sqrt{n}$

Input: point sequence $\mathcal{S}_{\mathrm{in}}$ of size n, compression level $\mathfrak{g}$, kernel $\mathbf{k}_{\mathrm{ALG}}$, failure probability δ
$\mathcal{S}_{\mathrm{C}} \leftarrow$ KT-COMPRESS$(\mathfrak{g}, \mathcal{S}_{\mathrm{in}}, \delta)$ // coreset of size $2^{\mathfrak{g}}\sqrt{n}$
$\mathcal{S}_{\mathrm{C++}} \leftarrow \mathrm{KT}(\mathcal{S}_{\mathrm{C}}, \mathbf{k}_{\mathrm{ALG}}, \frac{\mathfrak{g}}{\mathfrak{g}+2^{\mathfrak{g}}(\beta_n+1)}\delta)$ // coreset of size $\sqrt{n}$
return $\mathcal{S}_{\mathrm{C++}}$

Algorithm 2: KT-COMPRESS – Identify coreset of size $2^{\mathfrak{g}}\sqrt{n}$

Input: point sequence $\mathcal{S}_{\mathrm{in}}$ of size n, compression level $\mathfrak{g}$, kernel $\mathbf{k}_{\mathrm{ALG}}$, failure probability δ
return COMPRESS$(\mathcal{S}_{\mathrm{in}}, \mathfrak{g}, \mathbf{k}_{\mathrm{ALG}}, \frac{\delta}{n4^{\mathfrak{g}+1}(\log_4 n - \mathfrak{g})})$

function COMPRESS$(\mathcal{S}, \mathfrak{g}, \mathbf{k}_{\mathrm{ALG}}, \delta)$*:*
  if $|\mathcal{S}| = 4^{\mathfrak{g}}$ **then return** $\mathcal{S}$
  Partition $\mathcal{S}$ into four arbitrary subsequences $\{\mathcal{S}_i\}_{i=1}^4$ each of size $n/4$
  for $i = 1, 2, 3, 4$ **do**
    $\widetilde{\mathcal{S}}_i \leftarrow$ COMPRESS$(\mathcal{S}_i, \mathfrak{g}, \mathbf{k}_{\mathrm{ALG}}, \delta)$ // run COMPRESS recursively to return coresets of size $2^{\mathfrak{g}} \cdot \sqrt{\frac{|\mathcal{S}|}{4}}$
  end
  $\widetilde{\mathcal{S}} \leftarrow$ CONCATENATE$(\widetilde{\mathcal{S}}_1, \widetilde{\mathcal{S}}_2, \widetilde{\mathcal{S}}_3, \widetilde{\mathcal{S}}_4)$ // combine the coresets to obtain a coreset of size $2 \cdot 2^{\mathfrak{g}} \cdot \sqrt{|\mathcal{S}|}$
return $\mathrm{KT}(\widetilde{\mathcal{S}}, \mathbf{k}_{\mathrm{ALG}}, |\widetilde{\mathcal{S}}|^2\delta)$ // halve the coreset to size $2^{\mathfrak{g}}\sqrt{|\mathcal{S}|}$ via symmetrized kernel thinning

function KT$(\mathcal{S}, \mathbf{k}_{\mathrm{ALG}}, \delta)$*:*
  // Identify kernel thinning coreset containing $\lfloor |\mathcal{S}|/2 \rfloor$ input points
  $\mathcal{S}_{\mathrm{KT}} \leftarrow$ KT-SWAP$(\mathbf{k}_{\mathrm{ALG}},$ KT-SPLIT$(\mathbf{k}_{\mathrm{ALG}}, \mathcal{S}, \delta))$
return $\mathcal{S}_{\mathrm{KT}}$ with probability $\frac{1}{2}$ and the complementary coreset $\mathcal{S} \setminus \mathcal{S}_{\mathrm{KT}}$ otherwise

Define the event

$$\mathcal{E}_{\mathrm{KT},\delta} \triangleq \{\text{KT-COMPRESS++ succeeds}\}. \tag{23}$$

Dwivedi and Mackey [10, Thm. 1, Rmk. 4] show that

$$\mathbb{P}(\mathcal{E}_{\mathrm{KT},\delta}) \geq 1 - \delta.$$

We restate [10, Thm. 4] in our notation:

Lemma 1 (L^∞ guarantee for KT-COMPRESS++)**.** *Let $\mathcal{Z} \subset \mathbb{R}^d$ and consider a reproducing kernel $\mathbf{k}_{\mathrm{ALG}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Assume $n/n_{\mathrm{out}} \in 2^{\mathbb{N}}$. Let $\mathcal{S}_{\mathrm{KT}} \triangleq$ KT-COMPRESS++$(\mathcal{S}_{\mathrm{in}}, \mathfrak{g}, \mathbf{k}_{\mathrm{ALG}}, \delta)$ and*

Algorithm 3a: KT-SPLIT $-$ Divide points into candidate coresets of size $\lfloor n/2 \rfloor$

Input: kernel $\mathbf{k}_{\text{split}}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, failure probability δ

$\mathcal{S}^{(1)}, \mathcal{S}^{(2)} \leftarrow \{\}$ // Initialize empty coresets: $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}$ have size i after round i
$\sigma \leftarrow 0$ // Initialize swapping parameter
for $i = 1, 2, \ldots, \lfloor n/2 \rfloor$ **do**
  // Consider two points at a time
  $(x, x') \leftarrow (x_{2i-1}, x_{2i})$
  // Compute swapping threshold $\mathfrak{a}_i$
  $\mathfrak{a}_i, \sigma \leftarrow$ `get_swap_params`$(\sigma, \mathfrak{b}, \frac{\delta}{n})$ with $\mathfrak{b}^2 = \mathbf{k}_{\text{split}}(x, x) + \mathbf{k}_{\text{split}}(x', x') - 2\mathbf{k}_{\text{split}}(x, x')$
  // Assign one point to each coreset after probabilistic swapping
  $\theta \leftarrow \sum_{j=1}^{2i-2} (\mathbf{k}_{\text{split}}(x_j, x) - \mathbf{k}_{\text{split}}(x_j, x')) - 2\sum_{z \in \mathcal{S}^{(1)}} (\mathbf{k}_{\text{split}}(z, x) - \mathbf{k}_{\text{split}}(z, x'))$
  $(x, x') \leftarrow (x', x)$ *with probability* $\min(1, \frac{1}{2}(1 - \frac{\theta}{\mathfrak{a}_i})_+)$
  $\mathcal{S}^{(1)}$`.append`(x); $\mathcal{S}^{(2)}$`.append`(x')
end
return $(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$, candidate coresets of size $\lfloor n/2 \rfloor$

function `get_swap_params`$(\sigma, \mathfrak{b}, \delta)$*:*
  $\mathfrak{a}_i \leftarrow \max(\mathfrak{b}\sigma\sqrt{2\log(2/\delta)}, \mathfrak{b}^2)$
  $\sigma^2 \leftarrow \sigma^2 + \mathfrak{b}^2(1 + (\mathfrak{b}^2 - 2\mathfrak{a}_i)\sigma^2/\mathfrak{a}_i^2)_+$
return $(\mathfrak{a}_i, \sigma)$

Algorithm 3b: KT-SWAP $-$ Identify and refine the best candidate coreset

Input: kernel $\mathbf{k}_{\text{ALG}}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, candidate coresets $(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$

$\mathcal{S}^{(0)} \leftarrow$ `baseline_coreset`$(\mathcal{S}_{\text{in}}, \texttt{size} = \lfloor n/2 \rfloor)$ // Compare to baseline (e.g., standard thinning)

$\mathcal{S}_{\text{KT}} \leftarrow \mathcal{S}^{(\ell^\star)}$ for $\ell^\star \leftarrow \operatorname{argmin}_{\ell \in \{0,1,2\}} \text{MMD}_{\mathbf{k}_{\text{ALG}}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(\ell)})$ // Select best coreset

// Swap out each point in $\mathcal{S}_{\text{KT}}$ for best alternative in $\mathcal{S}_{\text{in}}$ while ensuring no point is repeated in $\mathcal{S}_{\text{KT}}$
for $i = 1, \ldots, \lfloor n/2 \rfloor$ **do**
  $\mathcal{S}_{\text{KT}}[i] \leftarrow \operatorname{argmin}_{z \in \{\mathcal{S}_{\text{KT}}[i]\} \cup (\mathcal{S}_{\text{in}} \setminus \mathcal{S}_{\text{KT}})} \text{MMD}_{\mathbf{k}_{\text{ALG}}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT}} \text{ with } \mathcal{S}_{\text{KT}}[i] = z)$
end
return $\mathcal{S}_{\text{KT}}$, refined coreset of size $\lfloor n/2 \rfloor$

define $\mathbb{P}_{\text{in}} \triangleq \frac{1}{n}\sum_{z \in \mathcal{S}_{\text{in}}} \delta_z$ *and* $\mathbb{Q}_{\text{out}} \triangleq \frac{1}{n_{\text{out}}}\sum_{z \in \mathcal{S}_{\text{KT}}} \delta_z$*. Then on event* $\mathcal{E}_{KT,\delta}$*, the following bound holds:*

$$\|(\mathbb{P}_{\text{in}} - \mathbb{Q}_{\text{out}})\mathbf{k}_{\text{ALG}}\|_\infty \leq c\frac{\|\mathbf{k}_{\text{ALG}}\|_{\infty,\text{in}}}{n_{\text{out}}}\mathfrak{M}_{\mathbf{k}_{\text{ALG}}}(n, n_{\text{out}}, d, \delta, R), \quad \textit{where} \tag{24}$$

$$\mathfrak{M}_{\mathbf{k}_{\text{ALG}}}(n, n_{\text{out}}, d, \delta, R) \triangleq \sqrt{\log\left(\frac{n_{\text{out}}\log_2(n/n_{\text{out}})}{\delta}\right)} \times \left[\sqrt{\log\left(\tfrac{1}{\delta}\right)} + \sqrt{d\log\left(1 + \frac{L_{\mathbf{k}_{\text{ALG}}}}{\|\mathbf{k}_{\text{ALG}}\|_\infty}(R_{\mathbf{k}_{\text{ALG}},n} + R)\right)}\right], \tag{25}$$

$$L_{\mathbf{k}_{\text{ALG}}} \triangleq \sup_{z_1,z_2,z_3 \in \mathcal{Z}} \frac{|\mathbf{k}_{\text{ALG}}(z_1,z_2) - \mathbf{k}_{\text{ALG}}(z_1,z_3)|}{\|z_2 - z_3\|_2}, \quad \textit{and} \tag{26}$$

$$R_{\mathbf{k}_{\text{ALG}},n} \triangleq \inf\left\{r : \sup_{\substack{z_1,z_2 \in \mathcal{Z} \\ \|z_1 - z_2\|_2 \geq r}} |\mathbf{k}_{\text{ALG}}(z_1, z_2)| \leq \frac{\|\mathbf{k}_{\text{ALG}}\|_\infty}{n}\right\}, \tag{27}$$

for some universal positive constant c*.*

Proof. The claim follows by replacing $\mathbf{k}$ [10, Thm. 4] with $\mathbf{k}_{\text{ALG}}$, replacing the sub-Gaussian constant of KT with that of KT-COMPRESS++ in [22, Ex. 5], and replacing $\|\mathbf{k}_{\text{ALG}}\|_\infty$ with $\|\mathbf{k}_{\text{ALG}}\|_{\infty,\text{in}} \triangleq \sup_{z \in \mathcal{S}_{\text{in}}} \mathbf{k}_{\text{ALG}}(z, z)$ throughout. □

We restate [9, Thm. 2] in our notation:

Lemma 2 (MMD guarantee for KT-COMPRESS++). *Let $\mathcal{Z} \subset \mathbb{R}^d$ and consider a reproducing kernel $\mathbf{k}_{\mathrm{ALG}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Assume $n/n_{\mathrm{out}} \in 2^{\mathbb{N}}$. Let $\mathcal{S}_{\mathrm{KT}} \triangleq$ KT-COMPRESS++$(\mathbf{k}_{\mathrm{ALG}}, \mathfrak{g})(\mathcal{S}_{\mathrm{in}})$ and define $\mathbb{P}_{\mathrm{in}} \triangleq \frac{1}{n}\sum_{z\in\mathcal{S}_{\mathrm{in}}} \delta_z$ and $\mathbb{Q}_{\mathrm{out}} \triangleq \frac{1}{n_{\mathrm{out}}}\sum_{z\in\mathcal{S}_{\mathrm{KT}}} \delta_z$. Then on event $\mathcal{E}_{KT,\delta}$, the following bound holds:*

$$\sup_{\substack{h\in\mathcal{H}(\mathbf{k}_{\mathrm{ALG}}):\\ \|h\|_{\mathbf{k}_{\mathrm{ALG}}}\le 1}} |(\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})h| \le \inf_{\substack{\epsilon\in(0,1)\\ \mathcal{S}_{\mathrm{in}}\subset\mathcal{A}}} \left\{ 2\epsilon + \frac{2\|\mathbf{k}_{\mathrm{ALG}}\|^{1/2}_{\infty,\mathrm{in}}}{n_{\mathrm{out}}} \mathfrak{W}_{\mathbf{k}_{\mathrm{ALG}}}(n, n_{\mathrm{out}}, \delta, \mathcal{A}, \epsilon) \right\} \quad \textit{where}$$

$$\mathfrak{W}_{\mathbf{k}_{\mathrm{ALG}}}(n, n_{\mathrm{out}}, \delta, R, \epsilon) \triangleq c\sqrt{\log\left(\tfrac{n_{\mathrm{out}}\log(n/n_{\mathrm{out}})}{\delta}\right) \cdot \left[\log\left(\tfrac{1}{\delta}\right) + \log \mathcal{N}_{\mathbf{k}_{\mathrm{ALG}}}(\mathcal{A}, \epsilon)\right]}. \tag{28}$$

for some universal postive constant c.

Proof. The claim follows from replacing $\mathbf{k}$ in [9, Thm. 2] with $\mathbf{k}_{\mathrm{ALG}}$ and replacing the sub-Gaussian constant of KT with that of KT-COMPRESS++ in [22, Ex. 5]. □

B Proof of Thm. 1: KT-NW

Our primary goal is to bound $\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}[(\widehat{f}_{\mathrm{KT}}(x_0) - f^\star(x_0))^2]$ for a fixed $x_0 \in \mathcal{X}$. Once we have this bound, bounding $\|\widehat{f}_{\mathrm{KT}} - f^\star\|_2^2$ is as straightforward as integrating over $x_0 \in \mathcal{X}$.

Consider the following decomposition:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - f^\star(x_0)\right)^2\right] &= \mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0) + \widehat{f}(x_0) - f^\star(x_0)\right)^2\right] \\
&\le 2\,\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2\right] \qquad (29)\\
&\quad + 2\,\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}(x_0) - f^\star(x_0)\right)^2\right]. \qquad (30)
\end{aligned}$$

Define the random variables

$$\eta_i \triangleq \mathbf{1}\left\{\tfrac{\|X_i - x_0\|}{h} \le 1\right\} \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Also define the event

$$\mathcal{E} \triangleq \{\textstyle\sum_{i=1}^n \eta_i > 0\}. \tag{31}$$

Since X_i are i.i.d. samples from $\mathbb{P}$, it follows that η_i are i.i.d. Bernoulli random variables with parameter

$$\overline{p} \triangleq \mathbb{P}(\eta_i = 1) \ge c_0 p_{\min} h^d, \tag{32}$$

where $c_0 > 0$ depends only on d and κ (see Assum. 2). Denote the denominator terms in $\widehat{f}$ and $\widehat{f}_{\mathrm{KT}}$ by

$$\widehat{p}(\cdot) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbf{k}(\cdot, x_i) \quad \text{and} \quad \widehat{p}_{\mathrm{KT}}(\cdot) \triangleq \tfrac{1}{n_{\mathrm{out}}}\sum_{j=1}^{n_{\mathrm{out}}} \mathbf{k}(\cdot, x'_i), \tag{33}$$

respectively, and the numerator terms in $\widehat{f}$ and $\widehat{f}_{\mathrm{KT}}$ by

$$\widehat{A}(\cdot) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbf{k}(\cdot, x_i) y_i \quad \text{and} \quad \widehat{A}_{\mathrm{KT}}(\cdot) \triangleq \tfrac{1}{n_{\mathrm{out}}}\sum_{j=1}^{n_{\mathrm{out}}} \mathbf{k}(\cdot, x'_i) y'_i, \tag{34}$$

respectively.

We now consider two cases depending on the event $\mathcal{E}$.

Case I: Suppose event $\mathcal{E}^c$ is satisfied. It follows from (33) that $\widehat{p}(x_0) = 0$, in which case $\widehat{f}(x_0) = 0$. Since $\mathcal{S}_{\mathrm{out}} \subset \mathcal{S}_{\mathrm{in}}$, it necessarily follows that $\widehat{p}_{\mathrm{KT}}(x_0) = 0$ and $\widehat{f}_{\mathrm{KT}}(x_0) = 0$. Thus, we can bound (29) and (30) by

$$\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}^c]\right] = 0 \quad \text{and}$$

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_{\text{in}}}\left[\left(\widehat{f}(x_0)-f^\star(x_0)\right)^2\mathbb{I}[\mathcal{E}^c]\right] &= \mathbb{E}_{\mathcal{S}_{\text{in}}}\left[(0-f^\star(x_0))^2\,\mathbb{I}[\mathcal{E}^c]\right]\\ &\le (f^\star)^2(x_0)\mathbb{P}(\mathcal{E}^c)\\ &\le (f^\star)^2(x_0)(1-\overline{p})^n\\ &\le (f^\star)^2(x_0)\exp\{-Cnh^d\}\end{aligned}$$

for some positive constant C that does not depend on n. Note that these are low-order terms compared to the rest of the calculations, so we may ignore them in the final bound.

Case II: Otherwise, we may assume event $\mathcal{E}$ is satisfied. Let us first bound (29). On event $\mathcal{E}_{\text{KT},\delta}$ (23), we claim that

$$\mathbb{E}_{\mathcal{S}_{\text{in}}}\left[\left(\widehat{f}_{\text{KT}}(x_0)-\widehat{f}(x_0)\right)^2\mathbb{I}[\mathcal{E}]\right] \le \tfrac{Cd\log^2 n}{nh^{2d}} \quad \text{whenever} \quad \overline{p}=\omega(\sqrt{\tfrac{d}{n}}\log n). \tag{35}$$

We defer the proof to App. B.1.

Letting $X \triangleq (X_1,\dots,X_n)$ and $Y \triangleq (Y_1,\dots,Y_n)$ denote the x and y components of $\mathcal{S}_{\text{in}}$, respectively, we can further decompose (30) by

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_{\text{in}}}\left[\left(\widehat{f}(x_0)-f^\star(x_0)\right)^2\mathbb{I}[\mathcal{E}]\right] &= \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\left(\widehat{f}(x_0)-\mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right]\right)^2\right]\mathbb{I}[\mathcal{E}]\right]\\ &\quad+\mathbb{E}_X\left[\left(\mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right]-f^\star(x_0)\right)^2\mathbb{I}[\mathcal{E}]\right],\end{aligned}$$

where the first RHS term corresponds to the variance and the second RHS term corresponds to the bias. We claim that

$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\left(\widehat{f}(x_0)-\mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right]\right)^2\right]\mathbb{I}[\mathcal{E}]\right] \le \sigma_\xi^2\left(n\exp\{-Cnh^d\}+\tfrac{C}{nh^d}\right), \tag{36}$$

$$\mathbb{E}_X\left[\left(\mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right]-f^\star(x_0)\right)^2\mathbb{I}[\mathcal{E}]\right] \le C\cdot L_f^2h^{2\beta} \tag{37}$$

for some constant $C>0$ that does not depend on either n or h. We defer the proofs to App. B.2 and B.3.

Combining (35) to (37), we have

$$\mathbb{E}_{\mathcal{S}_{\text{in}}}\left[\left(\widehat{f}_{\text{KT}}(x_0)-f^\star(x_0)\right)^2\mathbb{I}[\mathcal{E}]\right] \le \underbrace{\tfrac{Cd\log^2 n}{nh^{2d}}}_{\text{KT bound}}+\underbrace{2\sigma_\xi^2\left(ne^{-Cnh^d}+\tfrac{C}{nh^d}\right)}_{\text{Variance bound}}+\underbrace{2CL_f^2h^{2\beta}}_{\text{Bias bound}}. \tag{38}$$

Note that $h^d\le 1$, so the $\frac{Cd\log^2 n}{nh^{2d}}$ term dominates the $\frac{C}{nh^d}$ term. Thus, the optimal choice of bandwidth h comes from balancing

$$\tfrac{C}{nh^{2d}} \sim 2L_f^2h^{2\beta} \quad\Longrightarrow\quad h=cn^{-\frac{1}{2\beta+2d}}. \tag{39}$$

Finally, we must verify our growth rate assumption on $\overline{p}$ in (35) is satisfied. Since $\beta>0$, we have

$$\overline{p} \overset{(32)}{\ge} c_0p_{\min}h^d \overset{(39)}{=} c_0'n^{-\frac{d}{2\beta+2d}} \quad\Longrightarrow\quad \lim_{n\to\infty}\frac{\overline{p}}{\sqrt{\frac{d}{n}}\log n}=\infty.$$

Plugging (39) into (38) yields the advertised bound (17).

B.1 Proof of claim (35)

We first provide a generic result for approximating the numerator and denominator terms defined in (33) and (34).

Lemma 3 (Simultaneous L^∞ bound using KT-COMPRESS++ with $\mathbf{k}_{\text{NW}}$)**.** *Suppose* $\mathbf{k}$ *satisfies Assum. 2. Given* $\mathcal{S}_{\text{in}}$*, the following bounds hold on the event* $\mathcal{E}_{KT,\delta}$*:*

$$\|\widehat{p}-\widehat{p}_{\text{KT}}\|_\infty \le c_p\sqrt{\tfrac{d}{n}}(\log n+\log(1/\delta)) \tag{40}$$

$$\|\widehat{A}-\widehat{A}_{\text{KT}}\|_\infty \le c_p\sqrt{\tfrac{d}{n}}(\log n+\log(1/\delta)), \tag{41}$$

where $c_a, c_p>0$ *are constants that do not depend on* d *or* n*.*

See App. B.1.1 for the proof. In the sequel, we will simply treat the $\log(1/\delta)$ term as a constant, meaning the $\log n$ terms dominate in the expressions.

With this lemma in hand, let us prove the claim (35). Define the following events:

$$\mathcal{A} \triangleq \{\widehat{p}_{\mathrm{KT}}(x_0) = 0\} \qquad \mathcal{B} \triangleq \{\widehat{p}_{\mathrm{KT}}(x_0) \neq 0\} \qquad \mathcal{C} \triangleq \left\{\widehat{p}(x_0) \geq \tfrac{\overline{p}}{2}\right\}.$$

On event $\mathcal{E}$, consider the following decomposition:

$$\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta}]\right] = \mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{C}^c]\right] \tag{42}$$

$$+ \mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{A} \cap \mathcal{C}]\right] \tag{43}$$

$$+ \mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{B} \cap \mathcal{C}]\right]. \tag{44}$$

Bounding (42). Note that almost surely, we have

$$|\widehat{f}(x_0)| \leq Y_{\max} \quad \text{and} \quad |\widehat{f}_{\mathrm{KT}}(x_0)| \leq Y_{\max}.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{C}^c]\right] &\leq 4Y_{\max}^2\, \mathbb{P}\Big(n\,\widehat{p}(x_0) < \tfrac{n\overline{p}}{2}\Big) \\
&\overset{(i)}{=} 4Y_{\max}^2 \mathbb{P}\Big(\textstyle\sum_{i=1}^n \eta_i - n\overline{p} < \tfrac{n\overline{p}}{2} - n\overline{p}\Big) \\
&\overset{(ii)}{\leq} c_0 \exp\{-c_1 n h^d\},
\end{aligned}$$

where (i) follows from subtracting $n\overline{p}$ from both sides of the probability statement and (ii) follows from concentration of Bernoulli random variables (see App. B.2).

Bounding (43). Note that on event $\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{C}$, we have

$$\begin{aligned}
\widehat{p}_{\mathrm{KT}}(x_0) &\geq \widehat{p}(x_0) - \|\widehat{p} - \widehat{p}_{\mathrm{KT}}\|_\infty \\
&\overset{(i)}{\geq} \tfrac{\overline{p}}{2} - c_p\sqrt{\tfrac{d}{n}}\log n \\
&\overset{(35)}{\geq} c_1\overline{p} \overset{(32)}{\geq} c_2 p_{\min} h^d > 0.
\end{aligned} \tag{45}$$

where step (i) follows from applying (40) and substituting $\widehat{p} \geq \frac{\overline{p}}{2}$. Hence the events $\mathcal{E}_{\mathrm{KT},\delta}$ and $\mathcal{A} \cap \mathcal{C}$ are mutually exclusive with probability 1, thereby yielding

$$\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{A} \cap \mathcal{C}]\right] = 0.$$

Bounding (44). On the event $\mathcal{B} \cap \mathcal{C}$, we have $\widehat{f}_{\mathrm{KT}}(x_0) = \frac{\widehat{A}_{\mathrm{KT}}(x_0)}{\widehat{p}_{\mathrm{KT}}(x_0)}$ and $\widehat{f}(x_0) = \frac{\widehat{A}(x_0)}{\widehat{p}(x_0)}$, which yields

$$\begin{aligned}
\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0) &= \tfrac{\widehat{A}_{\mathrm{KT}}}{\widehat{p}_{\mathrm{KT}}} - \tfrac{\widehat{A}(x_0)}{\widehat{p}(x_0)} = \tfrac{\widehat{A}_{\mathrm{KT}}(x_0)\cdot\widehat{p}(x_0) - \widehat{A}(x_0)\cdot\widehat{p}_{\mathrm{KT}}(x_0)}{\widehat{p}(x_0)\cdot\widehat{p}_{\mathrm{KT}}(x_0)} \\
&= \tfrac{(\widehat{A}_{\mathrm{KT}}(x_0) - \widehat{A}(x_0))\cdot\widehat{p}(x_0) + \widehat{A}(x_0)\cdot(\widehat{p}(x_0) - \widehat{p}_{\mathrm{KT}}(x_0))}{\widehat{p}(x_0)\cdot\widehat{p}_{\mathrm{KT}}(x_0)} \\
&\leq \tfrac{|\widehat{A}_{\mathrm{KT}}(x_0) - \widehat{A}(x_0)|\cdot\widehat{p}(x_0) + \widehat{A}(x_0)\cdot|\widehat{p}(x_0) - \widehat{p}_{\mathrm{KT}}(x_0)|}{\widehat{p}(x_0)\cdot\widehat{p}_{\mathrm{KT}}(x_0)}
\end{aligned}$$

We can invoke (40) and (41) to bound $|\widehat{p}(x_0) - \widehat{p}_{\mathrm{KT}}(x_0)|$ and $|\widehat{A}_{\mathrm{KT}}(x_0) - \widehat{A}(x_0)|$ respectively. Thus, we have

$$\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0) - \widehat{f}(x_0)\right)^2 \mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{B} \cap \mathcal{C}]\right] \leq \left(\tfrac{c_a\sqrt{\frac{d}{n}}\log n\cdot\widehat{p}(x_0) + c_p\sqrt{\frac{d}{n}}\log n\cdot\widehat{A}(x_0)}{\widehat{p}(x_0)\cdot\widehat{p}_{\mathrm{KT}}(x_0)}\right)^2$$

$$\leq \frac{2d\cdot\log^2 n}{n}\left[\left(\frac{c_a}{\widehat{p}_{\mathrm{KT}}(x_0)}\right)^2+\left(\frac{\widehat{A}(x_0)}{\widehat{p}(x_0)}\right)^2\left(\frac{c_p}{\widehat{p}_{\mathrm{KT}}(x_0)}\right)^2\right]$$
$$\overset{(i)}{\leq} \frac{2d\cdot\log^2 n}{n}\left[\frac{c_a^2+Y_{\max}^2 c_p^2}{\widehat{p}_{\mathrm{KT}}(x_0)}\right]^2$$
$$\overset{(ii)}{\leq} \frac{Cd\log^2 n}{nh^{2d}},$$

for some positive constant C that does not depend on n, where step (i) uses the fact that $\frac{\widehat{A}(x_0)}{\widehat{p}(x_0)} \leq Y_{\max}$ and step (ii) uses the lower bound on $\widehat{p}_{\mathrm{KT}}(x_0)$ from (45). Combining (42) to (44), we have

$$\mathbb{E}_{\mathcal{S}_{\mathrm{in}}}\left[\left(\widehat{f}_{\mathrm{KT}}(x_0)-\widehat{f}(x_0)\right)^2\mathbb{I}[\mathcal{E}_{\mathrm{KT},\delta}]\right] \leq c_0\ \exp\{-c_1 nh^d\}+\frac{Cd\log n}{nh^{2d}}.$$

Note that the second term dominates so that we may drop the first term with slight change to the value of the constant C in the bound (35).

B.1.1 Proof of Lem. 3: Simultaneous L^∞ bound using KT-COMPRESS++ with $\mathbf{k}_{\mathrm{NW}}$

We first decompose $\mathbf{k}_{\mathrm{NW}}$ as

$$\mathbf{k}_{\mathrm{NW}}((x_1,y_1),(x_2,y_2)) = \mathbf{k}_1((x_1,y_1),(x_2,y_2))+\mathbf{k}_2((x_1,y_1),(x_2,y_2)), \quad \text{where}$$
$$\mathbf{k}_1((x_1,y_1),(x_2,y_2)) \triangleq \mathbf{k}(x_1,x_2) \quad \text{and} \tag{46}$$
$$\mathbf{k}_2((x_1,y_1),(x_2,y_2)) \triangleq \mathbf{k}(x_1,x_2)\cdot y_1 y_2. \tag{47}$$

and note that

$$\mathcal{H}(\mathbf{k}_{\mathrm{NW}}) = \mathcal{H}(\mathbf{k}_1)\oplus\mathcal{H}(\mathbf{k}_2). \tag{48}$$

This fact will be useful later for proving simultaneous L^∞ approximation guarantees for $\widehat{A}$ and $\widehat{p}$.

Given that $\mathbf{k}$ satisfies Assum. 2, we want to show that $\mathbf{k}_{\mathrm{NW}}$ defined by (14) satisfies the Lipschitz and tail decay properties, so that we may apply Lem. 1. Note that

$$\|\mathbf{k}_{\mathrm{NW}}\|_\infty = \|\mathbf{k}\|_\infty(1+Y_{\max}^2). \tag{49}$$

We claim that kernel $\mathbf{k}_{\mathrm{NW}}$ satisfies

$$L_{\mathbf{k}_{\mathrm{NW}}} \leq L_{\mathbf{k}} + Y_{\max}(\|\mathbf{k}\|_\infty + L_{\mathbf{k}}Y_{\max}) \quad \text{and} \tag{50}$$
$$R_{\mathbf{k}_{\mathrm{NW}},n} \leq R_{\mathbf{k},n} + 2Y_{\max} \tag{51}$$

By [10, Rmk. 8], we have

$$\frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty} \leq \frac{L_\kappa}{h} \quad \text{and} \quad R_{\mathbf{k},n} \leq h\kappa^\dagger(1/n), \tag{52}$$

where $\kappa^\dagger$ is defined by (16). Applying (49) and (50), we have

$$\frac{L_{\mathbf{k}_{\mathrm{NW}}}}{\|\mathbf{k}_{\mathrm{NW}}\|_\infty} \leq \frac{L_{\mathbf{k}}+Y_{\max}\left(\|\mathbf{k}\|_\infty+L_{\mathbf{k}}Y_{\max}\right)}{\|\mathbf{k}\|_\infty(1+Y_{\max}^2)}$$
$$\leq \frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}+\frac{1}{Y_{\max}}+\frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}$$
$$\overset{(52)}{\leq} \frac{2L_\kappa}{h}+\frac{1}{Y_{\max}}.$$

Finally, we have

$$\frac{L_{\mathbf{k}_{\mathrm{NW}}}R_{\mathbf{k}_{\mathrm{NW}},n}}{\|\mathbf{k}_{\mathrm{NW}}\|_\infty} \leq \left(\frac{2L_\kappa}{h}+\frac{1}{Y_{\max}}\right)\left(h\kappa^\dagger(1/n)+2Y_{\max}\right)$$
$$= 2L_\kappa\kappa^\dagger(1/n)+\frac{4L_\kappa Y_{\max}}{h}+\frac{h\kappa^\dagger(1/n)}{Y_{\max}}+2$$
$$\leq 4\max\{1,L_\kappa Y_{\max}\}\cdot\frac{\kappa^\dagger(1/n)}{h}$$
$$\leq 4\max\{1,L_\kappa Y_{\max}\}\cdot c' n^\alpha, \tag{53}$$

where the last inequality follows from Assum. 2 for some universal positive constant c'.

Since Assum. 1 is satisfied, R is constant. Applying (53) to $\mathfrak{M}_{\mathbf{k}_{\mathrm{NW}}}(n, n_{\mathrm{out}}, d, \delta, R)$ as defined by (25), we have the bound

$$\mathfrak{M}_{\mathbf{k}_{\mathrm{NW}}}(n, n_{\mathrm{out}}, d, \delta, R) \leq c'' \sqrt{\log\left(\tfrac{n_{\mathrm{out}}}{\delta}\right)} \left[\sqrt{\log\left(\tfrac{8}{\delta}\right)} + 5\sqrt{d \log n}\right]$$

for some positive constant c''. Substituting this into (24), we have

$$\begin{aligned}
\left\|(\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})\mathbf{k}_{\mathrm{NW}}\right\|_\infty &\leq c_1 \tfrac{\|\mathbf{k}\|_\infty (1+Y_{\max}^2)}{n_{\mathrm{out}}} \sqrt{\log\left(\tfrac{n_{\mathrm{out}}}{\delta}\right)} \left[\sqrt{\log\left(\tfrac{8}{\delta}\right)} + 5\sqrt{d \log n}\right] \\
&\leq c_2 \tfrac{\|\mathbf{k}\|_\infty (1+Y_{\max}^2)}{n_{\mathrm{out}}} \sqrt{d} (\sqrt{\log n} + \sqrt{\log(1/\delta)})^2,
\end{aligned}$$

for some positive constants c_1, c_2. By definition,

$$\left\|(\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})\mathbf{k}_{\mathrm{NW}}\right\|_\infty = \sup_{z \in \mathcal{Z}} \langle (\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})\mathbf{k}_{\mathrm{NW}}, \mathbf{k}_{\mathrm{NW}}(\cdot, z)\rangle_{\mathbf{k}_{\mathrm{NW}}}.$$

Define $\mathbf{k}_1$ and $\mathbf{k}_2$ by (46) and (47), respectively, and note that $\mathbf{k}_1(\cdot, z), \mathbf{k}_2(\cdot, z) \in \mathcal{H}(\mathbf{k}_{\mathrm{NW}})$ for all $z \in \mathcal{Z}$. We want to show that

$$\begin{aligned}
&\sup_{z \in \mathcal{Z}} \langle (\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})\mathbf{k}_{\mathrm{NW}}, \mathbf{k}_1(\cdot, z)\rangle_{\mathbf{k}_{\mathrm{NW}}} \leq c_2 \tfrac{\|\mathbf{k}\|_\infty (1+Y_{\max}^2)}{n_{\mathrm{out}}} \sqrt{d} (\sqrt{\log n} + \sqrt{\log(1/\delta)})^2 \quad \text{and} \\
&\sup_{z \in \mathcal{Z}} \langle (\mathbb{P}_{\mathrm{in}} - \mathbb{Q}_{\mathrm{out}})\mathbf{k}_{\mathrm{NW}}, \mathbf{k}_2(\cdot, z)\rangle_{\mathbf{k}_{\mathrm{NW}}} \leq c_2 \tfrac{\|\mathbf{k}\|_\infty (1+Y_{\max}^2)}{n_{\mathrm{out}}} \sqrt{d} (\sqrt{\log n} + \sqrt{\log(1/\delta)})^2,
\end{aligned}$$

which would imply (40) and (41) (after simplifying all terms besides n, d, and δ).

The first inequality follows from replacing all occurrences of the test function $\mathbf{k}_{\mathrm{NW}}(\cdot, (x, y))$ in the proof of Lem. 1 with the function $\mathbf{k}_1(\cdot, x)$ and noting that $\langle \mathbf{k}_{\mathrm{NW}}(\cdot, (x_i, y_i)), \mathbf{k}_1(\cdot, (x, y))\rangle_{\mathbf{k}_{\mathrm{NW}}} = \langle \mathbf{k}_1(\cdot, x_i), \mathbf{k}_1(\cdot, x)\rangle_{\mathbf{k}_1}$ from the fact that $\mathcal{H}(\mathbf{k}_{\mathrm{NW}}) = \mathcal{H}(\mathbf{k}_1) \oplus \mathcal{H}(\mathbf{k}_2)$ (48).

The second inequality follows from replacing all occurrences of the test function $\mathbf{k}_{\mathrm{NW}}(\cdot, (x, y))$ in the proof of Lem. 1 with the function $\mathbf{k}_2(\cdot, x)$ and noting that $\langle \mathbf{k}_{\mathrm{NW}}(\cdot, (x_i, y_i)), \mathbf{k}_2(\cdot, (x, y))\rangle_{\mathbf{k}_{\mathrm{NW}}} = \langle \mathbf{k}_2(\cdot, (x_i, y_i)), \mathbf{k}_2(\cdot, (x, y))\rangle_{\mathbf{k}_2}$, again from the fact that $\mathcal{H}(\mathbf{k}_{\mathrm{NW}}) = \mathcal{H}(\mathbf{k}_1) \oplus \mathcal{H}(\mathbf{k}_2)$ (48).

Proof of claim (50). We leverage the fact that the Lipschitz constants defined by (26) satisfies the following additivity property. Letting $\mathcal{Z} = \mathcal{S}_{\mathrm{in}}$, we have

$$\begin{aligned}
L_{\mathbf{k}_{\mathrm{NW}}} &= \sup_{z_1, z_2, z_3 \in \mathcal{Z}} \tfrac{|\mathbf{k}_{\mathrm{NW}}(z_1, z_2) - \mathbf{k}_{\mathrm{NW}}(z_1, z_3)|}{\|z_2 - z_3\|_2} \\
&\leq \sup_{z_1, z_2, z_3 \in \mathcal{Z}} \tfrac{|\mathbf{k}_1(z_1, z_2) - \mathbf{k}_1(z_1, z_3)|}{\|z_2 - z_3\|_2} + \sup_{z_1, z_2, z_3 \in \mathcal{Z}} \tfrac{|\mathbf{k}_2(z_1, z_2) - \mathbf{k}_2(z_1, z_3)|}{\|z_2 - z_3\|_2} \\
&= L_{\mathbf{k}_1} + L_{\mathbf{k}_2}.
\end{aligned}$$

We proceed to bound $L_{\mathbf{k}_1}$ and $L_{\mathbf{k}_2}$ separately. Note that

$$L_{\mathbf{k}_1} = L_{\mathbf{k}}.$$

Applying the definition (26) to $L_{\mathbf{k}_2}$, we have

$$\begin{aligned}
L_{\mathbf{k}_2} &= \sup_{\substack{z_1 = (x_1, y_1) \\ z_2 = (x_2, y_2) \\ z_3 = (x_3, y_3)}} \tfrac{|\mathbf{k}(x_1, x_2) y_1 y_2 - \mathbf{k}(x_1, x_3) y_1 y_3|}{\sqrt{\|x_2 - x_3\|^2 + \|y_2 - y_3\|^2}} \\
&= \sup_{\substack{z_1 = (x_1, y_1) \\ z_2 = (x_2, y_2) \\ z_3 = (x_3, y_3)}} \tfrac{|y_1| \cdot |\mathbf{k}(x_1, x_2) y_2 - \mathbf{k}(x_1, x_2) y_3 + \mathbf{k}(x_1, x_2) y_3 - \mathbf{k}(x_1, x_3) y_3|}{\sqrt{\|x_2 - x_3\|^2 + \|y_2 - y_3\|^2}} \\
&= \sup_{\substack{z_1 = (x_1, y_1) \\ z_2 = (x_2, y_2) \\ z_3 = (x_3, y_3)}} \tfrac{|y_1| \cdot |\mathbf{k}(x_1, x_2)(y_2 - y_3) + (\mathbf{k}(x_1, x_2) - \mathbf{k}(x_1, x_3)) y_3|}{\sqrt{\|x_2 - x_3\|^2 + \|y_2 - y_3\|^2}} \\
&\leq \sup_{\substack{z_1 = (x_1, y_1) \\ z_2 = (x_2, y_2) \\ z_3 = (x_3, y_3)}} \tfrac{|y_1| \cdot |\mathbf{k}(x_1, x_2)(y_2 - y_3)|}{\sqrt{\|x_2 - x_3\|^2 + \|y_2 - y_3\|^2}} + \sup_{\substack{z_1 = (x_1, y_1) \\ z_2 = (x_2, y_2) \\ z_3 = (x_3, y_3)}} \tfrac{|y_1| \cdot |(\mathbf{k}(x_1, x_2) - \mathbf{k}(x_1, x_3)) y_3|}{\sqrt{\|x_2 - x_3\|^2 + \|y_2 - y_3\|^2}} \\
&\leq Y_{\max} \|\mathbf{k}\|_\infty + L_{\mathbf{k}} Y_{\max}^2.
\end{aligned}$$

Putting together the pieces yields the claimed bound.

Proof of claim (51). We aim to show that $R_{\mathbf{k}_{\mathrm{NW}},n}$ is not much larger than $R_{\mathbf{k},n}$. Note that $\mathbf{k}_{\mathrm{NW}}$ can be rewritten as

$$\mathbf{k}_{\mathrm{NW}}((x_1,y_1),(x_2,y_2)) = (1+y_1y_2)\mathbf{k}(x_1,x_2).$$

We define the sets

$$\Gamma \triangleq \left\{ r : \sup_{\substack{x_1,x_2: \\ \|x_1-x_2\|_2 \geq r}} |\mathbf{k}(x_1,x_2)| \leq \tfrac{\|\mathbf{k}\|_\infty}{n} \right\} \quad \text{and} \tag{54}$$

$$\Gamma^\star \triangleq \left\{ r^\star : \sup_{\substack{(x_1,y_1),(x_2,y_2): \\ \|x_1-x_2\|^2+\|y_1-y_2\|^2 \geq (r^\star)^2}} |\mathbf{k}(x_1,x_2)\cdot y_1y_2| \leq \tfrac{\|\mathbf{k}_\star\|_\infty}{n} \right\}, \tag{55}$$

noting that

$$R_{\mathbf{k},n} = \inf \Gamma \quad \text{and} \quad R_{\mathbf{k}_{\mathrm{NW}},n} = \inf \Gamma^\star$$

by definition (27).

Suppose $r \in \Gamma$. Then for any $(x_1,y_1),(x_2,y_2)$ such that

$$\|x_1-x_2\|^2 + \|y_1-y_2\|^2 \geq r^2 + 4Y_{\max}^2,$$

it must follow that $\|x_1-x_2\|^2 \geq r^2$ (since $\|y_1-y_2\|^2 \leq 4Y_{\max}^2$ by triangle inequality). Since r satisfies (54), it must follow that

$$|\mathbf{k}(x_1,x_2)\cdot(y_1y_2+1)| \leq |\mathbf{k}(x_1,x_2)|(Y_{\max}^2+1) \leq \tfrac{\|\mathbf{k}\|_\infty}{n}(Y_{\max}^2+1) \overset{(49)}{=} \tfrac{\|\mathbf{k}_{\mathrm{NW}}\|_\infty}{n},$$

meaning $\sqrt{r^2+4Y_{\max}^2} \in \Gamma^\star$, where recall $\Gamma^\star$ is defined by (55). Thus, we have

$$R_{\mathbf{k}_{\mathrm{NW}},n} \leq \sqrt{R_{\mathbf{k},n}+4Y_{\max}^2} \leq R_{\mathbf{k},n} + 2Y_{\max}$$

as desired.

B.2 Proof of claim (36)

Suppose event $\mathcal{E}$ (31) is satisfied. Define the shorthand for the variance:

$$\sigma^2(x_0;X) \triangleq \mathbb{E}_{Y|X}\left[\left(\widehat{f}(x_0) - \mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right]\right)^2\right].$$

Conditioned on $X_1=x_1, X_2=x_2, \ldots, X_n=x_n$, we have

$$\mathbb{E}_{Y|X}\left[\widehat{f}(x_0)\right] = \mathbb{E}_{Y_1|X_1,\ldots,Y_n|X_n}\left[\frac{\sum_{i=1}^n Y_i\mathbf{k}(X_i,x_0)}{\sum_{i=1}^n \mathbf{k}(X_i,x_0)}\right] = \frac{\sum_{i=1}^n f^\star(X_i)\mathbf{k}(X_i,x_0)}{\sum_{i=1}^n \mathbf{k}(X_i,x_0)}, \tag{56}$$

where we have used the fact that $\mathbb{E}[Y \mid X=\cdot] = f^\star(\cdot)$ by assumption (1).

Note that on event $\mathcal{E}$, we have

$$\begin{aligned}
\sigma^2(x_0;X) &\overset{(56)}{=} \mathbb{E}_{Y|X}\left[\left(\frac{\sum_{i=1}^n Y_i\mathbf{k}(X_i,x_0)}{\sum_{i=1}^n \mathbf{k}(X_i,x_0)} - \frac{\sum_{i=1}^n f^\star(X_i)\mathbf{k}(X_i,x_0)}{\sum_{i=1}^n \mathbf{k}(X_i,x_0)}\right)^2\right] \\
&= \mathbb{E}_{Y|X}\left[\left(\frac{\sum_{i=1}^n v_i\mathbf{k}(X_i,x_0)}{\sum_{i=1}^n \mathbf{k}(X_i,x_0)}\right)^2\right] \\
&= \mathrm{Var}[v_1]\cdot\sum_{i=1}^n \frac{\mathbf{k}(X_i,x_0)^2}{\left(\sum_{i=1}^n \mathbf{k}(X_i,x_0)\right)^2},
\end{aligned}$$

where recall $v_1,\ldots,v_n$ are i.i.d. random variables with $\mathrm{Var}[v_i]=\sigma^2$ by (1). Taking the expectation w.r.t. $X_1,\ldots,X_n$ and leveraging symmetry, we have

$$\mathbb{E}_X\left[\sigma^2(x_0;X)\,\mathbb{I}[\mathcal{E}]\right] = n\sigma^2\cdot\sigma_X^2, \quad \text{where} \quad \sigma_X^2 \triangleq \mathbb{E}_X\left[\frac{\mathbf{k}^2(X_1,x_0)}{\left(\sum_{i=1}^n \mathbf{k}(X_i,x_0)\right)^2}\right]. \tag{57}$$

σ_X^2 can be bounded by

$$\sigma_X^2 \leq \mathbb{E}_X\left[\frac{\mathbf{k}^2(X_1,x_0)}{\left(\sum_{i=1}^n \mathbf{k}(X_i,x_0)\right)^2}\mathbb{I}\left[\sum_{i=1}^n \eta_i \leq \tfrac{n\bar{p}}{2}\right]\right] + \left(\tfrac{2}{n\bar{p}}\right)^2\mathbb{E}_X\left[\mathbf{k}^2(X_1,x_0)\right]$$

$$\overset{(i)}{\leq} \mathbb{P}\Big(\sum_{i=1}^n \eta_i \leq \tfrac{n\overline{p}}{2}\Big) + \Big(\tfrac{2}{n\overline{p}}\Big)^2 \int_{\mathcal{X}} \mathbf{k}^2(x_1, x_0) p(dx_1),$$

where step (i) follows from the fact that $\frac{\mathbf{k}^2(X_1,x_0)}{\left(\sum_{i=1}^n \mathbf{k}(X_i,x_0)\right)^2} \leq 1$. Using Bernstein's inequality [28, Prop. 2.14], the first term can be bounded by

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=1}^n \eta_i \leq \tfrac{n\overline{p}}{2}\Big) &= \mathbb{P}\Big(\sum_{i=1}^n \eta_i - n\overline{p} \leq -\tfrac{n\overline{p}}{2}\Big) \\ &\leq \exp\Big\{-\tfrac{(n\overline{p})^2}{2(n\overline{p}(1-\overline{p})+n\overline{p}/3)}\Big\} \leq \exp\big\{-c_1 n h^d\big\}\end{aligned}$$

for some universal positive constant c. Applying the fact that p is bounded by Assum. 1 and κ is square-integrable by Assum. 2, we can bound the second term by

$$\begin{aligned}\Big(\tfrac{2}{n\overline{p}}\Big)^2 \int_{\mathcal{X}} \mathbf{k}^2(x_1, x_0) p(dx_1) &\overset{(i)}{\leq} \Big(\tfrac{2}{n\overline{p}}\Big)^2 p_{\max} h^d \int_{\mathbb{R}^d} \kappa^2(u) du \\ &\overset{(ii)}{\leq} \tfrac{4}{(n\overline{p})^2}\big(c_1 h^d\big) \leq \tfrac{c_2}{n^2 h^d},\end{aligned}$$

for some positive constant c_2 that does not depend on either h or n. Substituting these expressions into (57) yields a bound on $\mathbb{E}_X\big[\sigma^2(x_0; X)\, \mathbb{I}[\mathcal{E}]\big]$ as desired.

B.3 Proof of claim (37)

Define the following shorthand for the bias:

$$b(x_0; X) \triangleq \mathbb{E}_{Y|X}\Big[\widehat{f}(x_0)\Big] - f^\star(x_0).$$

We state a more detailed version of the claim.

Lemma 4 (Bias of Nadaraya-Watson)**.** *Suppose Assum. 1 and 2 are satisfied and the event $\mathcal{E}$ (31) holds.*

If $f^\star \in \Sigma(\beta, L_f)$ for $\beta \in (0, 1]$, $L_f > 0$, then the following statements hold true for any $x_0 \in \mathcal{X}$:

- *(I.a) If $\mathbf{k}$ is compactly supported, then $b^2(x_0; X) \leq L_f^2 h^{2\beta}$.*
- *(I.b) If $\mathbf{k}$ has tail decay satisfying (16), then $b^2(x_0; X) \leq c \cdot L_f^2 (h^{2\beta} \vee h^{2\beta - d})$ for some positive constant c.*

Now suppose $f^\star \in \Sigma(\beta, L_f)$ for $\beta \in (1, 2]$, $L_f > 0$ and the density p of the marginal distribution of X satisfies $p \in \Sigma(\beta - 1, L_p)$, $L_p > 0$. Let σ_X be defined by (57). Then the following statements hold for any $x_0 \in \mathcal{X}$:

- *(II.a) If $\mathbf{k}$ is compactly supported, then $b^2(x_0; X) \leq \big(L_f + \|\nabla f(x_0)\| L_p p_{\min}^{-1}\big) h^{2\beta} + \sigma_X^2$.*
- *(II.b) If $\mathbf{k}$ has tail decay satisfying (16), then $b^2(x_0; X) \leq c \cdot \big(L_f + \|\nabla f(x_0)\| L_p p_{\min}^{-1}\big)(h^{2\beta} \vee h^{2\beta - d}) + \sigma_X^2$ for some positive constant c.*

We proceed to prove Case I and Case II in App. B.3.1 and App. B.3.2, respectively.

B.3.1 Bias when $\beta \in (0, 1]$

Proof of claim (I.a). For completeness, we state the proof from Belkin et al. [3, Lem. 2]. On the event $\mathcal{E}$, we have

$$b(x_0; X) \overset{(56)}{=} \frac{\sum_{i=1}^n (f^\star(X_i) - f^\star(x_0)) \mathbf{k}(X_i, x_0)}{\sum_{i=1}^n \mathbf{k}(X_i, x_0)} \overset{(i)}{\leq} \frac{\sum_{i=1}^n L_f \|X_i\|^\beta \mathbf{k}(X_i, x_0)}{\sum_{i=1}^n \mathbf{k}(X_i, x_0)} \overset{(ii)}{\leq} L_f h^\beta, \tag{58}$$

where step (i) follows from our assumption that $f^\star \in \Sigma(\beta, L_f)$ and step (ii) follows from our assumption that $\mathbf{k}$ is compactly supported, so $\mathbf{k}(x, x_0) = 0$ whenever $\|x - x_0\| > h$.

Proof of claim (I.b). Consider the following decomposition of the bias:

$$b(x_0;X) \le \frac{\sum_{i=1}^n \mathbf{k}(X_i,x_0)\mathbb{I}[\|X_i-x_0\|\le h](f^\star(X_i)-f^\star(x_0))}{\sum_{k=1}^n \mathbf{k}(X_k,x_0)\mathbb{I}[\|X_i-x_0\|\le h]} + \frac{\sum_{i=1}^n \mathbf{k}(X_i,x_0)\mathbb{I}[\|X_i-x_0\|> h](f^\star(X_i)-f^\star(x_0))}{\sum_{k=1}^n \mathbf{k}(X_k,x_0)}.$$

Note that the first RHS term can be bounded using (58). To bound the second RHS term, we introduce the following high-probability event

$$\overline{\mathcal{E}} \triangleq \{|\mathcal{I}| = O(nh^{d/2})\}, \quad \text{where} \quad \mathcal{I} \triangleq \{i \in [n] : \|X_i - x_0\| = O(h^{1/2})\}.$$

On this event, we may apply the Lipschitz property of $f^\star$ again to obtain

$$\begin{aligned}\frac{\sum_{i=1}^n \mathbf{k}(X_i,x_0)\mathbb{I}[\|X_i-x_0\|>h](f(X_i)-f(x_0))}{\sum_{k=1}^n \mathbf{k}(X_k,x_0)} &\le \frac{L_f\sum_{i=1}^n \mathbf{k}(X_i,x_0)\|X_i-x_0\|^\beta}{\sum_{k\in\mathcal{I}} \mathbf{k}(X_k,x_0)} \\ &\overset{(i)}{\le} \frac{L_f\sum_{i=1}^n \kappa(\frac{\|X_i-x_0\|}{h})\|X_i-x_0\|^\beta}{\sum_{k\in\mathcal{I}} \mathbf{k}(X_k,x_0)} \\ &\le \frac{L_f n \max_{u\ge 0}\kappa(u)(uh)^\beta}{\sum_{k\in\mathcal{I}} \mathbf{k}(X_k,x_0)},\end{aligned}$$

where step (i) follows from the fact that $\mathbf{k}$ is a shift-invariant kernel by Assum. 2. Note that when κ has tail decay satisfying (16), we have

$$\max_{x\ge 0}\kappa(u)(uh)^\beta = c\cdot h^\beta$$

for some positive constant c. This implies on the event $\overline{\mathcal{E}}$, the following bound holds

$$b(x_0;X) \le L_f h^\beta + L_f\cdot c\cdot \frac{nh^\beta}{nh^{d/2}} = c\cdot L_f(h^\beta \vee h^{\beta-d/2}).$$

B.3.2 Bias when $\beta \in (1,2]$

Proof of claim (II.a). For notational simplicity, let $x_0 = 0$. We further assume the density p of X satisfies (i) $p \in [p_\ell, p_u]$ for some $0 < p_\ell < p_u < \infty$ and $p \in \Sigma(\beta-1, L_p)$. Define $B \triangleq \|\nabla f(0)\|^2$ and denote $\theta \triangleq (L_p, L_f, p_\ell, p_u, B)$ to be the collection of parameters. Constants $c > 0$ in this proof may differ from line to line, but depends only up to parameters in θ.

By definition,

$$b^2(0) \triangleq \mathbb{E}\Big[\sum_{i,j=1}^n G_iG_j\mathbb{I}[\mathcal{E}_0]\Big]$$

where

$$G_i = \frac{(f(X_i)-f(0))\mathbf{k}(X_i)}{\sum_{k=1}^n \mathbf{k}(X_k)}.$$

Further define

$$A(X_i,X_j) \triangleq \frac{\mathbf{k}(X_i)\mathbf{k}(X_j)}{(\sum_{k=1}^n \mathbf{k}(X_k))^2}\mathbb{I}[\mathcal{E}_0] \ge 0 \quad \text{for} \quad i \ne j,$$

and denote $\mathbb{E}_{i,j}$ to be the conditional expectation of law X_i, X_j, while fixing all other randomness $X_k, k \ne i, k \ne j$. Under independence of X_i, $i \in [n]$, it is safe to say that $\mathbb{E}_{i,j}$ is the expectation of marginal law of X_i, X_j, while fixing other randomness as constants.

Define

$$R(x) = f(x) - f(0) - \langle \nabla f(0), x\rangle$$

so that by simple add-substract algebra, we have

$$\begin{aligned}\mathbb{E}_{i,j}[G_iG_j\mathbb{I}[\mathcal{E}_0]] &= \iint \langle\nabla f(0),x_i\rangle\langle\nabla f(0),x_j\rangle A(x_i,x_j)p(x_i)p(x_j)dx_idx_j \\ &\quad + 2\iint \langle\nabla f(0),x_i\rangle R(x_j)A(x_i,x_j)p(x_i)p(x_j)dx_idx_j \\ &\quad + \iint R(x_i)R(x_j)A(x_i,x_j)p(x_i)p(x_j)dx_idx_j.\end{aligned}$$

Observe that $A(x_i,x_j)$ are even functions for both arguments, and $x \mapsto \langle\nabla f(0),x\rangle$ is an odd function. So we can see

$$\int \langle\nabla f(0),x_i\rangle A(x_i,x_j)p(0)dx_i = 0.$$

Such observation allows us to write

$$\begin{aligned}\mathbb{E}_{i,j}[G_iG_j\mathbb{I}[\mathcal{E}_0]] &= \iint\langle\nabla f(0),x_i\rangle\langle\nabla f(0),x_j\rangle A(x_i,x_j)(p(x_i)-p(0))(p(x_j)-p(0))dx_idx_j\\ &+2\iint\langle\nabla f(0),x_i\rangle R(x_j)A(x_i,x_j)(p(x_i)-p(0))p(x_j)dx_idx_j\\ &+\iint R(x_i)R(x_j)A(x_i,x_j)p(x_i)p(x_j)dx_idx_j.\end{aligned}$$

From the assumptions, we see $|R(x_i)|\le L_f\|x_i\|^\beta$ so that we have

$$\begin{aligned}\mathbb{E}_{i,j}[G_iG_j\mathbb{I}[\mathcal{E}_0]] &\le c\iint\|x_i\|^\beta\|x_j\|^\beta A(x_i,x_j)dx_idx_j\\ &+c\iint\|x_i\|^\beta\|x_j\|^\beta A(x_i,x_j)p(x_i)dx_idx_j\\ &+c\iint\|x_i\|^\beta\|x_j\|^\beta A(x_i,x_j)p(x_i)p(x_j)dx_idx_j\\ &\le c\iint\|x_i\|^\beta\|x_j\|^\beta A(x_i,x_j)dx_idx_j.\end{aligned}$$

So overall we have

$$\mathbb{E}[G_iG_j\mathbb{I}[\mathcal{E}_0]]\le c\mathbb{E}\Big[\tfrac{\|Z_i\|^\beta\|Z_j\|^\beta\mathbf{k}(Z_i)\mathbf{k}(Z_j)}{(\sum_{k=1}^n\mathbf{k}(Z_k))^2}\mathbb{I}[\mathcal{E}_0]\Big],$$

where Z_i's are independent and uniform measures on the domain — such integration is established by pulling out all the p_u, p_ℓ for all p's of X_i's. So

$$b^2(0)\le c\mathbb{E}\left[\left(\sum_{i=1}^n\tfrac{|Z_i|^\beta\mathbf{k}(Z_i)}{\sum_{k=1}^n\mathbf{k}(Z_k)}\right)^2\mathbb{I}[\mathcal{E}_0]\right].$$

Now let's induce the same reasoning as done in $\beta\in(0,1]$ case, so that

$$b^2(0)\le O(h^{2\beta})\vee O(h^{2\beta-d}).$$

Proof of claim (II.b). The proof follows by similar logic as (II.a) combined with the truncation argument of (I.b).

C Proof of Thm. 2: KT-KRR for finite-dimensional RKHS

We rely on the localized Gaussian/Rademacher analysis of KRR from prior work [28]. Define the *Gaussian critical radius* $\varepsilon_n>0$ to be the smallest positive solution to the inequality

$$\widehat{\mathcal{G}}_n(\varepsilon;\mathcal{B}_\mathcal{H}(3))\le\tfrac{R}{2\sigma}\varepsilon^2,\quad\text{where}\quad\widehat{\mathcal{G}}_n(\varepsilon;\mathcal{F})\triangleq\mathbb{E}_w\left[\sup_{\substack{f\in\mathcal{F}:\\\|f\|_n\le\varepsilon}}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n w_if(x_i)\right|\right],\tag{59}$$

$\mathcal{B}_\mathcal{H}(3)$ is the $\|\cdot\|_\mathbf{k}$-ball of radius 3 and $w_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$.

Assumption 4. *Assume that $\|f^\star\|_\mathbf{k}\in\mathcal{B}_\mathbf{k}(R)$ and $\widehat{f}_{\mathrm{KT},\lambda'}\in\mathcal{B}_\mathbf{k}(c_\dagger R)$, for some constant $c_\dagger>0$.*

Note that for any $g\in\mathcal{B}_\mathbf{k}(c_\dagger R)$, we have

$$\|g\|_\infty\le\sup_{x\in\mathcal{X}}\langle g,\mathbf{k}(\cdot,x)\rangle_\mathbf{k}\overset{(i)}{\le}\sup_{x\in\mathcal{X}}\|g\|_\mathbf{k}\|\mathbf{k}(\cdot,x)\|_\mathbf{k}\le\|g\|_\mathbf{k}\sqrt{\|\mathbf{k}\|_\infty}\le c_\dagger R\sqrt{\|\mathbf{k}\|_\infty}\triangleq B,$$

where step (i) follows from Cauchy-Schwarz. Thus, the function class $\mathcal{B}_\mathbf{k}(c_\dagger R)$ is B-uniformly bounded. Now define the *Rademacher critical radius* $\delta_n>0$ to be the smallest positive solution to the inequality

$$\mathcal{R}_n(\delta;\mathcal{H})\le\delta^2,\quad\text{where}\quad\mathcal{R}_n(\delta;\mathcal{F})\triangleq\mathbb{E}_{x,\nu}\left[\sup_{\substack{f\in\mathcal{F}:\\\|f\|_2\le\delta}}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n\nu_if(x_i)\right|\right]\tag{60}$$

and $\nu_i=\pm1$ each with probability $1/2$.

Finally, we use the following shorthand to control the KT approximation error term,

$$\eta_{n,\mathbf{k}}\triangleq\tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2+\mathfrak{W}_\mathbf{k}(n,n_{\text{out}},\delta,\mathfrak{R}_{\text{in}},\tfrac{\mathfrak{a}}{n_{\text{out}}})),\quad\text{where}\tag{61}$$

$$\mathfrak{R}_{\text{in}}\triangleq\max_{x\in\mathcal{S}_{\text{in}}}\|x\|_2\quad\text{and}\quad\mathfrak{a}\triangleq\|\mathbf{k}\|_{\infty,\text{in}}+Y_{\max}^2\tag{62}$$

and $\mathfrak{W}_\mathbf{k}$ is an *inflaction factor* defined in (28) that scales with the covering number $\mathcal{N}_\mathbf{k}$ (see Def. 2). With these definitions in place, we are ready to state a detailed version of Thm. 2:

Theorem 4 (KT-KRR for finite-dimensional RKHS, detailed)**.** *Suppose the kernel operator associated with* $\mathbf{k}$ *and* $\mathbb{P}$ *has eigenvalues* $\mu_1 \geq \ldots \geq \mu_m > 0$ *(by Mercer's theorem). Define* $C_m \triangleq 1/\mu_m$*. Let* ε_n *and* δ_n *denote the solutions to* (60) *and* (59)*, respectively. Further assume* [4]

$$n\delta_n^2 > \log(4\log(1/\delta_n)). \tag{63}$$

Let $\widehat{f}_{\mathrm{KT},\lambda'}$ *denote the* KT-KRR *estimator with regularization parameter*

$$\lambda' \geq 2\xi_n^2 \quad \textit{where} \quad \xi_n \triangleq \varepsilon_n \vee \delta_n \vee 4\sqrt{C_m}(\|f^\star\|_{\mathbf{k}} + 1)\eta_{n,\mathbf{k}}.$$

Then with probability at least $1 - 2\delta - 2e^{-\frac{n\delta_n^2}{c_1(b^2+\sigma^2)}}$*, we have*

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \leq c\big\{\xi_n^2 + \lambda'\big\}\|f^\star\|_{\mathbf{k}}^2 + c\delta_n^2. \tag{64}$$

where recall δ *is the success probability of* KT-COMPRESS++ (23).

See App. D for the proof. We set $\lambda' = 2\xi_n^2$, so that (64) becomes

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \leq 3c\xi_n^2\|f^\star\|_{\mathbf{k}}^2 + c\delta_n^2. \tag{65}$$

It remains to bound the quantities ε_n (59), δ_n (60), and $\eta_{n,\mathbf{k}}$ (61). We claim that

$$\varepsilon_n \leq c_0 \tfrac{\sigma}{R}\sqrt{\tfrac{m}{n}} \tag{66}$$

$$\delta_n \leq c_1 b\sqrt{\tfrac{m}{n}} \tag{67}$$

$$\eta_{n,\mathbf{k}} \leq c_2 \tfrac{\sqrt{m\cdot\log n_{\mathrm{out}}\cdot\log(1/\delta)}}{n_{\mathrm{out}}}. \tag{68}$$

for some universal positive constants c_0, c_1, c_2. Now set

$$R = \|f^\star\|_{\mathbf{k}} \qquad \delta = e^{-1/R^4}.$$

Thus, we have

$$\xi_n \leq c'\big(\tfrac{\sigma}{\|f^\star\|_{\mathbf{k}}} \vee b \vee \tfrac{4\sqrt{C_m}}{\|f^\star\|_{\mathbf{k}}}\big)\tfrac{\sqrt{m}}{\sqrt{n}\wedge n_{\mathrm{out}}}.$$

for some universal positive constant c'. Substituting this into (65) leads to the advertised bound (18).

Proof of claim (66). For finite rank kernels, $\hat{\mu}_j = 0$ for $j > m$. Thus, we have $\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^n \min\{\varepsilon^2, \hat{\mu}_j\}} = \sqrt{\frac{2}{n}}\sqrt{m\varepsilon^2}$. From the critical radius condition (59), we want $\sqrt{\frac{2}{n}}\sqrt{m\varepsilon^2} \leq \frac{R}{4\sigma}\varepsilon^2$, so we may set $\varepsilon_n \simeq \frac{\sigma}{R}\sqrt{\frac{m}{n}}$.

Proof of claim (67). By similar logic as above, we have $\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^n \min\{\delta^2, \mu_j\}} = \sqrt{\frac{2}{n}}\sqrt{m\delta^2}$. From the critical radius condition (60), we want $\sqrt{\frac{2}{n}}\sqrt{m\delta^2} \leq \frac{1}{b}\delta^2$, so we may set $\delta_n \simeq b\sqrt{2}\sqrt{\frac{m}{n}}$.

Proof of claim (68). Consider the linear operator $T : \mathcal{H} \to \mathbb{R}^m$ that maps a function to the coefficients in the vector space spanned by $\{\phi_i\}_{i=1}^m$. Note that

$$\|T\| = \tfrac{\|Tf\|_\infty}{\|f\|_{\mathbf{k}}} \leq \sqrt{\|\mathbf{k}\|_\infty}$$

Since the image of T has dimension m, we have $\mathrm{rank}(T) \leq m$. Moreover, $\|\mathbf{k}\|_\infty \leq \mu_1 \cdot \mathfrak{R}_{\mathrm{in}}^2$. Now we can invoke [24, Eq. 14] with $\epsilon = \mathfrak{a}/n_{\mathrm{out}}$ to obtain

$$\mathcal{N}_{\mathbf{k}}(\mathcal{B}_2^d(\mathfrak{R}_{\mathrm{in}}), \mathfrak{a}/n_{\mathrm{out}}) \leq \mathcal{N}(T, \mathfrak{a}/n_{\mathrm{out}}) \leq (1 + \mu_1\mathfrak{R}_{\mathrm{in}}^2 n_{\mathrm{out}}/\mathfrak{a})^m.$$

Taking the log on both sides and substituting this bound into (61), we have

$$\begin{aligned}
\eta_{n,\mathbf{k}} &= \tfrac{\mathfrak{a}^2}{n_{\mathrm{out}}}\big(2 + \mathfrak{W}_{\mathbf{k}}(n, n_{\mathrm{out}}, \delta, \mathfrak{R}_{\mathrm{in}}, \tfrac{\mathfrak{a}}{n_{\mathrm{out}}})\big) \\
&\leq \tfrac{\mathfrak{a}^2}{n_{\mathrm{out}}}\Big(2 + \sqrt{\log\Big(\tfrac{n_{\mathrm{out}}\log(n/n_{\mathrm{out}})}{\delta}\Big)\cdot\Big[\log\big(\tfrac{1}{\delta}\big) + \log\mathcal{N}_{\mathbf{k}}(\mathcal{B}_2^d(\mathfrak{R}_{\mathrm{in}}), \tfrac{\mathfrak{a}}{n_{\mathrm{out}}})\Big]}\Big) \\
&\leq \tfrac{\mathfrak{a}^2}{n_{\mathrm{out}}}\Big(2 + \sqrt{\log\Big(\tfrac{n_{\mathrm{out}}\log(n/n_{\mathrm{out}})}{\delta}\Big)\cdot\Big[\log\big(\tfrac{1}{\delta}\big) + m\log\Big(1 + \tfrac{2\|T\|n_{\mathrm{out}}}{\mathfrak{a}}\Big)\Big]}\Big) \\
&\leq c\tfrac{\sqrt{m\cdot\log n_{\mathrm{out}}\cdot\log(1/\delta)}}{n_{\mathrm{out}}}
\end{aligned}$$

for some positive constant c that doesn't depend on $m, n_{\mathrm{out}}, \delta$.

[4] Note that when $\mathbf{k}$ is finite-rank, this condition is automatically satisfied.

D Proof of Thm. 4: KT-KRR for finite-dimensional RKHS, detailed

We rescale our observation model (1) by $\|f^\star\|_{\mathbf{k}}$, so that the noise variance is $(\sigma/\|f^\star\|_{\mathbf{k}})^2$ and our new regression function satisfies $\|f^\star\|_{\mathbf{k}} = 1$. Our final prediction error should then be multiplied by $\|f^\star\|_{\mathbf{k}}^2$ to recover a result for the original problem. For simplicity, denote

$$\widetilde{\sigma} = \sigma/\|f^\star\|_{\mathbf{k}}.$$

For notational convenience, define an event

$$\mathcal{E} = \big\{\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \le c(\xi_n^2 + \lambda')\big\},$$

and our goal is to show that $\mathcal{E}$ occurs with high-probability in terms of $\mathbb{P}$, the probability regarding all the randomness. For that end, we introduce several events that are used throughout,

$$\mathcal{E}_{\text{conc}} \triangleq \left\{ \sup_{g\in\mathcal{H}} \big|\|g\|_n - \|g\|_2\big| \le \tfrac{\delta_n}{2} \right\} \quad \text{and} \quad \mathcal{E}_{\text{lower}} \triangleq \Big\{\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2 > \delta_n\Big\}, \tag{69}$$

where δ_n is defined in (60) and $\mathcal{H}$ is the RKHS generated by $\mathbf{k}$ hence star-shaped. Further, we introduce two technical events $\mathcal{A}_{\mathrm{KT}}(u), \mathcal{B}_{\mathrm{KT}}$ defined in (82) and (95) respectively, which are proven to occur with small probability, and define a shorthand

$$\mathcal{E}_{\text{good}} \triangleq \mathcal{A}_{\mathrm{KT}}^c(\xi_n) \cap \mathcal{B}_{\mathrm{KT}}^c \cap \mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\mathrm{KT},\delta}.$$

Equipped with these shorthands, observe the following inequality,

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E} \cap \mathcal{E}_{\text{lower}}) + \mathbb{P}(\mathcal{E} \cap \mathcal{E}_{\text{lower}}^c) \\ &\ge \mathbb{P}(\mathcal{E} \cap \mathcal{E}_{\text{lower}}) + \mathbb{P}(\mathcal{E}_{\text{lower}}^c)\end{aligned} \tag{70}$$

where the second inequality is because $\mathcal{E}_{\text{lower}}^c \subseteq \mathcal{E}_{\text{lower}}^c \cap \mathcal{E}$ due to the assumption $\lambda' \ge 2\xi_n^2 \ge 2\delta_n^2$.

If we are able to show the set inclusion $\{\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}}\} \subseteq \{\mathcal{E} \cap \mathcal{E}_{\text{lower}}\}$ and that $\mathbb{P}(\mathcal{E}_{\text{good}}^c)$ is small, we are able refine (70) to the following

$$\mathbb{P}(\mathcal{E}) \ge \mathbb{P}(\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}}) + \mathbb{P}(\mathcal{E}_{\text{lower}}^c) \ge 1 - \mathbb{P}(\mathcal{E}_{\text{good}}^c) - \mathbb{P}(\mathcal{E}_{\text{lower}}^c) + \mathbb{P}(\mathcal{E}_{\text{lower}}) = 1 - \mathbb{P}(\mathcal{E}_{\text{good}}^c),$$

where the last quantity $1 - \mathbb{P}(\mathcal{E}_{\text{good}}^c)$ would be large.

To complete this proof strategy, we claim the set inclusion

$$\{\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}}\} \subseteq \{\mathcal{E} \cap \mathcal{E}_{\text{lower}}\} \tag{71}$$

to hold and prove it in App. D.1 and further claim

$$\mathbb{P}(\mathcal{E}_{\text{good}}^c) \le c''\big\{\delta + e^{-c' n\delta_n^2/(B_{\mathcal{H}}^2 \wedge \widetilde{\sigma}^2)}\big\} \tag{72}$$

which verify in App. D.2.

Putting the pieces together, claims (71) and (72) collectively implies

$$\mathbb{P}(\mathcal{E}) \ge 1 - c''\big\{\delta + e^{-c' n\delta_n^2/(B_{\mathcal{H}}^2 \wedge \widetilde{\sigma}^2)}\big\}$$

as desired.

D.1 Proof of claim (71)

There are several intermediary steps we take to show the set inclusion of interest (71). We introduce the shorthand

$$\widehat{\Delta}_{\mathrm{KT}} \triangleq \widehat{f}_{\mathrm{KT},\lambda'} - f^\star.$$

By invoking Propositions and basic inequalities to come, we successively show the following chain of set inclusions

$$\begin{aligned}\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}} &\subseteq \mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}} \cap \{\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\text{out}}}^2 \le c(\xi_n^2 + \lambda')\} && (73)\\ &\subseteq \mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}} \cap \{\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \le c(\xi_n^2 + \lambda')\} && (74)\\ &\subseteq \mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{lower}} \cap \mathcal{E} && (75)\\ &\subseteq \mathcal{E}_{\text{lower}} \cap \mathcal{E}. && (76)\end{aligned}$$

Note that step (76) is achieved trivially by dropping $\mathcal{E}_{\text{good}}$. Further note that (74) is the crucial intermediary step after which we may apply uniform concentration across n independent samples. Proof of (74) leverages on the Proposition to come (Prop. 1) that allows $\|\cdot\|_n$ and $\|\cdot\|_{n_{\text{out}}}$ to be exchangeable for finite rank kernels.

Recovering step (73) Since $\widehat{f}_{\mathrm{KT},\lambda'}$ and $f^\star$ are optimal and feasible, respectively for the central optimization problem of interest

$$\min_{f\in\mathcal{H}(\mathbf{k})} \tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}(y_i'-f(x_i'))^2+\lambda'\|f\|_{\mathbf{k}}^2,$$

we have the basic inequality

$$\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}\Big(y_i'-\widehat{f}_{\mathrm{KT},\lambda'}(x_i')\Big)^2+\lambda'\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}^2\le \tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}(y_i'-f^\star(x_i'))^2+\lambda'\|f^\star\|_{\mathbf{k}}^2, \quad (77)$$

With some algebra , may refine (77) to

$$\tfrac{1}{2}\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2\le\Big|\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}v_i'\widehat{\Delta}_{\mathrm{KT}}(x_i')\Big|+\lambda\Big\{\|f^\star\|_{\mathbf{k}}^2-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}^2\Big\}. \quad (78)$$

where $\widehat{\Delta}_{\mathrm{KT}}=\widehat{f}_{\mathrm{KT},\lambda'}-f^\star$.Suppose that $\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}<\xi_n$, then we trivially recover (73) by adding $\lambda'>0$. Thus, we assume that $\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}\ge\xi_n$.

Under the assumption $\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}\ge\xi_n$, which is without loss of generality, we utilize the basic inequality (78) and control its stochastic component

$$\Big|\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}v_i'\widehat{\Delta}_{\mathrm{KT}}(x_i')\Big|,$$

with a careful case work to follow, which is technical by nature.

Case where $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}\le 2$: Under such case, we introduce a technical event

$$\mathcal{A}_{\mathrm{KT}}(u)\triangleq\Big\{\exists g\in\mathcal{F}\setminus\mathcal{B}_2(\delta_n)\cap\{\|g\|_{n_{\mathrm{out}}}\ge u\}\text{ such that }\Big|\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}v_i'g(x_i')\Big|\ge 3\|g\|_{n_{\mathrm{out}}}u\Big\},$$

for any star-shaped function class $\mathcal{F}\subset\mathcal{H}$. Since $\|f^\star\|_{\mathbf{k}}=1$, triangle inequality implies $\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}\le\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}+\|f^\star\|_{\mathbf{k}}\le 3$. Moreover, on the event $\mathcal{E}_{\text{lower}}$ ($\subseteq\mathcal{E}_{\text{good}}$), we have $\|\widehat{\Delta}_{\mathrm{KT}}\|_2>\delta_n$. Thus, we may apply $\widehat{\Delta}_{\mathrm{KT}}$ to the event $\mathcal{A}_{\mathrm{KT}}^c(\xi_n)$ with $\mathcal{F}=\mathcal{B}_{\mathcal{H}}(3)$ (i.e., the $\mathcal{H}$-ball of radius 3) to attain

$$\Big|\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}v_i'\widehat{\Delta}_{\mathrm{KT}}(x_i')\Big|\le c_0\xi_n\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}\text{ on the event }\mathcal{A}_{\mathrm{KT}}^c(\xi_n)\cap\mathcal{E}_{\text{lower}}. \quad (79)$$

Upper bounding the stochastic component of the basic inequality (78) by (79) and dropping the $-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}^2$ term in (78), we have

$$\tfrac{1}{2}\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2\le c_0\xi_n\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}+\lambda'.$$

As a last step under the case $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}\le 2$, apply the quadratic formula (specifically, if $a,b\ge 0$ and $x^2-ax-b\le 0$, then $x\le a^2+b$) to obtain

$$\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2\le 4c_0^2\xi_n^2+2\lambda'.$$

Case where $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}>2$: Under such case, by assumption we have $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}>2>1\ge\|f^\star\|_{\mathbf{k}}$. Thus, we may derive the following

$$\|f^\star\|_{\mathbf{k}}-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}<0\quad\text{and}\quad\|f^\star\|_{\mathbf{k}}+\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}>1,$$

which further implies the following inequality

$$\|f^\star\|_{\mathbf{k}}^2-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}^2=\{\|f^\star\|_{\mathbf{k}}-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}\}\{\|f^\star\|_{\mathbf{k}}+\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}\}\le\|f^\star\|_{\mathbf{k}}-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}. \quad (80)$$

Further writing $\widehat{f}_{\mathrm{KT},\lambda'}=f^\star+\widehat{\Delta}_{\mathrm{KT}}$ and noting that $\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}-\|f^\star\|_{\mathbf{k}}\le\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}$ holds through triangle inequality, we may further refine (80) as

$$\|f^\star\|_{\mathbf{k}}-\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}}\le 2\|f^\star\|_{\mathbf{k}}-\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}\le 2-\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}},$$

so that the basic inequality in (78) reduces to

$$\tfrac{1}{2}\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2\le\Big|\tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}}v_i'\widehat{\Delta}_{\mathrm{KT}}(x_i')\Big|+\lambda'\{2-\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}\}. \quad (81)$$

We again introduce a technical event that controls the stochastic component of (81), which is

$$\mathcal{B}_{\mathrm{KT}} \triangleq \Big\{ \exists g \in \mathcal{F} \setminus \mathcal{B}_2(\delta_n) \cap \{\|g\|_{\mathbf{k}} \geq 1\} : \\ \Big| \tfrac{1}{n_{\mathrm{out}}} \textstyle\sum_{i=1}^{n_{\mathrm{out}}} v_i' g(x_i') \Big| > 4\xi_n \|g\|_{n_{\mathrm{out}}} + 2\xi_n^2 \|g\|_{\mathbf{k}} + \tfrac{1}{4}\|g\|_{n_{\mathrm{out}}}^2 \Big\}, \tag{82}$$

for a star-shaped function class $\mathcal{F} \subset \mathcal{H}$.

By triangle inequality, we have $\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} \geq \|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}} - \|f^\star\|_{\mathbf{k}} > 1$, and on event $\mathcal{E}_{\text{lower}}$ ($\subset \mathcal{E}_{\text{good}}$), we have $\|\widehat{\Delta}_{\mathrm{KT}}\|_2 > \delta_n$. Thus, we may apply $g = \widehat{\Delta}_{\mathrm{KT}}$ to the event $\mathcal{B}_{\mathrm{KT}}^c$, and the resulting refined basic inequality is

$$\begin{aligned} \tfrac{1}{2}\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2 &\leq 4\xi_n \|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}} + (2\xi_n^2 - \lambda')\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} + 2\lambda' \\ &\leq 4\xi_n \|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}} + 2\lambda' \quad \text{on the event } \mathcal{B}_{\mathrm{KT}}^c \cap \mathcal{E}_{\text{lower}} \end{aligned}$$

where the second inequality is due to the assumption that $\lambda' \geq 2\xi_n^2$. We apply the quadratic formula (specifically, if $a, b \geq 0$ and $x^2 - ax - b \leq 0$, then $x \leq a^2 + b$) to obtain

$$\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2 \leq 4c_0^2\xi_n^2 + 2\lambda'.$$

Putting the pieces together, we have shown

$$\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2 \leq c(\xi_n^2 + \lambda') \quad \text{on the event } \mathcal{A}_{\mathrm{KT}}^c(\xi_n) \cap \mathcal{B}_{\mathrm{KT}}^c \cap \mathcal{E}_{\text{lower}},$$

which is sufficient to recover (73).

Recovering step (74) We now upgrade events

$$\{\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2 \leq c(\xi_n^2 + \lambda')\} \quad \Longrightarrow \quad \{\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \leq c'(\xi_n^2 + \lambda')\}$$

by exploiting the events $\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\mathrm{KT},\delta}$ (subset of $\mathcal{E}_{\text{good}}$) that were otherwise not used when recovering (73). For this end, the following result is a crucial ingredient, which shows that $\|\cdot\|_{n_{\mathrm{out}}}$ and $\|\cdot\|_n$ are essentially exchangeable with high-probability,

Proposition 1 (Multiplicative guarantee for KT-COMPRESS++ with $\mathbf{k}_{\mathrm{RR}}$)**.** *Let $C_m \triangleq 1/\mu_m$ and suppose δ_n satisfies (63). Then on event $\mathcal{E}_{KT,\delta} \cap \mathcal{E}_{conc}$, where $\mathcal{E}_{KT,\delta}$ and $\mathcal{E}_{conc}$ are defined in (23) and (69) respectively, we have*

$$(1 - 4C_m \cdot \eta_{n,\mathbf{k}})\|g\|_n \leq \|g\|_{n_{\mathrm{out}}} \leq (1 + 4C_m \cdot \eta_{n,\mathbf{k}})\|g\|_n \tag{83}$$

uniformly over all $g \in \mathcal{H}$ such that $\|g\|_2 > \delta_n$.

See App. D.3 for the proof.

An immediate consequence of Prop. 1 is that

$$\{\|\widehat{\Delta}_{\mathrm{KT}}\|_{n_{\mathrm{out}}}^2 \leq c(\xi_n^2 + \lambda')\} \quad \Longrightarrow \quad \{\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \leq c'(\xi_n^2 + \lambda')\}$$

on the event $\mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{lower}}$,, which is sufficient to recover (74).

Recovering step (75) Our last step is to show

$$\{\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \leq c'(\xi_n^2 + \lambda')\} \quad \Longrightarrow \quad \{\|\widehat{\Delta}_{\mathrm{KT}}\|_2^2 \leq c''(\xi_n^2 + \lambda')\}.$$

Such result can be immediately shown on the event $\mathcal{E}_{\text{conc}}$ by observing that $\widehat{\Delta}_{\mathrm{KT}} \in \mathcal{H}$, by our assumption that $f^\star \in \mathcal{H}$ and by the definition

$$\widehat{f}_{\mathrm{KT},\lambda'} \in \operatorname{argmin}_{f \in \mathcal{H}(\mathbf{k})} L_{n_{\mathrm{out}}}(f) + \lambda' \|f\|_{\mathbf{k}}^2.$$

D.2 Proof of claim (72)

It suffices to show the appropriate bounds for the following four probability terms

$$\mathbb{P}(\mathcal{A}_{\mathrm{KT}}(\xi_n)), \quad \mathbb{P}(\mathcal{B}_{\mathrm{KT}}), \quad \mathbb{P}(\mathcal{E}^c_{\mathrm{conc}}), \quad \mathbb{P}(\mathcal{E}^c_{\mathrm{KT},\delta}).$$

Fix the shorthand

$$B_{\mathcal{H}} \triangleq \|\mathbf{k}\|^2_\infty R^2 < \infty.$$

We know from [10] that $\mathbb{P}(\mathcal{E}^c_{\mathrm{KT},\delta}|\mathcal{S}_{\mathrm{in}}) \le \delta$ and then we may apply [28, Thm. 14.1] to obtain a high probability statement,

$$\mathbb{P}(\mathcal{E}^c_{\mathrm{conc}}) \le e^{-c' n\delta_n^2/B^2_{\mathcal{H}}}. \tag{84}$$

Now we present two Lemmas that bound the $\mathbb{P}(\cdot \mid \mathcal{S}_{\mathrm{in}})$ probability of events $\mathcal{A}_{\mathrm{KT}}(\xi_n)$ and $\mathcal{B}_{\mathrm{KT}}$,

Lemma 5 (Controlling bad event when $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}} \le 2$)**.** *Suppose $u \ge \xi_n$. Then for some constant $c > 0$,*

$$\mathbb{P}(\mathcal{A}_{\mathrm{KT}}(u) \mid \mathcal{S}_{\mathrm{in}}) \le \delta + e^{-cn\delta_n^2/B^2_{\mathcal{H}}} + e^{-cnu^2/\widetilde{\sigma}^2} \tag{85}$$

where $\widetilde{\sigma} = \sigma/\|f^\star\|_{\mathbf{k}}$.

See App. D.4 for the proof. Note that by plugging in ξ_n into (85) results in a probability that depends on $\mathcal{S}_{\mathrm{in}}$ (as ξ_n depends on $\mathcal{S}_{\mathrm{in}}$). By invoking the definition of ξ_n, we may further refine the probability bound of $\mathcal{A}_{\mathrm{KT}}(\xi_n)$ by

$$\mathbb{P}(\mathcal{A}_{\mathrm{KT}}(\xi_n) \mid \mathcal{S}_{\mathrm{in}}) \le \delta + e^{-cn\delta_n^2/B^2_{\mathcal{H}}} + e^{-cn\delta_n^2/\widetilde{\sigma}^2} \tag{86}$$

Lemma 6 (Controlling bad event when $\|\widehat{f}_{\mathrm{KT},\lambda'}\|_{\mathbf{k}} > 2$)**.** *For some constants $c, c' > 0$,*

$$\mathbb{P}(\mathcal{B}_{\mathrm{KT}} \mid \mathcal{S}_{\mathrm{in}}) \le \delta + e^{-cn\delta_n^2/B^2_{\mathcal{H}}} + ce^{-n\xi_n^2/(c'\widetilde{\sigma}^2)} \tag{87}$$

where $\widetilde{\sigma} = \sigma/\|f^\star\|_{\mathbf{k}}$.

See App. D.5 for the proof. It is notable that ξ_n in the probability bound of (87) contains a term ε_n defined in (59) that is a function of $\mathcal{S}_{\mathrm{in}}$. Invoking the definition of ξ_n, we observe the probability upper bound (87) can be refined to

$$\mathbb{P}(\mathcal{B}_{\mathrm{KT}} \mid \mathcal{S}_{\mathrm{in}}) \le \delta + e^{-cn\delta_n^2/B^2_{\mathcal{H}}} + ce^{-n\delta_n^2/(c'\widetilde{\sigma}^2)}, \tag{88}$$

which does not depend on $\mathcal{S}_{\mathrm{in}}$.

Putting the pieces together, we have the following probability bound for some constants $c, c' > 0$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c_{\mathrm{good}} \mid \mathcal{S}_{\mathrm{in}}) &\le \mathbb{P}(\mathcal{A}_{\mathrm{KT}}(\xi_n) \mid \mathcal{S}_{\mathrm{in}}) + \mathbb{P}(\mathcal{B}_{\mathrm{KT}} \mid \mathcal{S}_{\mathrm{in}}) + \mathbb{P}(\mathcal{E}^c_{\mathrm{conc}} \mid \mathcal{S}_{\mathrm{in}}) + \mathbb{P}(\mathcal{E}^c_{\mathrm{KT},\delta} \mid \mathcal{S}_{\mathrm{in}}) \\ &\overset{(84)(86)(88)}{\le} c\big\{\delta + e^{-c' n\delta_n^2/(B^2_{\mathcal{H}} \wedge \widetilde{\sigma}^2)}\big\}\end{aligned}$$

thereby implying $\mathbb{P}(\mathcal{E}^c_{\mathrm{good}}) \le c''\big\{\delta + e^{-c' n\delta_n^2/(B^2_{\mathcal{H}} \wedge \widetilde{\sigma}^2)}\big\}$ for some constant c''.

D.3 Proof of Prop. 1: Multiplicative guarantee for KT-COMPRESS++ with $\mathbf{k}_{\mathrm{RR}}$

Fix $g \in \mathcal{H}$. Denote $\langle g, h\rangle = \int g(x)h(x)dx$ as the inner product in the L^2 sense. By Mercer's theorem [28, Cor. 12.26], the $\mathbf{k}$-norm of g has a basis expansion $\|g\|^2_{\mathbf{k}} = \sum_{i=1}^m \langle g, \phi_i\rangle^2/\lambda_i$ so that

$$\|g\|^2_{\mathbf{k}} \le \textstyle\sum_{i=1}^m \langle g, \phi_i\rangle^2/\lambda_m = C_m\|g\|_2^2 \quad \text{since} \quad C_m = 1/\lambda_m. \tag{89}$$

The assumption $\|g\|_2 \ge \delta_n$ implies that on the event $\mathcal{E}_{\mathrm{conc}}$ (69), we have

$$\tfrac{1}{2}\delta_n \le \|g\|_2 - \tfrac{1}{2}\delta_n \le \|g\|_n \tag{90}$$

Moreover, g must be a non-zero function. Note that $g^2 \in \mathcal{H}(\mathbf{k}_{\mathrm{RR}})$ (see App. F.3). Thus, we may apply Lem. 12 to $f_1 = f_2 = g$ and $a = 1, b = 0$ to obtain

$$\big|\|g\|_n^2 - \|g\|^2_{n_{\mathrm{out}}}\big| = \Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n g^2(x_i) - \frac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}} g^2(x'_i)\Big| \le \|g\|^2_{\mathbf{k}} \cdot \eta_{n,\mathbf{k}}. \tag{91}$$

The LHS can be expanded as

$$\big|\|g\|_n^2 - \|g\|_{n_{\text{out}}}^2\big| = \big|\|g\|_n - \|g\|_{n_{\text{out}}}\big| \cdot \underbrace{\big|\|g\|_n + \|g\|_{n_{\text{out}}}\big|}_{>0 \text{ by } (90)}.$$

Thus, we may rearrange (91) and combine with (89) to obtain

$$\big|\|g\|_n - \|g\|_{n_{\text{out}}}\big| \le \tfrac{C_m\|g\|_2^2}{\|g\|_n + \|g\|_{n_{\text{out}}}} \cdot \eta_{n,\mathbf{k}}. \tag{92}$$

On event $\mathcal{E}_{\text{conc}}$, we have

$$\|g\|_2^2 \overset{(69)}{\le} (\tfrac{1}{2}\delta_n + \|g\|_n)^2 \overset{(i)}{\le} \tfrac{\delta_n^2}{4} + \delta_n\|g\|_n + \|g\|_n^2. \tag{93}$$

Thus, we have

$$\begin{aligned}
\tfrac{\|g\|_2^2}{\|g\|_n + \|g\|_{n_{\text{out}}}} &\overset{(93)}{\le} \tfrac{\delta_n^2}{4\big|\|g\|_n + \|g\|_{n_{\text{out}}}\big|} + \tfrac{\delta_n\|g\|_n}{\|g\|_n + \|g\|_{n_{\text{out}}}} + \tfrac{\|g\|_n^2}{\|g\|_n + \|g\|_{n_{\text{out}}}} \\
&\overset{(90)}{\le} \tfrac{\delta_n^2}{2\delta_n} + \delta_n + \|g\|_n \cdot \tfrac{\|g\|_n}{\|g\|_n + \|g\|_{n_{\text{out}}}} \\
&\le \tfrac{3}{2}\delta_n + \|g\|_n \\
&\overset{(90)}{\le} 3\|g\|_n + \|g\|_n = 4\|g\|_n.
\end{aligned} \tag{94}$$

Using (94) to refine (92), we have on event $\mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$:

$$\big|\|g\|_n - \|g\|_{n_{\text{out}}}\big| \le 4C_m\|g\|_n \cdot \eta_{n,\mathbf{k}}.$$

With some algebra, this implies with probability at least $1 - \delta - \exp(-c'n\delta_n^2/B_{\mathcal{F}}^2)$:

$$(1 - 4C_m \cdot \eta_{n,\mathbf{k}})\|g\|_n \le \|g\|_{n_{\text{out}}} \le (1 + 4C_m \cdot \eta_{n,\mathbf{k}})\|g\|_n$$

uniformly over all non-zero $g \in \mathcal{H}$ such that $\|g\|_2 > \delta_n$.

D.4 Proof of Lem. 5: Controlling bad event when $\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}} \le 2$

Recall $\mathcal{E}_{\text{KT},\delta}$ and $\mathcal{E}_{\text{conc}}$ defined by (23) and (69). Also recall that $\mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$ combined with the assumption $\|g\|_2 \ge \delta_n$ invokes the event (83). Our aim is to show

$$\mathcal{A}_{\text{KT}}(u) \cap \mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}} \subseteq \{Z_n(2u) \ge 2u^2\}, \quad \text{where} \quad Z_n(t) \triangleq \sup_{\substack{g \in \mathcal{F}: \\ \|g\|_n \le t}} \big|\tfrac{\widetilde{\sigma}}{n}\textstyle\sum_{i=1}^n w_i g(x_i)\big| \tag{95}$$

so that we have a probability bound

$$\mathbb{P}(\mathcal{A}_{\text{KT}}(u)) \le \mathbb{P}(\mathcal{E}_{\text{KT},\delta}^c) + \mathbb{P}(\mathcal{E}_{\text{conc}}^c) + \mathbb{P}(Z_n(2u) \ge 2u^2).$$

The first RHS term can be bounded by δ (see (23)). The second RHS term can bounded by (84). The third term can be bounded by

$$\mathbb{P}(Z_n(2u) \ge 2u^2) = \mathbb{P}\big(Z_n(u) \ge u^2/2 + u^2/2\big) \overset{(i)}{\le} \mathbb{P}\big(Z_n(u) \ge u\varepsilon_n/2 + u^2/2\big) \overset{(ii)}{\le} e^{-\frac{nu^2}{8\widetilde{\sigma}^2}},$$

where (i) follows from our assumption that $u \ge \varepsilon_n$ and (ii) follows from applying generic concentration bounds on $Z_n(u)$ (see [28, Thm. 2.26, Eq. 13.66]). Putting together the pieces yields our desired probability bound (85).

Proof of claim (95). Consider the event $\mathcal{A}_{\text{KT}}(u) \cap \mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$. The norm equivalence established on the event $\mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$ in Prop. 1 is an important ingredient throughout.

Let $g \in \mathcal{H}$ be the function that satisfies three conditions: $\|g\|_2 \ge \delta_n$, $\|g\|_{n_{\text{out}}} \ge u$, and

$$\Big|\tfrac{1}{n_{\text{out}}}\textstyle\sum_{i=1}^{n_{\text{out}}} v_i' g(x_i')\Big| \ge 3\|g\|_{n_{\text{out}}} u.$$

Define the normalized function

$$\widetilde{g} = u \cdot g / \|g\|_{n_{\text{out}}}$$

so that it satisfies $\|\widetilde{g}\|_{n_{\text{out}}} = u$ and also

$$\left| \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} v_i' \widetilde{g}(x_i') \right| \geq 3u^2. \tag{96}$$

By triangle inequality, the LHS of (96) can be further upper bounded by

$$\left| \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} v_i' \widetilde{g}(x_i') \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} v_i \widetilde{g}(x_i) \right| + \frac{u}{\|g\|_{n_{\text{out}}}} \left| \frac{1}{n} \sum_{i=1}^{n} v_i g(x_i) - \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} v_i' g(x_i') \right|. \tag{97}$$

Recall the chosen g satisies $\|g\|_{n_{\text{out}}} \geq u$. Observe that

$$v_i g(x_i) \overset{(1)}{=} (y_i - f^\star(x_i)) g(x_i) = -f^\star(x_i) g(x_i) + y_i g(x_i),$$

so we may apply Lem. 12 with $f_1 = f^\star, f_2 = g$ and $a = -1, b = 1$. Thus, on the event $\mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} v_i g(x_i) - \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} v_i' g(x_i') \right| \leq \|g\|_{\mathbf{k}} (\|f^\star\|_{\mathbf{k}} + 1) \cdot \eta_{n,\mathbf{k}}. \tag{98}$$

Thus, we may rearrange (97) and combine with (96) and (98) to obtain

$$\left| \frac{1}{n} \sum_{i=1}^{n} v_i \widetilde{g}(x_i) \right| \geq 3u^2 - \frac{u}{\|g\|_{n_{\text{out}}}} \|g\|_{\mathbf{k}} (\|f^\star\|_{\mathbf{k}} + 1) \cdot \eta_{n,\mathbf{k}}$$

Note that

$$\frac{\|g\|_{\mathbf{k}}}{\|g\|_{n_{\text{out}}}} = \frac{\|g\|_{\mathbf{k}}}{\|g\|_2} \cdot \frac{\|g\|_2}{\|g\|_n} \cdot \frac{\|g\|_n}{\|g\|_{n_{\text{out}}}}.$$

We tackle each term in turn. First, $\frac{\|g\|_{\mathbf{k}}}{\|g\|_2} \overset{(89)}{\leq} \sqrt{C}_m$. Since we assume $\|g\|_2 \geq \delta_n$, we have $\frac{\|g\|_2}{\|g\|_n} \leq \frac{\delta_n/2 + \|g\|_n}{\|g\|_n} \overset{(90)}{\leq} 2$ on event $\mathcal{E}_{\text{conc}}$; and $\frac{\|g\|_n}{\|g\|_{n_{\text{out}}}} \overset{(103)}{\leq} 2$ on event $\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{KT},\delta}$. Taken together,

$$\frac{\|g\|_{\mathbf{k}}}{\|g\|_{n_{\text{out}}}} \leq 4\sqrt{C}_m \quad \text{on event} \quad \mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{KT},\delta}. \tag{99}$$

As $u \geq \xi_n \geq 4\sqrt{C}_m (\|f^\star\|_{\mathbf{k}} + 1) \eta_{n,\mathbf{k}}$ by assumption, we have therefore found $\widetilde{g}$ with norm $\|\widetilde{g}\|_{n_{\text{out}}} = u$ satisfying

$$\left| \frac{1}{n} \sum_{i=1}^{n} v_i \widetilde{g}(x_i) \right| \geq 3u^2 - u^2 = 2u^2.$$

We may further show that

$$\|\widetilde{g}\|_n = \frac{u}{\|g\|_{n_{\text{out}}}} \|g\|_n \leq u \frac{\|g\|_n}{\|g\|_{n_{\text{out}}}} \leq u \cdot 2 \quad \text{on event} \quad \mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{KT},\delta},$$

where the last inequality follows from the fact that $\|g\|_2 \geq \delta_n$ and by applying (103). So we observe

$$2u^2 \leq \left| \frac{1}{n} \sum_{i=1}^{n} v_i \widetilde{g}(x_i) \right| \leq \sup_{\|\widetilde{g}\|_n \leq 2u} \left| \frac{1}{n} \sum_{i=1}^{n} v_i \widetilde{g}(x_i) \right| = Z_n(2u)$$

D.5 Proof of Lem. 6: Controlling bad event when $\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}} > 2$

Our aim is to show for any $g \in \partial\mathcal{H}$ with $\|g\|_{\mathbf{k}} \geq 1$,

$$\left| \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} v_i' g(x_i') \right| \leq 2\xi_n \|g\|_{n_{\text{out}}} + 2\xi_n^2 \|g\|_{\mathbf{k}} + \frac{1}{16} \|g\|_{n_{\text{out}}}^2 \quad \text{with high probability.}$$

Note that it is sufficient to prove our aim for $g \in \partial\mathcal{H}$ with $\|g\|_{\mathbf{k}} = 1$—by proving only for g with $\|g\|_{\mathbf{k}} = 1$, then for any $h \in \partial\mathcal{H}$ with $\|h\|_{\mathbf{k}} \geq 1$, we may plug $g = h/\|h\|_{\mathbf{k}}$ into

$$\left|\frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} v_i' g(x_i')\right| \leq 2\xi_n \|g\|_{n_{\text{out}}} + 2\xi_n^2 + \frac{1}{16}\|g\|_{n_{\text{out}}}^2 \tag{100}$$

to recover the aim of interest. So without loss of generality, we show (100) for all g such that $g \in \partial\mathcal{H}$ and $\|g\|_{\mathbf{k}} = 1$.

Let $\mathcal{B}_{\text{KT}}$ denote the event where (100) is violated, i.e. there exists $g \in \partial\mathcal{H}$ with $\|g\|_{\mathbf{k}} = 1$ so that

$$\left|\frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} v_i' g(x_i')\right| > 3\xi_n \|g\|_{n_{\text{out}}} + 2\xi_n^2 + \frac{1}{4}\|g\|_{n_{\text{out}}}^2. \tag{101}$$

We prove the following set inclusion,

$$\begin{aligned}
&\mathcal{B}_{\text{KT}} \cap \mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}} \\
\subseteq &\left\{\exists\, g \in \partial\mathcal{H} \text{ s.t. } \|g\|_{\mathbf{k}} = 1 \text{ and } \left|\frac{1}{n}\sum_{i=1}^{n} v_i g(x_i)\right| > 2\varepsilon_n \|g\|_n + 2\varepsilon_n^2 + \frac{1}{16}\|g\|_n^2\right\},
\end{aligned} \tag{102}$$

where we know the RHS event of (102) has probability bounded by $ce^{-n\xi_n^2/(c'\tilde{\sigma}^2)}$ which is proven in [28, Lem. 13.23]. So the set inclusion (102) implies a bound over the event $\mathcal{B}_{\text{KT}}$,

$$\mathbb{P}(\mathcal{B}_{\text{KT}}) \leq \mathbb{P}(\mathcal{E}_{\text{KT},\delta}^c) + \mathbb{P}(\mathcal{E}_{\text{conc}}^c) + ce^{-n\xi_n^2/(c'\tilde{\sigma}^2)},$$

where $\mathbb{P}(\mathcal{E}_{\text{KT},\delta}^c) \leq \delta$ by (23) and $\mathbb{P}(\mathcal{E}_{\text{conc}}^c)$ by (84).

Choose g so that $\|g\|_{\mathbf{k}} = 1$ and (101) holds. Condition $\|g\|_{\mathbf{k}} = 1$ as well as the condition (89) resulting from a finite rank kernel $\mathbf{k}$ implies $\delta_n \leq 1 \leq \|g\|_{\mathbf{k}} \leq \sqrt{C_m}\|g\|_2$. Invoke Prop. 1 for the choice of g that satisfies $\|g\|_2 \geq \delta_n/\sqrt{C_m} \geq \delta_n$, so that on the event $\mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\text{conc}}$, we have the following norm equivalence,

$$\tfrac{1}{2}\|g\|_n \leq \|g\|_{n_{\text{out}}} \leq \tfrac{3}{2}\|g\|_n \quad \text{for any } n \text{ such that } C_m \cdot \eta_{n,\mathbf{k}} \leq 1/18. \tag{103}$$

Then we have the following chain of inequalities, which holds on event $\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{KT},\delta}$

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} v_i g(x_i)\right| &\overset{(i)}{\geq} \left|\frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} v_i' g(x_i')\right| - \left|\frac{1}{n}\sum_{i=1}^{n} v_i g(x_i) - \frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} v_i' g(x_i')\right| \\
&\overset{(ii)}{\geq} 3\xi_n\|g\|_{n_{\text{out}}} + 2\xi_n^2 + \frac{1}{4}\|g\|_{n_{\text{out}}}^2 - \|g\|_{\mathbf{k}}(\|f^\star\|_{\mathbf{k}} + 1)\cdot\eta_{n,\mathbf{k}} \\
&\overset{(99)}{\geq} 3\xi_n\|g\|_{n_{\text{out}}} + 2\xi_n^2 + \frac{1}{4}\|g\|_{n_{\text{out}}}^2 - 4\sqrt{C_m}\|g\|_{n_{\text{out}}}(\|f^\star\|_{\mathbf{k}} + 1)\cdot\eta_{n,\mathbf{k}} \\
&\overset{(103)}{\geq} (\tfrac{3}{2}\xi_n - 2\sqrt{C_m}(\|f^\star\|_{\mathbf{k}} + 1)\cdot\eta_{n,\mathbf{k}})\|g\|_n + 2\xi_n^2 + \frac{1}{16}\|g\|_n^2,
\end{aligned} \tag{104}$$

where step (i) follows from triangle inequality and step (ii) follows from our assumption (101) to bound the first term and our approximation guarantee (98) to bound the second term. By definition of ξ_n, we have

$$\tfrac{3}{2}\xi_n - 2\sqrt{C_m}(\|f^\star\|_{\mathbf{k}} + 1)\cdot\eta_{n,\mathbf{k}} \geq 2\xi_n.$$

Using this to refine (104), we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} v_i g(x_i)\right| \geq 2\xi_n\|g\|_n + 2\xi_n^2 + \frac{1}{16}\|g\|_n^2,$$

which directly implies the inclusion (102) as desired.

E Proof of Thm. 3: KT-KRR guarantee for infinite-dimensional RKHS

We state a more detailed version of the theorem:

Theorem 5 (KT-KRR guarantee for infinite-dimensional RKHS, detailed)**.** *Assume $f^\star \in \mathcal{H}(\mathbf{k})$ and Assum. 1 is satisfied. If $\mathbf{k}$ is* LOGGROWTH(α, β)*, then for some constant c (depending on d, α, β), $\widehat{f}_{\mathrm{KT},\lambda'}$ with $\lambda' = \mathcal{O}(1/n_{\mathrm{out}})$ satisfies*

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \leq c\Big(\frac{\log^\alpha n}{n} + \frac{\sqrt{\log^\alpha n_{\mathrm{out}}}}{n_{\mathrm{out}}}\Big) \cdot \big[\|f^\star\|_{\mathbf{k}} + 1\big]^2 \tag{105}$$

with probability at least $1 - 2\delta - 2e^{-\frac{n\delta_n^2}{c_1(\|f^\star\|_{\mathbf{k}}^2 + \sigma^2)}}$.

If $\mathbf{k}$ is POLYGROWTH(α, β) *with $\alpha \in (0, 2)$, then for some constant c (depending on d, α, β), $\widehat{f}_{\mathrm{KT},\lambda'}$ with $\lambda = \mathcal{O}(n_{\mathrm{out}}^{-\frac{2-\alpha}{2}})$ satisfies*

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \leq c\|f^\star\|_{\mathbf{k}}^{\frac{2}{2+\alpha}} n^{-\frac{2}{2+\alpha}} + \big[\|f^\star\|_{\mathbf{k}} + 1\big]^2 n_{\mathrm{out}}^{-\frac{2-\alpha}{2}} \log n_{\mathrm{out}} + c' b^{\frac{4}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \tag{106}$$

with probability at least $1 - 2\delta - 2e^{-\frac{n\delta_n^2}{c_1(\|f^\star\|_{\mathbf{k}}^2 + \sigma^2)}}$.

E.1 Generic KT-KRR guarantee

We state a generic result for infinite-dimensional RKHS that only depends on the Rademacher and Gaussian critical radii as well as the KT approximation term, all introduced in App. C.

Theorem 6 (KT-KRR)**.** *Let $f^\star \in \mathcal{H}(\mathbf{k})$ and Assum. 1 is satisfied. Let δ_n, ε_n denote the solutions to (59), (60), respectively. Denote $\widehat{f}_{\mathrm{KT},\lambda'}$ with regularization parameter $\lambda' \geq 2\eta_{n,\mathbf{k}}$, where $\eta_{n,\mathbf{k}}$ is defined by (61). Then with probability at least $1 - 2\delta - 2e^{-\frac{n\delta_n^2}{c(\|f^\star\|_{\mathbf{k}}^2 + \sigma^2)}} - c_1 e^{-c_2 \frac{n\|f^\star\|_{\mathbf{k}}^2 \varepsilon_n^2}{\sigma^2}}$, we have*

$$\begin{aligned}
\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 &\leq \mathbb{U}^{\mathrm{full}} + \mathbb{U}^{\mathrm{KT}}, \quad \textit{where} \\
\mathbb{U}^{\mathrm{full}} &\triangleq c\big(\varepsilon_n^2 + \lambda'\big)\big[\|f^\star\|_{\mathbf{k}} + 1\big]^2 + c\delta_n^2 \quad \textit{and} \\
\mathbb{U}^{\mathrm{KT}} &\triangleq c \cdot \eta_{n,\mathbf{k}} \big[\|f^\star\|_{\mathbf{k}} + 1\big]^2.
\end{aligned}$$

See App. F for the proof. The term $\mathbb{U}^{\mathrm{full}}$ follows from the excess risk bound of FULL-KRR $\widehat{f}_{\mathrm{full},\lambda}$. The term $\mathbb{U}^{\mathrm{KT}}$ follows from our KT approximation. Clearly, the best rates are achieved when we choose $\lambda = 2\eta_{n,\mathbf{k}}$.

E.2 Proof of explicit rates

The strategy for each setting is as follows:

1. Bound the Gaussian critical radius (60) using [28, Cor. 13.18], which reduces to finding $\varepsilon > 0$ satisfying the inequality
$$\sqrt{\tfrac{2}{n}}\sqrt{\textstyle\sum_{j=1}^n \min\{\varepsilon^2, \hat{\mu}_j\}} \leq \beta\varepsilon^2, \quad \text{where} \quad \beta \triangleq \tfrac{\|f^\star\|_{\mathbf{k}}}{4\sigma} \tag{107}$$
and $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \ldots \geq \hat{\mu}_n \geq 0$ are the eigenvalues of the normalized kernel matrix $\mathbf{K}/n$, where $\mathbf{K}$ is defined by (6).
2. Bound the Rademacher critical radius (59) using [28, Cor. 14.5], which reduces to solving the inequality
$$\sqrt{\tfrac{2}{n}}\sqrt{\textstyle\sum_{j=1}^\infty \min\{\delta^2, \mu_j\}} \leq \tfrac{\delta^2}{b}, \tag{108}$$
where $(\mu_j)_{j=1}^\infty$ are the eigenvalues of the $\mathbf{k}$ according to Mercer's theorem [28, Thm. 12.20] and b is the uniform bound on the function class.
3. Bound $\eta_{n,\mathbf{k}}$ (61) using the covering number bound $\mathcal{N}(\mathcal{B}_2^d(\mathfrak{R}_{\mathrm{in}}), \epsilon)$ from Assum. 3.

In the sequel, we make use of the following notation. Let

$$R_n \triangleq 1 + \sup_{x \in \mathcal{S}_{\mathrm{in}}} \|x_i\|_2 \overset{(62)}{=} 1 + \mathfrak{R}_{\mathrm{in}} \quad \text{and} \quad L_{\mathbf{k}}(r) \triangleq \tfrac{\mathfrak{C}_d}{\log 2} r^\beta$$

according to [15, Eq. 6], where $\mathfrak{C}_d$ is the constant that appears in Assum. 3.

E.2.1 Proof of (106)

We begin by solving (107).

Lemma 7 (Critical Gaussian radius for PolyGrowth kernels)**.** *Suppose Assum. 1 is satisfied and* $\mathbf{k}$ *is* PolyGrowth *with* $\alpha < 2$ *as defined by Assum. 3. Then the Gaussian critical radius satisfies*

$$\varepsilon_n^2 \simeq \left(\tfrac{2c}{\|f^\star\|_{\mathbf{k}}/4\sigma}\right)^{\frac{4}{2+\alpha}} \left(2^{-\alpha} L_{\mathbf{k}}(R_n)(1+\tfrac{32\alpha}{2-\alpha})\right)^{\frac{2}{2+\alpha}} \cdot n^{-\frac{2}{2+\alpha}}. \tag{109}$$

Proof. [15, Cor. B.1] implies that

$$\hat{\mu}_j \le 4\Big(\tfrac{L_{\mathbf{k}}(R_n)}{j-1}\Big)^{\frac{2}{\alpha}} \quad \text{for all} \quad j > L_{\mathbf{k}}(R_n)+1$$

Let k be the smallest integer such that

$$k > L_{\mathbf{k}}(R_n)+1 \quad \text{and} \quad 4\Big(\tfrac{L_{\mathbf{k}}(R_n)}{k-1}\Big)^{\frac{2}{\alpha}} \le \varepsilon^2.$$

By Assum. 1, R_n is a constant, so the first inequality is easily satisfied for large enough n

$$k \ge 2^{-\alpha} L_{\mathbf{k}}(R_n)\varepsilon^{-\alpha} + 1. \tag{110}$$

Then

$$\begin{aligned}
\tfrac{2}{\sqrt{n}}\sqrt{\textstyle\sum_{j=1}^n \min\{\varepsilon^2, \hat{\mu}_j\}} &\le \tfrac{2}{\sqrt{n}}\sqrt{k\varepsilon^2 + \textstyle\sum_{j=k+1}^n 4\Big(\tfrac{L_{\mathbf{k}}(R_n)}{j-1}\Big)^{\frac{2}{\alpha}}} \\
&\overset{(i)}{\le} \tfrac{2}{\sqrt{n}}\sqrt{k\varepsilon^2 + \tfrac{4L_{\mathbf{k}}(R_n)^{2/\alpha}}{2/\alpha-1}k^{1-2/\alpha}} \\
&\overset{(110)}{\le} \tfrac{2}{\sqrt{n}}\sqrt{2^{-\alpha}L_{\mathbf{k}}(R_n)\varepsilon^{2-\alpha} + \tfrac{4\cdot 2^{2-\alpha}L_{\mathbf{k}}(R_n)}{2/\alpha-1}\varepsilon^{2-\alpha}},
\end{aligned}$$

where step (i) follows from the approximation

$$\textstyle\sum_{j=k}^{n-1} 4\Big(\tfrac{L_{\mathbf{k}}(R_n)}{j}\Big)^{\frac{2}{\alpha}} \le 4L_{\mathbf{k}}(R_n)^{2/\alpha}\int_k^\infty t^{-2/\alpha}dt = 4L_{\mathbf{k}}(R_n)^{2/\alpha}\tfrac{1}{2/\alpha-1}k^{1-\frac{2}{\alpha}}.$$

To solve (107), it suffices to solve

$$\begin{aligned}
&\tfrac{2c}{\sqrt{n}}\sqrt{2^{-\alpha}L_{\mathbf{k}}(R_n)(1+\tfrac{16}{2/\alpha-1})\varepsilon^{2-\alpha}} \le \beta\varepsilon^2 \\
\implies \quad &\varepsilon \ge \left(\tfrac{2c}{\beta}\right)^{\frac{2}{2+\alpha}} \left(2^{-\alpha}L_{\mathbf{k}}(R_n)(1+\tfrac{32\alpha}{2-\alpha})\right)^{\frac{1}{2+\alpha}} \cdot n^{-\frac{1}{2+\alpha}}.
\end{aligned}$$

Since ε_n is the smallest such solution to (107) by definition, we have (109) as desired. □

We proceed to solve (108).

Lemma 8. *Suppose Assum. 1 is satisfied and* $\mathbf{k}$ *is* PolyGrowth *with* $\alpha < 2$ *as defined by Assum. 3. Then the Rademacher critical radius satisfies*

$$\varepsilon_n^2 \simeq b^{\frac{4}{2+\alpha}} \left(2^{-\alpha}L_{\mathbf{k}}(R_n)(1+\tfrac{32\alpha}{2-\alpha})\right)^{\frac{2}{2+\alpha}} n^{-\frac{2}{2+\alpha}}.$$

Proof. Thus, we can solve the following inequality

$$\sqrt{\tfrac{2}{n}}\sqrt{\textstyle\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \le \tfrac{1}{b}\delta^2,$$

Following the same logic as in the proof of Lem. 7 but with $\beta = 1/b$ yields the desired bound. □

Finally, it remains to bound (61). We have

$$\begin{aligned}
\eta_{n,\mathbf{k}} &= \tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2 + \mathfrak{W}_{\mathbf{k}}(n, n_{\text{out}}, \delta, \mathfrak{R}_{\text{in}}, \tfrac{\mathfrak{a}}{n_{\text{out}}})) \\
&\le \tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2 + \sqrt{\log\Big(\tfrac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\Big) \cdot \Big[\log\big(\tfrac{1}{\delta}\big) + \log \mathcal{N}_{\mathbf{k}}(\mathcal{B}_2^d(\mathfrak{R}_{\text{in}}), \tfrac{\mathfrak{a}}{n_{\text{out}}})\Big]})
\end{aligned}$$

$$\leq \frac{\mathfrak{a}^2}{n_{\text{out}}}\Big(2+\sqrt{\log\Big(\frac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\Big)\cdot\Big[\log\big(\frac{1}{\delta}\big)+\mathfrak{C}_d\big(\frac{n_{\text{out}}}{\mathfrak{a}}\big)^{\alpha}(\mathfrak{R}_{\text{in}}+1)^{\beta})\Big]}$$

$$\leq \frac{\mathfrak{a}^2}{n_{\text{out}}}\left(2+\sqrt{\log\Big(\frac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\Big)}\cdot\left[\sqrt{\log\big(\frac{1}{\delta}\big)}+\sqrt{\mathfrak{C}_d\frac{(\mathfrak{R}_{\text{in}}+1)^{\beta}}{\mathfrak{a}^{\alpha}}n_{\text{out}}^{\frac{\alpha}{2}}}\right]\right)$$

$$\leq n_{\text{out}}^{\frac{\alpha}{2}-1}\cdot\mathfrak{a}^2\left(2+\sqrt{\log\Big(\frac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\Big)}\cdot\sqrt{\mathfrak{C}_d\frac{(\mathfrak{R}_{\text{in}}+1)^{\beta}}{\mathfrak{a}^{\alpha}}}\right)$$

for some universal positive constant c.

In summary, there exists positive constants c_0, c_1, c_2 such that

$$\varepsilon_n^2 \leq c_0\Big(\frac{\sigma}{\|f^\star\|_{\mathbf{k}}}\Big)^{\frac{4}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \quad \delta_n^2 \leq c_1 b^{\frac{4}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \quad \eta_{n,\mathbf{k}} \leq c_2\mathfrak{a}^2 n_{\text{out}}^{-\frac{2-\alpha}{2}}\log n_{\text{out}}$$

Setting $\lambda' = c_2\mathfrak{a}^2 n_{\text{out}}^{-\frac{2-\alpha}{2}}\log n_{\text{out}}$, we have

$$\begin{aligned}\|\widehat{f}_{\text{KT},\lambda'} - f^\star\|_2^2 &\leq c\big(\varepsilon_n^2+\lambda'+\eta_{n,\mathbf{k}}\big)\cdot\big[\|f^\star\|_{\mathbf{k}}+1\big]^2 + c'\delta_n^2 \\ &\leq c\|f^\star\|_{\mathbf{k}}^{\frac{2}{2+\alpha}} n^{-\frac{2}{2+\alpha}} + \big[\|f^\star\|_{\mathbf{k}}+1\big]^2 n_{\text{out}}^{-\frac{2-\alpha}{2}}\log n_{\text{out}} + c' b^{\frac{4}{2+\alpha}} n^{-\frac{2}{2+\alpha}}.\end{aligned}$$

E.2.2 Proof of (105)

We begin by solving (107).

Lemma 9 (Critical Gaussian radius for LOGGROWTH kernels)**.** *Under Assum. 1 and* LOGGROWTH *version of Assum. 3, Gaussian critical radius satisfies*

$$\varepsilon_n^2 \simeq \frac{\sigma^2}{\|f^\star\|_{\mathbf{k}}^2}\frac{\log(2e\cdot\frac{\|f^\star\|_{\mathbf{k}}}{4\sigma}\sqrt{n})^{\alpha}}{n}\cdot L_{\mathbf{k}}(R_n)C''_{\alpha}$$

for some constant C''_α that only depends on α. where we ignore log-log factors.

Proof. [15, Cor. B.1] implies that

$$\hat{\mu}_j \leq 4\exp\left(2-2\Big(\frac{j-1}{L_{\mathbf{k}}(R_n)}\Big)^{\frac{1}{\alpha}}\right) \quad \text{for all} \quad j > L_{\mathbf{k}}(R_n)+1$$

Let k be the smallest integer such that

$$k > L_{\mathbf{k}}(R_n)+1 \quad \text{and} \quad 4\exp\left(2-2\Big(\frac{j-1}{L_{\mathbf{k}}(R_n)}\Big)^{\frac{1}{\alpha}}\right) \leq \varepsilon^2.$$

By Assum. 1, R_n is a constant, so the first inequality is easily satisfied for large enough n

$$k \geq L_{\mathbf{k}}(R_n)\log\big(\tfrac{2e}{\varepsilon}\big)^{\alpha}+1. \tag{111}$$

Thus, $k = \lceil L_{\mathbf{k}}(R_n)\log\big(\frac{2e}{\varepsilon}\big)^{\alpha}+1\rceil$. Then

$$\frac{2}{\sqrt{n}}\sqrt{\textstyle\sum_{j=1}^n \min\{\varepsilon^2,\hat{\mu}_j\}} \leq \frac{2}{\sqrt{n}}\sqrt{k\varepsilon^2+\sum_{j=k+1}^n 4\exp\left(2-2\Big(\frac{j-1}{L_{\mathbf{k}}(R_n)}\Big)^{\frac{1}{\alpha}}\right)} \tag{112}$$

Consider the following approximation:

$$\textstyle\sum_{\ell=k}^{n-1} 4\exp\left(2-2\Big(\frac{\ell}{L_{\mathbf{k}}(R_n)}\Big)^{\frac{1}{\alpha}}\right) \leq 4e^2\int_k^{\infty} e^{-\frac{2t}{L_{\mathbf{k}}(R_n)}^{1/\alpha}}dt = \int_{k-1}^{\infty} c^{t^{1/\alpha}}dt,$$

where $c \triangleq \exp(-(L_{\mathbf{k}}(R_n)/2)^{-1/\alpha}) \in (0,1)$. Defining $m \triangleq -\log c > 0$ and $k' \triangleq k-1$, we have

$$\int_{k'}^{\infty} c^{t^{1/\alpha}}dt \leq C_\alpha(k' b^{-1} + b^{\alpha-1}m^{-\alpha})e^{-mk'^{1/\alpha}}, \tag{113}$$

by Li et al. [15, Eq. 50], where $C_\alpha > 0$ is a constant satisfying $(x+y)^\alpha \leq C_\alpha(x^\alpha+y^\alpha)$ for any $x, y > 0$ and b is a known constant depending only on α. Plugging in $k' = \lceil L_{\mathbf{k}}(R_n)\log\big(\frac{2e}{\varepsilon}\big)^{\alpha}\rceil$, we can bound the exponential by

$$e^{-mk'^{1/\alpha}} \leq e^{-mL_k(R_n)^{1/\alpha}\log\left(\frac{2e}{\varepsilon}\right)} = \big(\tfrac{2e}{\varepsilon}\big)^{-mL_{\mathbf{k}}(R_n)^{1/\alpha}}.$$

Note that we can simplify the exponent by $-mL_{\mathbf{k}}(R_n)^{1/\alpha} = -(L_{\mathbf{k}}(R_n)/2)^{-1/\alpha}L_{\mathbf{k}}(R_n)^{1/\alpha} = -2^{1/\alpha}$. Note that $k' = k-1 \geq L(R_n) = 2m^{-\alpha}$. Thus, we can absorb the $b^{\alpha-1}m^{-\alpha}$ term in (113) into k and obtain the following bound

$$\textstyle\sum_{\ell=k}^{n-1} 4\exp\left(2 - 2\left(\frac{\ell}{L_{\mathbf{k}}(R_n)}\right)^{\frac{1}{\alpha}}\right) \leq C'_\alpha k' \left(\frac{2e}{\varepsilon}\right)^{-2^{1/\alpha}},$$

where C'_α depends only on α. Plugging this bound into (112), we have

$$\begin{aligned}
\tfrac{2}{\sqrt{n}}\sqrt{\textstyle\sum_{j=1}^n \min\{\varepsilon^2, \hat{\mu}_j\}} &\leq \tfrac{2}{\sqrt{n}}\sqrt{k\varepsilon^2 + C'_\alpha k\left(\tfrac{\varepsilon}{2e}\right)^{2^{1/\alpha}}} \\
&\overset{(111)}{\leq} \tfrac{2c}{\sqrt{n}}\sqrt{L_{\mathbf{k}}(R_n)\log\left(\tfrac{2e}{\varepsilon}\right)^{\alpha}\left(\varepsilon^2 + C'_\alpha\left(\tfrac{\varepsilon}{2e}\right)^{2^{1/\alpha}}\right)} \\
&\leq \tfrac{2c}{\sqrt{n}}\sqrt{L_{\mathbf{k}}(R_n)\log\left(\tfrac{2e}{\varepsilon}\right)^{\alpha} C''_\alpha \varepsilon^{2^{1/(1\vee\alpha)}}}
\end{aligned}$$

for some constant C''_α that only depends on α and universal positive constant c. To solve (107), it suffices to solve

$$\tfrac{2c}{\sqrt{n}}\sqrt{L_{\mathbf{k}}(R_n)\log\left(\tfrac{2e}{\varepsilon}\right)^{\alpha} C''_\alpha \varepsilon^{2^{1/(1\vee\alpha)}}} \leq \beta\varepsilon^2,$$

which is implied by the looser bound

$$\tfrac{1}{\beta^2}\cdot\tfrac{4c^2}{n}\cdot L_{\mathbf{k}}(R_n)C''_\alpha \leq \varepsilon^2\log\left(\tfrac{2e}{\varepsilon}\right)^{-\alpha}.$$

The solution to (107) (up to log-log factors) is

$$\varepsilon \simeq \tfrac{\log(2e\cdot\beta\sqrt{n})^{\alpha/2}}{\sqrt{n}}\sqrt{\tfrac{4c^2}{\beta^2}\cdot L_{\mathbf{k}}(R_n)C''_\alpha}.$$

□

We proceed to solve (108).

Lemma 10 (Critical Gaussian radius for LOGGROWTH kernels)**.** *Under Assum. 1 and* LOGGROWTH *version of Assum. 3, the Rademacher critical radius satisfies*

$$\delta_n^2 \simeq b^2\tfrac{\log(\frac{2e}{b}\cdot\sqrt{n})^{\alpha}}{n}\cdot L_{\mathbf{k}}(R_n)C''_\alpha.$$

Proof. Following the same logic as in the proof of Lem. 9 but with $\beta = 1/b$ yields the desired bound. □

Finally, it remains to bound (61). We have

$$\begin{aligned}
\eta_{n,\mathbf{k}} &= \tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2 + \mathfrak{W}_{\mathbf{k}}(n, n_{\text{out}}, \delta, \mathfrak{R}_{\text{in}}, \tfrac{\mathfrak{a}}{n_{\text{out}}})) \\
&\leq \tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2 + \sqrt{\log\left(\tfrac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\right)\cdot\left[\log\left(\tfrac{1}{\delta}\right) + \log\mathcal{N}_{\mathbf{k}}(\mathcal{B}_2^d(\mathfrak{R}_{\text{in}}), \tfrac{\mathfrak{a}}{n_{\text{out}}})\right]}) \\
&\leq \tfrac{\mathfrak{a}^2}{n_{\text{out}}}(2 + \sqrt{\log\left(\tfrac{n_{\text{out}}\log(n/n_{\text{out}})}{\delta}\right)\cdot\left[\log\left(\tfrac{1}{\delta}\right) + \mathfrak{C}_d\log(\tfrac{en_{\text{out}}}{\mathfrak{a}})^{\alpha}(\mathfrak{R}_{\text{in}}+1)^{\beta}\right]})
\end{aligned}$$

for some universal positive constant c.

In summary, there exists universal positive constants c_0, c_1, c_2 such that

$$\varepsilon_n^2 \leq c_0\tfrac{\sigma^2}{\|f^\star\|_{\mathbf{k}}^2}\tfrac{\log(2e\cdot\frac{\|f^\star\|_{\mathbf{k}}}{4\sigma}\sqrt{n})^{\alpha}}{n} \quad \delta_n^2 \leq c_1 b^2\tfrac{\log(\frac{2e}{b}\cdot\sqrt{n})^{\alpha/2}}{n} \quad \eta_{n,\mathbf{k}} \leq c_2\tfrac{\mathfrak{a}}{n_{\text{out}}}\log(\tfrac{en_{\text{out}}}{\mathfrak{a}})^{\alpha/2}\mathfrak{R}_{\text{in}}^{\beta/2}.$$

Setting $\lambda' = 2c_2\frac{\mathfrak{a}}{n_{\text{out}}}\log(\frac{en_{\text{out}}}{\mathfrak{a}})^{\alpha/2}$, we have

$$\begin{aligned}
\|\widehat{f}_{\text{KT},\lambda'} - f^\star\|_2^2 &\leq c\left(\varepsilon_n^2 + \lambda' + \eta_{n,\mathbf{k}}\right)\cdot\left[\|f^\star\|_{\mathbf{k}} + 1\right]^2 + c'\delta_n^2 \\
&\leq c\tfrac{\log(2e\cdot\frac{\|f^\star\|_{\mathbf{k}}}{4\sigma}\sqrt{n})^{\alpha}}{n} + \left[\|f^\star\|_{\mathbf{k}} + 1\right]^2\tfrac{c}{n_{\text{out}}}\log(\tfrac{en_{\text{out}}}{\mathfrak{a}})^{\alpha/2} + c'b^2\tfrac{\log(\frac{2e}{b}\cdot\sqrt{n})^{\alpha/2}}{n}.
\end{aligned}$$

F Proof of Thm. 6: KT-KRR

Our first goal is to bound the in-sample prediction error. We relate $\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2$ to $\|\widehat{f}_{\mathrm{full},\lambda} - f^\star\|_n^2$, where the latter quantity has well known properties from standard analyses of the KRR estimator (refer to [28]). Note that regularization parameter λ' of KT based estimator $\widehat{f}_{\mathrm{KT},\lambda'}$ is independently chosen from the regularization parameter λ of the estimator based on original samples $\widehat{f}_{\mathrm{full},\lambda}$. For $\widehat{f}_{\mathrm{full},\lambda}$, we choose the regularization parameter

$$\lambda = 2\varepsilon_n^2, \tag{114}$$

which is known to yield optimal L^2 error rates.

Define the main event of interest,

$$\mathcal{E} \triangleq \{\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \le c(\varepsilon_n^2 + \delta_n^2 + \lambda' + \eta_{n,\mathbf{k}})[\|f^\star\|_{\mathbf{k}} + 1]^2\}.$$

Our goal is to show $\mathcal{E}$ occurs with high probability. For that end, we introduce several additional events that are used throughout this proof.

For some constant $c >$, define the event of an appealing in-sample prediction error of $\widehat{f}_{\mathrm{KT},\lambda'}$,

$$\mathcal{E}_{\mathrm{KT},n}(t) \triangleq \left\{\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2 \le c\big[t^2 + \lambda' + \eta_{n,\mathbf{k}}\big] \cdot (\|f^\star\|_{\mathbf{k}} + 1)^2\right\} \quad \text{for} \quad t \ge \varepsilon_n.$$

where $\eta_{n,\mathbf{k}}$ is defined in (61). Recall $\mathcal{E}_{\mathrm{KT},\delta}$ is the event where KT-COMPRESS++ succeeds as defined by (23).

Further as $f^\star$ and $\widehat{f}_{\mathrm{KT},\lambda'}$ are both in $\{f \in \mathcal{H} : \|f\|_{\mathbf{k}} \le R\}$, we may deduce that all the functions under consideration satisfies $\|f\|_\infty \le \|\mathbf{k}\|_\infty \|f\|_{\mathbf{k}} \le \|\mathbf{k}\|_\infty R$ where $\|\mathbf{k}\|_\infty < \infty$. Accordingly, we define a uniform concentration event,

$$\mathcal{E}'_{\mathrm{conc}} \triangleq \{\sup_{f\in\mathcal{F}} \big|\|f\|_2^2 - \|f\|_n^2\big| \le \|f\|_2^2/2 + \delta_n^2/2\} \quad \text{where} \quad \mathcal{F} = \{f \in \mathcal{H} : \|f\|_\infty \le 2\|\mathbf{k}\|_\infty R\}. \tag{115}$$

Event (115) is analogous to the event $\mathcal{E}_{\mathrm{conc}}$ previously defined in (69) when dealing with finite rank kernels.

We first show that

$$\mathcal{E}_{\mathrm{KT},n}(\varepsilon_n \vee \delta_n) \cap \mathcal{E}'_{\mathrm{conc}} \subseteq \mathcal{E}. \tag{116}$$

Notice that almost surely we have

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_\infty \le 2\|\mathbf{k}\|_\infty R,$$

thereby implying

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 \le 2\,\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2 + \delta_n^2 \quad \text{on the event } \mathcal{E}'_{\mathrm{conc}}. \tag{117}$$

Next invoking the event $\mathcal{E}_{\mathrm{KT},n}(\varepsilon_n \vee \delta_n)$ along with (117), we have

$$\begin{aligned}\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_2^2 &\le 2c[(\varepsilon_n \vee \delta_n)^2 + \lambda' + \eta_{n,\mathbf{k}}] \cdot (\|f^\star\|_{\mathbf{k}} + 1)^2 + \delta_n^2 \\ &\le c(\varepsilon_n^2 + \delta_n^2 + \lambda' + \eta_{n,\mathbf{k}})[\|f^\star\|_{\mathbf{k}} + 1]^2.\end{aligned}$$

which recovers the event of $\mathcal{E}$.

The remaining task is to show $\mathcal{E}$ is of high-probability, which amounts to showing events $\mathcal{E}_{\mathrm{KT},n}(t)$ and $\mathcal{E}'_{\mathrm{conc}}$ are of high-probability by reflecting on (116). From [28, Thm. 14.1], we may immediately derive

$$\mathbb{P}(\mathcal{E}'_{\mathrm{conc}}) \ge 1 - c_1 e^{-c_2 \frac{n\delta_n^2}{\|\mathbf{k}\|_\infty^2 R^2}}$$

for some constants $c_1, c_2 > 0$.

We further claim that

$$\mathbb{P}(\mathcal{E}_{\mathrm{KT},n}(t) \mid \mathcal{S}_{\mathrm{in}}) \ge 1 - \delta - e^{-\frac{nt^2}{c_0\sigma^2}} - c_1 e^{-c_2 \frac{n\|f^\star\|_{\mathbf{k}}^2 t^2}{\sigma^2}} \tag{118}$$

for some constants $c_0, c_1, c_2 > 0$. Proof of claim (118) is deferred to App. F.1. Plugging in $t = \varepsilon_n \vee \delta_n$ into (118), and invoking inequality $\varepsilon_n \vee \delta_n \geq \delta_n$ so as to decouple the dependence on $\mathcal{S}_{\text{in}}$, we have

$$\mathbb{P}(\mathcal{E}_{\text{KT},n}(\varepsilon_n \vee \delta_n) \mid \mathcal{S}_{\text{in}}) \geq 1 - \delta - e^{-\frac{n\delta_n^2}{c_0\sigma^2}} - c_1 e^{-c_2 \frac{\|f^\star\|_{\mathbf{k}}^2 n\delta_n^2}{\sigma^2}}$$

which further implies

$$\mathbb{P}(\mathcal{E}_{\text{KT},n}(\varepsilon_n \vee \delta_n)) \geq 1 - \delta - e^{-\frac{n\delta_n^2}{c_0\sigma^2}} - c_1 e^{-c_2 \frac{\|f^\star\|_{\mathbf{k}}^2 n\delta_n^2}{\sigma^2}}.$$

Putting the pieces together, for some constants $c_0, c_1 > 0$, we have

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta - c_0 e^{-c_1 \frac{n\delta_n^2}{\sigma^2 \wedge (\sigma^2/\|f^\star\|_{\mathbf{k}}^2) \wedge (\|\mathbf{k}\|_\infty^2 R^2)}}. \tag{119}$$

Overall, (116) and (119) collectively yields the desired result.

F.1 Proof of claim (118)

To prove claim (118), we introduce two new intermediary and technical events. For some positive constant c_0, define the event [5] when in-sample prediction error of $\widehat{f}_{\text{full},\lambda}$ is appealing

$$\mathcal{E}_{\text{full},n}(t) \triangleq \left\{ \|\widehat{f}_{\text{full},\lambda} - f^\star\|_n^2 \leq 3c_0 \|f^\star\|_{\mathbf{k}}^2 t^2 \right\} \quad \text{for} \quad t \geq \varepsilon_n. \tag{120}$$

The second intermediary event, denoted as $\mathcal{E}_{\widehat{\Delta}_{\text{KT}}}(t)$, is the intersection of (128) and (129), which we do not elaborate here due to its technical nature—event $\mathcal{E}_{\widehat{\Delta}_{\text{KT}}}(t)$ plays an analogous role to $\mathcal{A}_{\text{KT}}^c \cap \mathcal{B}_{\text{KT}}^c$ defined in (82) and (95) respectively.

Our goal here is two-folds: first is to show

$$\{\mathcal{E}_{\text{full},n}(t) \cap \mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\widehat{\Delta}_{\text{KT}}}(t)\} \quad \Longrightarrow \quad \mathcal{E}_{\text{KT},n}(t)$$

and second is to prove the following bound

$$\mathbb{P}\Big(\mathcal{E}_{\text{full},n}(t) \cap \mathcal{E}_{\text{KT},\delta} \cap \mathcal{E}_{\widehat{\Delta}_{\text{KT}}}(t) \mid \mathcal{S}_{\text{in}}\Big) \geq 1 - \delta - e^{-\frac{nt^2}{c_0\sigma^2}} - c_1 e^{-c_2 \frac{n\|f^\star\|_{\mathbf{k}}^2 t^2}{\sigma^2}},$$

from which (118) follows. Note that Wainwright [28, Thm. 13.17] show

$$\mathbb{P}(\mathcal{E}_{\text{full},n}(t)) \geq 1 - c_1 e^{-c_2 \frac{n\|f^\star\|_{\mathbf{k}}^2 t^2}{\sigma^2}}$$

for some constants $c_1, c_2 > 0$ and that $\mathbb{P}(\mathcal{E}_{\text{KT},\delta} \mid \mathcal{S}_{\text{in}}) \geq 1 - \delta$. So it remains to bound the probability of event $\mathcal{E}_{\widehat{\Delta}_{\text{KT}}}(t)$, which we show below.

Given f, define the following quantities

$$L_n(f) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n (f^2(x_i) - 2f(x_i)y_i) + \tfrac{1}{n}\sum_{i=1}^n y_i^2 \quad \text{and}$$
$$L_{n_{\text{out}}}(f) \triangleq \tfrac{1}{n_{\text{out}}}\textstyle\sum_{i=1}^{n_{\text{out}}} (f^2(x_i') - 2f(x_i')y_i') + \tfrac{1}{n}\sum_{i=1}^n y_i^2.$$

In the sequel, we repeatedly make use of the following fact: on event $\mathcal{E}_{\text{KT},\delta}$ defined in (23), we have

$$|L_n(f) - L_{n_{\text{out}}}(f)| \leq \big(\|f\|_{\mathbf{k}}^2 + 2\big) \cdot \eta_{n,\mathbf{k}} \quad \text{for all non-zero} \quad f \in \mathcal{H}. \tag{121}$$

The claim of (121) is deferred to the end of this section. Given f, we can show with some algebra that

$$L_n(f) = \tfrac{1}{n}\textstyle\sum_{i=1}^n (f(x_i) - y_i)^2 = \|f - f^\star\|_n^2 - \tfrac{2}{n}\langle Z, \boldsymbol{\xi}\rangle + \tfrac{1}{n}\sum_{i=1}^n \xi_i^2, \tag{122}$$

where $Z \triangleq (f(x_1) - f^\star(x_1), \ldots, f(x_n) - f^\star(x_n))$ and $\boldsymbol{\xi} \triangleq (\xi_1, \ldots, \xi_n)$ are vectors in $\mathbb{R}^n$. Define the shorthands

$$\widehat{\Delta}_{\text{KT}} \triangleq \widehat{f}_{\text{KT},\lambda'} - f^\star \quad \text{and} \quad \widehat{\Delta}_{\text{full}} \triangleq \widehat{f}_{\text{full},\lambda} - f^\star.$$

[5] Since the input points in $\mathcal{S}_{\text{in}}$ are fixed, the randomness in $\widehat{f}_{\text{full},\lambda}$ originates entirely from the randomness of the noise variables $\boldsymbol{\xi}$.

In the sequel, we use the following shorthands:

$$Z_{full} \triangleq (\widehat{\Delta}_{\text{full}}(x_1), \ldots, \widehat{\Delta}_{\text{full}}(x_n)) \quad \text{and} \quad Z_{KT} \triangleq (\widehat{\Delta}_{\text{KT}}(x_1), \ldots, \widehat{\Delta}_{\text{KT}}(x_n)).$$

Now for the main argument to bound $\|\widehat{f}_{\text{KT},\lambda'} - f^\star\|_n^2$. When $\|\widehat{\Delta}_{\text{KT}}\|_n < t$, we immediately have $\|\widehat{\Delta}_{\text{KT}}\|_n^2 < t^2$, which implies (118). Thus, we may assume that $\|\widehat{\Delta}_{\text{KT}}\|_n \geq t$. Note that

$$\begin{aligned}\|\widehat{\Delta}_{\text{KT}}\|_n^2 &\overset{(122)}{=} L_n(\widehat{f}_{\text{KT},\lambda'}) + \tfrac{2}{n}\langle Z_{KT}, \xi\rangle - \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 \\ &= L_n(\widehat{f}_{\text{full},\lambda}) + \Big[L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big] + \tfrac{2}{n}\langle Z_{KT}, \xi\rangle - \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2.\end{aligned}$$

Given the optimality of $\widehat{f}_{\text{full},\lambda}$ on the objective (4), we have

$$L_n(\widehat{f}_{\text{full},\lambda}) \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 + \lambda\Big\{\|f^\star\|_{\mathbf{k}}^2 - \|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2\Big\} \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 + \lambda\|f^\star\|_{\mathbf{k}}^2,$$

where the last inequality follows trivially from dropping the $-\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2$ term. Thus,

$$\begin{aligned}\|\widehat{\Delta}_{\text{KT}}\|_n^2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 + \lambda\|f^\star\|_{\mathbf{k}}^2 + \Big[L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big] + \tfrac{2}{n}\langle Z_{KT}, \xi\rangle - \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 \\ &\leq \tfrac{2}{n}\langle Z_{KT}, \xi\rangle + \lambda\|f^\star\|_{\mathbf{k}}^2 + \Big[L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big]. \qquad (123)\end{aligned}$$

Using standard arguments to bound the term $\frac{2}{n}\langle Z_{KT}, \xi\rangle$, we claim that on the event $\mathcal{E}_{\widehat{\Delta}_{\text{KT}}}$, we have

$$\|\widehat{\Delta}_{\text{KT}}\|_n^2 \leq ct^2(\|f^\star\|_{\mathbf{k}} + 1)^2 + c'\Big[L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big] \qquad (124)$$

for some positive constants c, c', and that $\mathbb{P}(\mathcal{E}_{\widehat{\Delta}_{\text{KT}}} \mid \mathcal{S}_{\text{in}}) \geq 1 - e^{-\frac{nt^2}{2\sigma^2}}$. We defer the proof of claim (124) to the end of this section.

Now we bound the stochastic term $\Big[L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big]$ in (124)—first observe the following decomposition:

$$L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda}) = \Big(L_n(\widehat{f}_{\text{KT},\lambda'}) - L_{n_{\text{out}}}(\widehat{f}_{\text{KT},\lambda'})\Big) + \Big(L_{n_{\text{out}}}(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda})\Big).$$

On the event $\mathcal{E}_{\text{KT},\delta}$ (23), the first term in the display can be bounded by

$$L_n(\widehat{f}_{\text{KT},\lambda'}) - L_{n_{\text{out}}}(\widehat{f}_{\text{KT},\lambda'}) \overset{(121)}{\leq} (\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2 + 2)\,\eta_{n,\mathbf{k}}.$$

Note that $\widehat{f}_{\text{KT},\lambda'}$ is the solution to the following optimization problem,

$$\min_{f\in\mathcal{H}(\mathbf{k})} L_{n_{\text{out}}}(f) + \lambda'\|f\|_{\mathbf{k}}^2,$$

so the second term in the display can be bounded by the following basic inequality

$$\begin{aligned}&L_{n_{\text{out}}}(\widehat{f}_{\text{KT},\lambda'}) + \lambda'\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2 \leq L_{n_{\text{out}}}(\widehat{f}_{\text{full},\lambda}) + \lambda'\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \\ \text{so that}\quad &L_{n_{\text{out}}}(\widehat{f}_{\text{KT},\lambda'}) - L_{n_{\text{out}}}(\widehat{f}_{\text{full},\lambda}) \leq \lambda'\Big\{\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 - \|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2\Big\}.\end{aligned}$$

Thus, on event $\mathcal{E}_{\text{KT},\delta}$, we have

$$\begin{aligned}L_n(\widehat{f}_{\text{KT},\lambda'}) - L_n(\widehat{f}_{\text{full},\lambda}) &\leq (\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2 + 2)\,\eta_{n,\mathbf{k}} + \lambda'\Big\{\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 - \|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2\Big\} \\ &= 2\eta_{n,\mathbf{k}} + \lambda'\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 + \{\eta_{n,\mathbf{k}} - \lambda'\}\cdot\|\widehat{f}_{\text{KT},\lambda'}\|_{\mathbf{k}}^2 \\ &\overset{(i)}{\leq} 2\eta_{n,\mathbf{k}} + \lambda'\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \overset{(ii)}{\leq} \lambda'(\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 + 1)\end{aligned}$$

where steps (i) and (ii) both follow from the fact that $\lambda' \geq 2\eta_{n,\mathbf{k}}$ (see assumptions in Thm. 6). To bound $\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2$, we use the following lemma:

Lemma 11 (RKHS norm of $\widehat{f}_{\text{full},\lambda}$)**.** *On event $\mathcal{E}_{full,n}$ (120), we have the following bound*

$$\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \leq c_0(\|f^\star\|_{\mathbf{k}} + 1)^2 \qquad (125)$$

for some constant $c_0 > 0$.

See App. F.2 for the proof. Putting things together, we have

$$L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda}) \le c\lambda'(\|f^\star\|_{\mathbf{k}}+1)^2$$

for some constant c—substituting this bound into (124) yields

$$\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2 \le ct^2(\|f^\star\|_{\mathbf{k}}+1)^2 + c'\lambda'(\|f^\star\|_{\mathbf{k}}+1)^2,$$

for some constants c, c', which directly implies (118), i.e. implying

$$\{\mathcal{E}_{\mathrm{full},n}(t) \cap \mathcal{E}_{\mathrm{KT},\delta} \cap \mathcal{E}_{\widehat{\Delta}_{\mathrm{KT}}}(t)\} \quad \Longrightarrow \quad \mathcal{E}_{\mathrm{KT},n}(t).$$

Proof of claim (121). Given f, define the function

$$\ell'_f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}, \quad \text{where} \quad \ell'_f(x,y) \triangleq f^2(x) - 2y\cdot f(x) \tag{126}$$

and note that

$$L_n(f) - L_{n_{\mathrm{out}}}(f) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell'_f(x_i,y_i) - \tfrac{1}{n_{\mathrm{out}}}\sum_{i=1}^{n_{\mathrm{out}}} \ell'_f(x'_i,y'_i).$$

We first prove a generic technical lemma:

Lemma 12 (KT-COMPRESS++ approximation bound using $\mathbf{k}_{\mathrm{RR}}$)**.** *Suppose $f_1, f_2 \in \mathcal{H}(\mathbf{k})$ and $a, b \in \mathbb{R}$. Then the function*

$$\ell_{f_1,f_2} : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}, \quad \textit{where} \quad \ell_{f_1,f_2}(x,y) \triangleq a\cdot f_1(x)f_2(x) + b\cdot y f_1(x) \tag{127}$$

lies in the RKHS $\mathcal{H}(\mathbf{k}_{\mathrm{RR}})$. Moreover, on event $\mathcal{E}_{KT,\delta}$, we have

$$\mathbb{P}_{\mathrm{in}}\ell_{f_1,f_2} - \mathbb{Q}_{\mathrm{out}}\ell_{f_1,f_2} \le (|a|\cdot\|f_1\|_{\mathbf{k}}\|f_2\|_{\mathbf{k}} + |b|\cdot\|f_2\|_{\mathbf{k}})\cdot\eta_{n,\mathbf{k}}.$$

uniformly for all non-zero $f_1, f_2 \in \mathcal{H}(\mathbf{k})$.

See App. F.3 for the proof. Applying the lemma with $f_1 \triangleq f$, $f_2 \triangleq g$ and $a=1, b=-2$, we have

$$\mathbb{P}_{\mathrm{in}}\ell'_f - \mathbb{Q}_{\mathrm{out}}\ell'_f \le (\|f\|_{\mathbf{k}}^2 + 2)\cdot\eta_{n,\mathbf{k}},$$

which combined with the observation (127) yields the desired claim.

Proof of claim (124). Case I: First suppose that $\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} \le 1$. Recall that $\|\widehat{\Delta}_{\mathrm{KT}}\|_n \ge t \ge \varepsilon_n$ by assumption. Thus, we may apply [28, Lem. 13.12] to obtain

$$\tfrac{1}{n}\langle Z_{KT}, \boldsymbol{\xi}\rangle \le 2\|\widehat{\Delta}_{\mathrm{KT}}\|_n t \quad \text{w.p. at least} \quad 1 - e^{-\frac{nt^2}{2\sigma^2}} \tag{128}$$

Plugging the above bound into (123), we have with probability at least $1-e^{-\frac{nt^2}{2\sigma^2}}$:

$$\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \le 4\|\widehat{\Delta}_{\mathrm{KT}}\|_n t + \lambda\|f^\star\|_{\mathbf{k}}^2 + \Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big].$$

We can solve for $\|\widehat{\Delta}_{\mathrm{KT}}\|_n$ using the quadratic formula. Specifically, if $a, b \ge 0$ and $x^2 - ax - b \le 0$, then $x \le a + \sqrt{b}$. Thus, we have with probability at least $1-e^{-\frac{n\varepsilon_n^2}{2\sigma^2}}$:

$$\begin{aligned}\|\widehat{\Delta}_{\mathrm{KT}}\|_n &\le a + \sqrt{b}, \quad \text{where}\\ a &\triangleq 4t \quad \text{and}\\ b &\triangleq \lambda\|f^\star\|_{\mathbf{k}}^2 + \Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big].\end{aligned}$$

Using the fact that $(a+\sqrt{b})^2 \le 2a^2 + 2b$, we have with probability at least $1-e^{-\frac{nt^2}{2\sigma^2}}$:

$$\begin{aligned}\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2 &\le 32t^2 + 2\lambda\|f^\star\|_{\mathbf{k}}^2 + 2\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big]\\ &\overset{(114)}{\le} ct^2(\|f^\star\|_{\mathbf{k}}+1)^2 + 2\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big]\end{aligned}$$

Case II: Otherwise, we may assume that $\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} > 1$. Now we apply [28, Thm. 13.23] to obtain

$$\tfrac{1}{n}\langle Z_{KT}, \boldsymbol{\xi}\rangle \le 2t\|\widehat{\Delta}_{\mathrm{KT}}\|_n + 2t^2\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} + \tfrac{1}{16}\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \quad \text{w.p. at least} \quad 1 - c_1 e^{-\frac{nt^2}{c_2\sigma^2}}, \tag{129}$$

for some universal positive constants c_1, c_2. Plugging the above bound into (123) and collecting terms, we have with probability at least $1 - c_1 e^{-\frac{nt^2}{c_2\sigma^2}}$:

$$\tfrac{7}{8}\|\widehat{\Delta}_{\mathrm{KT}}\|_n^2 \le 4t\|\widehat{\Delta}_{\mathrm{KT}}\|_n + 4t^2\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}} + \lambda\|f^\star\|_{\mathbf{k}}^2 + \Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big].$$

Solving for $\|\widehat{\Delta}_{\mathrm{KT}}\|_n$ using the quadratic formula, we have with probability at least $1 - c_1 e^{-\frac{nt^2}{c_2\sigma^2}}$:

$$\begin{aligned}
\|\widehat{\Delta}_{\mathrm{KT}}\|_n &\le a + \sqrt{b}, \quad \text{where} \\
a &\triangleq \tfrac{32}{7}t \quad \text{and} \\
b &\triangleq \tfrac{32}{7}t^2\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}^2 + \tfrac{8}{7}\lambda\|f^\star\|_{\mathbf{k}}^2 + \tfrac{8}{7}\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big].
\end{aligned}$$

Using the fact that $(a+\sqrt{b})^2 \le 2a^2 + 2b$, we have with probability at least $1 - c_1 e^{-\frac{nt^2}{c_2\sigma^2}}$:

$$\begin{aligned}
\|\widehat{f}_{\mathrm{KT},\lambda'} - f^\star\|_n^2 &\le 42t^2 + 10t^2\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}^2 + 2.3\lambda\|f^\star\|_{\mathbf{k}}^2 + 2.3\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big] \\
&\overset{(i)}{\le} 42t^2 + 10t^2\|\widehat{\Delta}_{\mathrm{KT}}\|_{\mathbf{k}}^2 + 4.6t^2\|f^\star\|_{\mathbf{k}}^2 + 2.3\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big] \\
&\overset{(114),(125)}{\le} c_3 t^2(\|f^\star\|_{\mathbf{k}} + 1)^2 + c_4\Big[L_n(\widehat{f}_{\mathrm{KT},\lambda'}) - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big]
\end{aligned}$$

for some positive constants c_3, c_4, where step (i) follows from that fact that $\lambda = 2\varepsilon_n^2$ by (114).

F.2 Proof of Lem. 11: RKHS norm of $\widehat{f}_{\mathrm{full},\lambda}$

Given the optimality of $\widehat{f}_{\mathrm{full},\lambda}$ on the objective (4), we have the following basic inequality

$$\begin{aligned}
& L_n(\widehat{f}_{\mathrm{full},\lambda}) + \lambda\|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}}^2 \le \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 + \lambda\|f^\star\|_{\mathbf{k}}^2 \\
\implies \quad & \|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}}^2 \le \|f^\star\|_{\mathbf{k}}^2 + \tfrac{1}{\lambda}\Big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 - L_n(\widehat{f}_{\mathrm{full},\lambda})\Big).
\end{aligned}$$

Since $\|\widehat{f}_{\mathrm{full},\lambda} - f^\star\|_n^2 \ge 0$, we also have the trivial lower bound

$$\begin{aligned}
L_n(\widehat{f}_{\mathrm{full},\lambda}) &\overset{(122)}{=} \|\widehat{f}_{\mathrm{full},\lambda} - f^\star\|_n^2 - \tfrac{2}{n}\langle Z_{full}, \boldsymbol{\xi}\rangle + \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2 \\
&\ge -\tfrac{2}{n}\langle Z_{full}, \boldsymbol{\xi}\rangle + \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i^2.
\end{aligned}$$

Thus,

$$\|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}}^2 \le \|f^\star\|_{\mathbf{k}}^2 + \tfrac{1}{\lambda}\big(\tfrac{2}{n}\langle Z_{full}, \xi\rangle\big) \tag{130}$$

and it remains to bound $\frac{2}{n}\langle Z_{full}, \xi\rangle$.

Case I: First, suppose that $\|\widehat{\Delta}_{\mathrm{full}}\|_{\mathbf{k}} > 1$. Then we may apply [28, Lem. 13.23] to obtain

$$\tfrac{1}{n}\langle Z_{full}, \xi\rangle \le 2\varepsilon_n\|\widehat{\Delta}_{\mathrm{full}}\|_n + 2\varepsilon_n^2\|\widehat{\Delta}_{\mathrm{full}}\|_{\mathbf{k}} + \tfrac{1}{16}\|\widehat{\Delta}_{\mathrm{full}}\|_n^2 \quad \text{w.p. at least} \quad 1 - c_1 e^{-\frac{n\varepsilon_n^2}{c_2\sigma^2}}.$$

Combining this bound with (130), we have with probability at least $1 - c_1 e^{-\frac{n\varepsilon_n^2}{c_2\sigma^2}}$:

$$\begin{aligned}
\|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}}^2 &\le \|f^\star\|_{\mathbf{k}}^2 + \tfrac{2\varepsilon_n^2}{\lambda}\|\widehat{\Delta}_{\mathrm{full}}\|_{\mathbf{k}} + \tfrac{2}{\lambda}\Big(2\varepsilon_n\|\widehat{\Delta}_{\mathrm{full}}\|_n + \tfrac{1}{16}\|\widehat{\Delta}_{\mathrm{full}}\|_n^2\Big) \\
&\overset{(i)}{\le} \|f^\star\|_{\mathbf{k}}^2 + \tfrac{2\varepsilon_n^2}{\lambda}(\|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}} + \|f^\star\|_{\mathbf{k}}) + \tfrac{2}{\lambda}\Big(2\varepsilon_n\|\widehat{\Delta}_{\mathrm{full}}\|_n + \tfrac{1}{16}\|\widehat{\Delta}_{\mathrm{full}}\|_n^2\Big) \\
&\overset{(114)}{=} \|f^\star\|_{\mathbf{k}}^2 + \|\widehat{f}_{\mathrm{full},\lambda}\|_{\mathbf{k}} + \|f^\star\|_{\mathbf{k}} + \tfrac{2}{\lambda}\Big(2\varepsilon_n\|\widehat{\Delta}_{\mathrm{full}}\|_n + \tfrac{1}{16}\|\widehat{\Delta}_{\mathrm{full}}\|_n^2\Big),
\end{aligned}$$

where step (i) follows from triangle inequality. Solving for $\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}$ using the quadratic formula, we have

$$\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \leq 2 + \|f^\star\|_{\mathbf{k}}^2 + \|f^\star\|_{\mathbf{k}} + \tfrac{2}{\lambda}\Big(2\varepsilon_n\|\widehat{\Delta}_{\text{full}}\|_n + \tfrac{1}{16}\|\widehat{\Delta}_{\text{full}}\|_n^2\Big).$$

On the event $\mathcal{E}_{\text{full},n}$ (120), we have $\|\widehat{\Delta}_{\text{full}}\|_n \leq c\|f^\star\|_{\mathbf{k}}\varepsilon_n$ for some positive constant c, which implies the claimed bound (125) after some algebra.

Case II(a): Otherwise, assume $\|\widehat{\Delta}_{\text{full}}\|_{\mathbf{k}} \leq 1$ and $\|\widehat{\Delta}_{\text{full}}\|_n \leq \varepsilon_n$. Applying [28, Thm. 2.26] to the function $\sup_{\substack{\|g\|_{\mathbf{k}}\leq 1\\ \|g\|_n \leq \varepsilon_n}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i g(x_i)\right|$, we have

$$\tfrac{1}{n}\langle Z_{full}, \boldsymbol{\xi}\rangle \leq \tfrac{\varepsilon_n^2}{2} \quad \text{w.p. at least} \quad 1 - e^{-\frac{n\varepsilon_n^2}{8\sigma^2}}$$

Combining this bound with (130), we obtain

$$\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \leq \|f^\star\|_{\mathbf{k}}^2 + \tfrac{1}{\lambda}\varepsilon_n^2 \overset{(114)}{=} \|f^\star\|_{\mathbf{k}}^2 + \tfrac{1}{2},$$

which immediately implies the claimed bound (125).

Case II(b): Finally, assume $\|\widehat{\Delta}_{\text{full}}\|_{\mathbf{k}} \leq 1$ and $\|\widehat{\Delta}_{\text{full}}\|_n > \varepsilon_n$. Applying [28, Lem. 13.12] with $u = \varepsilon_n$, we have

$$\tfrac{1}{n}\langle Z_{full}, \boldsymbol{\xi}\rangle \leq 2\varepsilon^2 \quad \text{w.p. at least} \quad 1 - e^{-\frac{n\varepsilon_n^2}{2\sigma^2}}.$$

Combining this bound with (130), we obtain

$$\|\widehat{f}_{\text{full},\lambda}\|_{\mathbf{k}}^2 \leq \|f^\star\|_{\mathbf{k}}^2 + \tfrac{4}{\lambda}\varepsilon_n^2 \overset{(114)}{=} \|f^\star\|_{\mathbf{k}}^2 + 2,$$

which immediately implies the claimed bound (125).

F.3 Proof of Lem. 12: KT-COMPRESS++ approximation bound using $\mathbf{k}_{\text{RR}}$

By Grünewälder [12, Lem. 4], ℓ_{f_1,f_2} lies in the RKHS $\mathcal{H}(\mathbf{k}_{\text{RR}})$, which is a direct sum of two RKHS:

$$\mathcal{H}(\mathbf{k}_{\text{RR}}) = \mathcal{H}(\mathbf{k}_1) \oplus \mathcal{H}(\mathbf{k}_2),$$

where $\mathbf{k}_1, \mathbf{k}_2 : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ are the kernels defined by

$$\mathbf{k}_1((x_1,y_1),(x_2,y_2)) \triangleq \mathbf{k}^2(x_1,x_2) \quad \text{and} \quad \mathbf{k}_2((x_1,y_1),(x_2,y_2)) \triangleq \mathbf{k}(x_1,x_2)\cdot y_1y_2.$$

Applying Lem. 2 with

$$\mathcal{Z} = \mathcal{X}\times\mathcal{Y}, \quad \mathbf{k}_{\text{ALG}} = \mathbf{k}_{\text{RR}}, \quad \text{and} \quad \epsilon^\star = \tfrac{(\|\mathbf{k}\|_{\infty,\text{in}}^{1/2} + Y_{\max})^2}{n_{\text{out}}},$$

yields the following bound on event $\mathcal{E}_{\text{KT},\delta}$ (23):

$$\sup_{\substack{h\in\mathcal{H}(\mathbf{k}_{\text{RR}}):\\ \|h\|_{\mathbf{k}_{\text{RR}}}\leq 1}} |(\mathbb{P}_{\text{in}} - \mathbb{Q}_{\text{out}})h| \leq 2\epsilon^\star + \tfrac{\|\mathbf{k}_{\text{RR}}\|_{\infty,\text{in}}^{1/2}}{n_{\text{out}}}\cdot \mathfrak{W}_{\mathbf{k}_{\text{RR}}}(n, n_{\text{out}}, \delta, \mathfrak{R}_{\text{in}}, \epsilon^\star).$$

We claim that

$$\|\mathbf{k}_{\text{RR}}\|_{\infty,\text{in}}^{1/2} \leq \|\mathbf{k}\|_{\infty,\text{in}} + Y_{\max}^2 \quad \text{and} \tag{131}$$

$$\log \mathcal{N}^\dagger_{\mathbf{k}_{\text{RR}}}(\mathcal{S}_{\text{in}}, \epsilon^\star) \leq c\cdot \log \mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \tfrac{\|\mathbf{k}\|_{\infty,\text{in}}^{1/2} + Y_{\max}}{n_{\text{out}}}), \tag{132}$$

for some positive constant c, where $\mathcal{N}^\dagger_{\mathbf{k}_{\text{RR}}}$ is the cardinality of the cover of $\mathcal{B}^\dagger_{\mathbf{k}_{\text{RR}}} \triangleq \left\{\ell'_f/\|\ell'_f\|_{\mathbf{k}_{\text{RR}}} : f\in\mathcal{H}(\mathbf{k})\right\}$ for ℓ'_f defined by (126). Proof of the claims (131) and (132) are deferred to the end of this section. By definition of $\mathfrak{W}_{\mathbf{k}_{\text{RR}}}$, we have

$$\mathfrak{W}_{\mathbf{k}_{\text{RR}}}(n, n_{\text{out}}, \delta, \mathfrak{R}_{\text{in}}, \epsilon^\star) \overset{(28)}{\leq} \sqrt{c}\cdot \mathfrak{W}_{\mathbf{k}}(n, n_{\text{out}}, \delta, \mathfrak{R}_{\text{in}}, \tfrac{\|\mathbf{k}\|_{\infty,\text{in}}^{1/2} + Y_{\max}}{n_{\text{out}}}) \triangleq \mathfrak{W}'_{\mathbf{k}}.$$

On event $\mathcal{E}_{\mathrm{KT},\delta}$, we have

$$\sup_{\substack{h\in\mathcal{H}(\mathbf{k}_{\mathrm{RR}}):\\ \|h\|_{\mathbf{k}_{\mathrm{RR}}}\le 1}} |(\mathbb{P}_{\mathrm{in}}-\mathbb{Q}_{\mathrm{out}})h| \le \frac{2(\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}+Y_{\max})^2}{n_{\mathrm{out}}} + \frac{\|\mathbf{k}\|_{\infty,\mathrm{in}}+Y_{\max}^2}{n_{\mathrm{out}}}\cdot\mathfrak{W}'_{\mathbf{k}}$$
$$\overset{(i)}{=} \frac{\|\mathbf{k}\|_{\infty,\mathrm{in}}+Y_{\max}^2}{n_{\mathrm{out}}}\cdot[2+\mathfrak{W}'_{\mathbf{k}}], \tag{133}$$

where step (i) follows from the fact that $(\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}+Y_{\max})^2 \le 2(\|\mathbf{k}\|_{\infty,\mathrm{in}}+Y_{\max}^2)$.

Since f_1, f_2 are non-zero, we have $\|\ell_{f_1,f_2}\|_{\mathbf{k}} > 0$. Thus, the function $h \triangleq \ell_f/\|\ell_f\|_{\mathbf{k}_{\mathrm{RR}}} \in \mathcal{H}(\mathbf{k}_{\mathrm{RR}})$ is well-defined and satisfies $\|h\|_{\mathbf{k}_{\mathrm{RR}}} = 1$. Applying (133), we obtain

$$|\mathbb{P}_{\mathrm{in}}h - \mathbb{Q}_{\mathrm{out}}h| \le \frac{\|\mathbf{k}\|_{\infty,\mathrm{in}}+Y_{\max}^2}{n_{\mathrm{out}}}\cdot(2+\mathfrak{W}'_{\mathbf{k}}) \quad \text{on event} \quad \mathcal{E}_{\mathrm{KT},\delta}.$$

Multiplying both sides by $\|\ell_f\|_{\mathbf{k}_{\mathrm{RR}}}$ and noting that

$$\begin{aligned}\|\ell_{f,g}\|_{\mathbf{k}_{\mathrm{RR}}}^2 &= \|a\cdot f_1 f_2\|_{\widehat{\mathcal{H}\odot\mathcal{H}}}^2 + \|b\cdot f_2\otimes\langle\cdot,1\rangle_{\mathbb{R}}\|_{\mathcal{H}\otimes\mathcal{R}}^2\\ &\le a^2\|f_1\|_{\mathbf{k}}^2\|f_2\|_{\mathbf{k}}^2 + b^2\|f_2\|_{\mathbf{k}}^2\\ &\le (|a|\cdot\|f_1\|_{\mathbf{k}}\|f_2\|_{\mathbf{k}} + |b|\cdot\|f_2\|_{\mathbf{k}})^2,\end{aligned}$$

we have on event $\mathcal{E}_{\mathrm{KT},\delta}$,

$$\mathbb{P}_{\mathrm{in}}\ell_{f_1,f_2} - \mathbb{Q}_{\mathrm{out}}\ell_{f_1,f_2} \le (|a|\cdot\|f_1\|_{\mathbf{k}}\|f_2\|_{\mathbf{k}} + |b|\cdot\|f_2\|_{\mathbf{k}})\cdot\frac{\|\mathbf{k}\|_{\infty,\mathrm{in}}+Y_{\max}^2}{n_{\mathrm{out}}}\cdot(2+\mathfrak{W}'_{\mathbf{k}}),$$

which directly implies the bound (121) after applying the shorthand (61).

Proof of (131) Define $Y_{\max} \triangleq \sup_{y\in(\mathcal{S}_{\mathrm{in}})_y} y$. We have

$$\begin{aligned}\|\mathbf{k}_{\mathrm{ALG}}\|_{\infty,\mathrm{in}} &= \sup_{(x_1,y_1),(x_2,y_2)\in\mathcal{S}_{\mathrm{in}}}\{\mathbf{k}(x_1,x_2)^2 + \mathbf{k}(x_1,x_2)\cdot y_1y_2 + (y_1y_2)^2\}\\ &\le \sup_{x_1,x_2\in(\mathcal{S}_{\mathrm{in}})_x}\mathbf{k}(x_1,x_2)^2 + \sup_{x_1,x_2\in(\mathcal{S}_{\mathrm{in}})_x}\mathbf{k}(x_1,x_2)\cdot\sup_{y_1,y_2\in(\mathcal{S}_{\mathrm{in}})_y} y_1y_2\\ &\quad + \sup_{y_1,y_2\in(\mathcal{S}_{\mathrm{in}})_y}(y_1y_2)^2\\ &= \|\mathbf{k}\|_{\infty,\mathrm{in}}^2 + \|\mathbf{k}\|_{\infty,\mathrm{in}}\cdot Y_{\max}^2 + Y_{\max}^4\\ &\le \left(\|\mathbf{k}\|_{\infty,\mathrm{in}} + Y_{\max}^2\right)^2.\end{aligned}$$

Proof of (132) Since $\mathcal{H}(\mathbf{k}_{\mathrm{RR}})$ is a direct sum, we have

$$\log\mathcal{N}^{\dagger}_{\mathbf{k}_{\mathrm{RR}}}(\mathcal{S}_{\mathrm{in}},\epsilon^\star) \le \log\mathcal{N}^{\dagger}_{\mathbf{k}_1}(\mathcal{S}_{\mathrm{in}},\epsilon^\star/2) + \log\mathcal{N}^{\dagger}_{\mathbf{k}_2}(\mathcal{S}_{\mathrm{in}},\epsilon^\star/2), \tag{134}$$

where $\mathcal{N}^{\dagger}_{\mathbf{k}_1}$ and $\log\mathcal{N}^{\dagger}_{\mathbf{k}_2}$ are the covering numbers of $\mathcal{B}^{\dagger}_{\mathbf{k}_1} \triangleq \{f^2/\|f^2\|_{\mathbf{k}_1} : f\in\mathcal{H}(\mathbf{k})\}$ and $\mathcal{B}^{\dagger}_{\mathbf{k}_2} \triangleq \{f\otimes\langle\cdot,y\rangle_{\mathbb{R}}/\|f\otimes\langle\cdot,y\rangle_{\mathbb{R}}\|_{\mathbf{k}_2} : f\otimes\langle\cdot,y\rangle_{\mathbb{R}}\in\mathcal{H}(\mathbf{k}_2)\}$, respectively.

Note that

$$\begin{aligned}\log\mathcal{N}^{\dagger}_{\mathbf{k}_1}(\mathcal{S}_{\mathrm{in}},\epsilon^\star) &\le 2\log\mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}},\epsilon^\star/(2\|\mathbf{k}\|_{\infty}^{1/2}))\\ &\le 2\log\mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}},(1+\tfrac{Y_{\max}}{\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}})\tfrac{\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}+Y_{\max}}{2n_{\mathrm{out}}})\\ &\le 2\log\mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}},\tfrac{\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}+Y_{\max}}{2n_{\mathrm{out}}}).\end{aligned}$$

Define a kernel on $\mathbb{R}$ by $\mathbf{k}_{\mathbb{R}}(y_1,y_2) \triangleq y_1y_2$. When $\sup_{y\in(\mathcal{S}_{\mathrm{in}})_y}|y| \le Y_{\max}$, we have

$$\mathcal{N}_{\mathbf{k}_{\mathbb{R}}}([-Y_{\max},Y_{\max}],\epsilon) = \mathcal{O}(Y_{\max}^2/\epsilon) \quad \text{for} \quad \epsilon>0.$$

Similarly, note that

$$\log\mathcal{N}^{\dagger}_{\mathbf{k}_2}(\mathcal{S}_{\mathrm{in}},\epsilon^\star) \le \log\mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}},\epsilon^\star/(\|\mathbf{k}\|_{\infty}^{1/2}+\|\mathbf{k}_{\mathbb{R}}\|_{\infty}^{1/2})) + \log\mathcal{N}_{\mathbf{k}_{\mathbb{R}}}(\mathcal{S}_{\mathrm{in}},\epsilon^\star/(\|\mathbf{k}\|_{\infty,\mathrm{in}}^{1/2}+\|\mathbf{k}_{\mathbb{R}}\|_{\infty}^{1/2}))$$

$$\lesssim \log \mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \tfrac{\|\mathbf{k}\|_{\infty,\text{in}}^{1/2}+Y_{\max}}{n_{\text{out}}}) + \log\left(\tfrac{Y_{\max}^2(\|\mathbf{k}\|_{\infty}^{1/2}+Y_{\max})}{n_{\text{out}}}\right)$$

Substituting the above two log-covering number expressions into (134) yields

$$\begin{aligned}\log \mathcal{N}_{\mathbf{k}_{\text{ALG}}}(\mathcal{S}_{\text{in}}, \epsilon^\star) &\lesssim 3 \log \mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \tfrac{\|\mathbf{k}\|_{\infty,\text{in}}^{1/2}+Y_{\max}}{n_{\text{out}}}) + \log\left(\tfrac{Y_{\max}^2(\|\mathbf{k}\|_{\infty,\text{in}}^{1/2}+Y_{\max})}{n_{\text{out}}}\right). \\ &\leq c \cdot \log \mathcal{N}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \tfrac{\|\mathbf{k}\|_{\infty,\text{in}}^{1/2}+Y_{\max}}{n_{\text{out}}})\end{aligned}$$

for some universal positive constant c.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction give clear outlines on our contributions, and we present our contributions in the main text accordingly. We have also included pointers in the introduction that would link to the referred main text containing specific contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Sec. 6, we discuss limitations as well as future work to address these limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have clarified the assumptions and models required for all the theorems and corollaries provided in the main text and appendix. Also we provide a complete proof in the appendix for all the stated results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe details of the experiments in Sec. 5 and provide links to all code and data.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a link to our GitHub repository containing all code in Sec. 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide train-test splits and hyperparameters in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In all figures, we plot error bars representing standard deviation across 100 trials. In all tables, we report mean +/- standard error across 100 trials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate the computer resources for running all experiments in Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and our paper conforms with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work reduces the computational costs of classical methods and is applied to standard datasets. Thus, it has no outsize societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release models or data as part of this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We include URL and licenses for baseline code and datasets used in Sec. 5.2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code with documentation at `https://github.com/ag2435/npr` under a BSD-3 Clause license.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not have any studies or results regarding crowdsourcing experiments and human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not have any studies or results including study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.