Two-stage sampled learning theory on distributions

Zoltán Szabó[1] **Arthur Gretton**[1] **Barnabás Póczos**[2] **Bharath Sriperumbudur**[3]
[1]Gatsby Unit, UCL [2]Machine Learning Department, CMU [3]Department of Statistics, PSU

Abstract

We focus on the distribution regression problem: regressing to a real-valued response from a probability distribution. Although there exist a large number of similarity measures between distributions, very little is known about their generalization performance in specific *learning tasks*. Learning problems formulated on distributions have an inherent two-stage sampled difficulty: in practice only samples from sampled distributions are observable, and one has to build an estimate on similarities computed between sets of points. To the best of our knowledge, the only existing method with consistency guarantees for distribution regression requires kernel density estimation as an intermediate step (which suffers from slow convergence issues in high dimensions), and the domain of the distributions to be compact Euclidean. In this paper, we provide theoretical guarantees for a remarkably simple algorithmic alternative to solve the distribution regression problem: embed the distributions to a reproducing kernel Hilbert space, and learn a ridge regressor from the embeddings to the outputs. Our main contribution is to prove the consistency of this technique in the two-stage sampled setting under mild conditions (on separable, topological domains endowed with kernels). For a given total number of observations, we derive convergence rates as an explicit function of the problem difficulty. As a special case, we answer a 15-year-old open question: we establish the consistency of the classical set kernel [Haussler, 1999; Gärtner et. al, 2002] in regression, and cover more recent kernels on distributions, including those due to [Christmann and Steinwart, 2010].

1 INTRODUCTION

We address the learning problem of *distribution regression* in the two-stage sampled setting [1]: we regress from probability measures to real-valued responses, where we only have bags of samples from the probability distributions. Many classical problems in machine learning and statistics can be analysed in this framework. On the machine learning side, multiple instance learning [2, 3, 4] can be thought of in this way, in the case where each instance in a labeled bag is an i.i.d. (independent identically distributed) sample from a distribution. On the statistical side, tasks might include point estimation of statistics on a distribution (e.g., its entropy or a hyperparameter), where a supervised learning method can help in parameter estimation problems without closed form analytical expressions, or if simulation-based results are computationally expensive.

Before reviewing the existing techniques in the literature, let us start with a somewhat informal definition of the distribution regression problem, and an intuitive phrasing of our goal. Let us suppose that our data consist of $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l}$, where x_i is a probability distribution, $y_i \in \mathbb{R}$, and each (x_i, y_i) pair is i.i.d. sampled from a meta distribution $\mathcal{M}$. However, we do not observe x_i directly; rather, we observe a sample $x_{i,1}, \ldots, x_{i,N_i} \overset{i.i.d.}{\sim} x_i$. Thus the observed data are $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^{N_i}, y_i)\}_{i=1}^{l}$. Our goal is to predict a new y_{l+1} from a new batch of samples $x_{l+1,1}, \ldots, x_{l+1,N_{l+1}}$ drawn from a new distribution x_{l+1}. For example, in a medical application the i^{th} patient might be identified with a probability distribution (x_i), which can be periodicly accessed, measured by blood tests $(\{x_{i,n}\}_{n=1}^{N_i})$. We are also given some health indicator of the patient (y_i), which might be inferred from his/her blood measurements. Based on the observations $(\hat{\mathbf{z}})$, one might try to learn the mapping from the set of blood tests to the health indicator; and the hope is that by observing more patients (larger l) and performing a larger number of tests (larger N_i) the estimated mapping $(\hat{f} = \hat{f}(\hat{\mathbf{z}}))$ becomes more "precise".

The performance of the estimated mapping $(\hat{f})$ depends on the assumed function class $(\mathcal{H})$, the family of $\hat{f}$ candidates. Let $f_{\mathcal{H}}$ denote the best estimator from $\mathcal{H}$ given infinite training samples ($l = \infty$, $N_i = \infty$), and let $\mathcal{E}[f_{\mathcal{H}}]$ be its prediction error. Our goal is to

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. The ordering of the second through fourth authors is alphabetical.

obtain upper bounds for the $0 \leq \mathcal{E}[\hat{f}] - \mathcal{E}[f_{\mathcal{H}}]$ quantity which hold with high probability. More precisely, we are aiming at

1. deriving upper bounds on the excess risk, proving consistency: We construct $\mathcal{E}[\hat{f}] - \mathcal{E}[f_{\mathcal{H}}] \leq r(l, N, \lambda)$ bounds, where λ is a regularization parameter converging to zero as we see more samples ($l \to \infty$, $N = N_i \to \infty$), and choose the (l, N, λ) triplet appropriately to drive $r(l, N, \lambda)$ and hence $\mathcal{E}[\hat{f}] - \mathcal{E}[f_{\mathcal{H}}]$ to 0.
2. obtaining convergence rates: We establish convergence rates for a general prior family $\mathcal{P}(b, c)$ [5], where b captures the effective input dimension, and larger c means smoother $f_{\mathcal{H}}$. In particular, when $l = N^a$ ($a > 0$), the effective dimension is small (large b), and the total number of samples processed $t = lN = N^{a+1}$ is fixed, one obtains a rate of $1/t^{2/7}$ for a smooth regression function ($c = 2$), $1/t^{1/5}$ in the non-smooth case ($c = 1$).

The motivation for considering the $\mathcal{P}(b, c)$ family is two-fold:

1. it does not assume parametric distributions, still certain complexity terms can be explicitly upper bounded in the family. This property will be exploited in our analysis.
2. (for special input distributions) parameter b can be related to the spectral decay of Gaussian Gram matrices, thus available analysis techniques [6] might give alternative prior characterizations.

Briefly, we focus on the following question:

> Can the distribution regression problem be solved consistently under mild conditions?

Despite the large number of available "solutions" and applications of distribution regression dating back to 1999 [7], surprisingly this pretty fundamental question has hardly been touched. In our paper we give affirmative answer to the question by presenting the analysis of a *simple* kernel ridge regression approach [see Eq. (3)] in the two-stage sampled ($\mathcal{M} \to \mathbf{z} \to \hat{\mathbf{z}}$) setting.

Review of approaches to learning on distributions: A number of methods have been proposed over the years to compute the similarity of distributions or bags of samples. As a first approach, one could fit a parametric model to the bags, and estimate the similarity of the bags based on the obtained parameters. It is then possible to define learning algorithms on the basis of these similarities, which often take analytical form. Typical examples with explicit formulas include Gaussians, finite mixtures of Gaussians, and distributions from the exponential family (with known log-normalizer function and zero carrier measure) [8, 9, 10, 11]. A major limitation of these methods, however, is that they apply quite simple parametric assumptions, which may not be sufficient or verifiable in practise.

A heuristic related to the parametric approach is to assume that the training distributions are Gaussians in a reproducing kernel Hilbert space; see for example [10, 12] and references therein. This assumption is algorithmically appealing, as many divergence measures for Gaussians can be computed in closed form using only inner products, making them straightforward to kernelize. A fundamental shortfall of kernelized Gaussian divergences is the lack of their consistency analysis in specific learning algorithms.

A more theoretically grounded approach to learning on distributions has been to define positive definite kernels [13] on the basis of statistical divergence measures on distributions, or by metrics on non-negative numbers; these can then be used in kernel algorithms. This category includes work on semigroup kernels [14], nonextensive information theoretical kernel constructions [15], and kernels based on Hilbertian metrics [16]. For example, in [14] the intuition is as follows: if two measures or sets of points overlap, then their sum is expected to be more concentrated. The value of dispersion can be measured by entropy or inverse generalized variance. In the second type of approach [16], homogeneous Hilbert metrics on the non-negative real line are used to define the similarity of probability distributions. While these techniques guarantee to provide valid kernels on certain restricted domains of measures, the performance of learning algorithms based on finite sample estimates of these kernels remains a challenging open question. One might also plug into learning algorithms (based on similarities of distributions) consistent Rényi and Tsallis divergence estimates [17, 18], but these similarity indices are *not* kernels, and their consistency in specific learning tasks, similarly to the previous works, is open.

To the best of our knowledge, the only prior work addressing the consistency of regression on distributions requires kernel density estimation [1, 19], assumes that the response variable is scalar-valued[1], and the covariates are nonparametric continuous distributions on $\mathbb{R}^d$. As in our setting, the exact forms of these distributions are unknown; they are available only through finite sample sets. Póczos et al. estimated these distributions through a kernel density estimator (assuming these distributions to have a density) and then constructed a kernel regressor that acts on these kernel

[1] [20] considers the case where the responses are also distributions.

density estimates.[2] Using the classical bias-variance decomposition analysis for kernel regressors, they show the consistency of the constructed kernel regressor, and provide a polynomial upper bound on the rates, assuming the true regressor to be Hölder continuous, and the meta distribution that generates the covariates x_i to have finite doubling dimension [22].[3]

An alternative paradigm in learning when the inputs are "bags of objects" is to simply treat each input as a finite set, and to define kernel learning algorithms based on set kernels [23] (also called multi-instance kernels or ensemble kernels, and instances of convolution kernels [7]). In this case, the similarity of two sets is measured by the average pairwise point similarities between the sets. From a *theoretical* perspective, very little has been done to establish the consistency of set kernels in learning since their introduction in 1999 [7, 23]: i.e. in what sense (and with what rates) is the learning algorithm consistent, when the number of items per bag, and the number of bags, is allowed to increase?

It is possible, however, to view set kernels in a distribution setting, as they represent valid kernels between (mean) embeddings of empirical probability measures into a reproducing kernel Hilbert space (RKHS) [24]. The *population limits are well-defined* as being dot products between the embeddings of the generating distributions [25], and for characteristic kernels the distance between embeddings defines a *metric* on probability measures [26, 27]. When bounded kernels are used, mean embeddings *exist for all probability measures* [28]. When we consider the distribution regression setting, however, there is no reason to limit ourselves to set kernels. Embeddings of probability measures to RKHS are used by [29] in defining a yet larger class of easily computable kernels on distributions, via operations performed on the embeddings and their distances. Note that the relation between set kernels and kernels on distributions has been applied by [30] for classification on distribution-valued inputs, however consistency was not studied in that work.

Our **contribution** in this paper is to establish the consistency of an algorithmically simple, mean embedding based ridge regression method (described in Section 2) for the distribution regression problem. This result applies both to the basic set kernels of [7, 23], the distribution kernels of [29], and additional related kernels proposed herein. We provide two-stage sampled excess error bounds, consistency proof and convergence rates in Section 4, and break down the various tradeoffs arising in different sample size and problem difficulties. The principal challenge in proving theoretical guarantees arises from the two-stage sampled nature of the inputs. In our analysis, we make use of [5], who provide error bounds for the one-stage sample setup. These results will make our analysis somewhat shorter (but still rather challenging) by giving upper bounds for some of the upcoming objective terms. Even the verification of these conditions requires care (Section 3) since the inputs in the ridge regression are themselves distribution embeddings (i.e., functions in a reproducing kernel Hilbert space).

Due to the differences in the assumptions made and the loss function used, a direct comparison of our theoretical result and that of [1][3] remains an open question, however we make two observations. First, our approach is more general, since we may regress from any probability measure defined on a separable, topological domain endowed with a kernel. Póczos et al.'s work is restricted to compact domains of finite dimensional Euclidean spaces, and requires the distributions to admit probability densities; distributions on strings, time series, graphs, and other structured objects are disallowed. Second, density estimates in high dimensional spaces suffer from slow convergence rates [31, Section 6.5]. Our approach avoids this problem, as it works directly on distribution embeddings, and does not make use of density estimation as an intermediate step.

2 THE DISTRIBUTION REGRESSION PROBLEM

In this section, we define the distribution regression problem, for a general RKHS on distributions. In Section 3, we will provide examples of valid kernels for this RKHS, including set kernels [7, 23], the kernels from [29], and further related kernels. Below, we first introduce some notation and then formally discuss the distribution regression problem.

Notation: Let $(\mathcal{X}, \tau)$ be a topological space and let $\mathcal{B}(\mathcal{X}) := \mathcal{B}(\tau)$ be the Borel σ-algebra induced by the topology τ. $\mathcal{M}_1^+(\mathcal{X})$ denotes the set of Borel probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. The weak topology ($\tau_w = \tau_w(\mathcal{X}, \tau)$) on $\mathcal{M}_1^+(\mathcal{X})$ is defined as the weakest topology such that the L_h : $(\mathcal{M}_1^+(\mathcal{X}), \tau_w) \to \mathbb{R}$, $L_h(x) = \int_{\mathcal{X}} h(u)\mathrm{d}x(u)$ mapping is continuous for all $h \in C_b(\mathcal{X}) = \{(\mathcal{X}, \tau) \to \mathbb{R}$ bounded, continuous functions$\}$. Let $H = H(k)$ be the RKHS [6] with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing

[2] We would like to clarify that the kernels used in their work are classical smoothing kernels (extensively studied in non-parametric statistics [21]) and not the reproducing kernels that appear throughout our paper.

[3] Using a random kitchen sinks approach, with orthonormal basis projection estimators and RBF kernels [19] proposes a distribution regression algorithm that can computationally handle large scale datasets; as with [1], this approach is based on density estimation in $\mathbb{R}^d$.

kernel. Denote by

$$X = \mu\left(\mathcal{M}_1^+(\mathcal{X})\right) = \{\mu_x : x \in \mathcal{M}_1^+(\mathcal{X})\} \subseteq H$$

the set of $\mu_x = \int_{\mathcal{X}} k(\cdot,u)\mathrm{d}x(u) = \mathbb{E}_{u\sim x}[k(\cdot,u)] \in H$ mean embeddings [24] of the distributions to the space H, and let $Y = \mathbb{R}$. Intuitively, μ_x is the canonical feature map $[k(\cdot,u)]$ averaged according to the probability measure $[\mathrm{d}x(u)]$. Let $\mathcal{H} = \mathcal{H}(K)$ be the RKHS of functions with $K : X \times X \to \mathbb{R}$ as the reproducing kernel. $\mathcal{L}(\mathcal{H})$ is the space of $\mathcal{H} \to \mathcal{H}$ bounded linear operators, and δ_{μ_a} denotes the evaluation functional at μ_a ($a \in \mathcal{M}_1^+(\mathcal{X})$). For $M \in \mathcal{L}(\mathcal{H})$ the operator norm is defined as $\|M\|_{\mathcal{L}(\mathcal{H})} = \sup_{0\neq q\in\mathcal{H}} \frac{\|Mq\|_{\mathcal{H}}}{\|q\|_{\mathcal{H}}}$. Given $(U_1, \mathcal{S}_1)$ and $(U_2, \mathcal{S}_2)$ measurable spaces the $\mathcal{S}_1 \otimes \mathcal{S}_2$ product σ-algebra [6, page 480] on the product space $U_1 \times U_2$ is the σ-algebra generated by the cylinder sets $U_1 \times S_2$, $S_1 \times U_2$ ($S_1 \in \mathcal{S}_1$, $S_2 \in \mathcal{S}_2$). $\mathbb{E}[\cdot]$ denotes expectation.

Distribution regression: In the *distribution regression* problem, we are given samples $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^{N_i}, y_i)\}_{i=1}^l$ with $x_{i,1},\ldots,x_{i,N_i} \overset{i.i.d.}{\sim} x_i$ where $\mathbf{z} = \{(x_i,y_i)\}_{i=1}^l$ with $x_i \in \mathcal{M}_1^+(\mathcal{X})$ and $y_i \in Y$ drawn i.i.d. from a joint meta distribution $\mathcal{M}$ defined on the measurable space $(\mathcal{M}_1^+(\mathcal{X}) \times \mathbb{R}, \mathcal{B}(\tau_w) \otimes \mathcal{B}(\mathbb{R}))$. Unlike in classical supervised learning problems, the problem at hand involves two levels of randomness, wherein first $\mathbf{z}$ is drawn from $\mathcal{M}$ and then $\hat{\mathbf{z}}$ is generated by sampling points from x_i for all $i = 1,\ldots,l$. The **goal** is to learn the relation between the random distribution x and scalar response y based on the observed $\hat{\mathbf{z}}$. For notational simplicity, we will assume that $N = N_i$ ($\forall i$).

As in the classical regression task ($\mathbb{R}^d \to \mathbb{R}$), distribution regression can be tackled as a kernel ridge regression problem (using squared loss as the discrepancy criterion). The kernel (say $k_{\mathcal{G}}$) is defined on $\mathcal{M}_1^+(\mathcal{X})$, and the regressor is then modelled by an element in the RKHS $\mathcal{G} = \mathcal{G}(k_{\mathcal{G}})$ of functions mapping from $\mathcal{M}_1^+(\mathcal{X})$ to $\mathbb{R}$. In this paper, we choose $k_{\mathcal{G}}(x,x') = K(\mu_x, \mu_{x'})$ where $x, x' \in \mathcal{M}_1^+(\mathcal{X})$ and so that the function (in $\mathcal{G}$) to describe the (x,y) random relation is constructed as a composition

$$\mathcal{M}_1^+(\mathcal{X}) \xrightarrow{\mu} X(\subseteq H = H(k)) \xrightarrow{f\in\mathcal{H}=\mathcal{H}(K)} \mathbb{R}.$$

In other words, the distribution $x \in \mathcal{M}_1^+(\mathcal{X})$ is first mapped to $X \subseteq H$ by the mean embedding μ, and the result is composed with f, an element of the RKHS $\mathcal{H} = \mathcal{H}(K)$. Assuming that $f_{\mathcal{G}}$, the minimizer of the expected risk ($\mathcal{E}$) over $\mathcal{G}$ exists, then a function $f_{\mathcal{H}}$ also exists, and satisfies

$$\begin{aligned}\mathcal{E}\left[f_{\mathcal{H}}\right] &= \inf_{f\in\mathcal{H}} \mathcal{E}[f] = \inf_{f\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{M}}[f(\mu_x) - y]^2 \\ &= \inf_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim\mathcal{M}}[g(x) - y]^2 = \inf_{g\in\mathcal{G}} \mathcal{E}[g] = \mathcal{E}\left[g_{\mathcal{G}}\right].\end{aligned}$$

The classical regularization approach is to optimize

$$f_{\mathbf{z}}^{\lambda} = \arg\min_{f\in\mathcal{H}} \frac{1}{l}\sum_{i=1}^{l}\left[f(\mu_{x_i}) - y_i\right]^2 + \lambda\|f\|_{\mathcal{H}}^2 \quad (1)$$

instead of $\mathcal{E}$, based on samples $\mathbf{z}$. Since $\mathbf{z}$ is not accessible, we consider the objective function defined by the observable quantity $\hat{\mathbf{z}}$,

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f\in\mathcal{H}} \frac{1}{l}\sum_{i=1}^{l}\left[f(\mu_{\hat{x}_i}) - y_i\right]^2 + \lambda\|f\|_{\mathcal{H}}^2, \quad (2)$$

where $\hat{x}_i = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_{i,n}}$ is the empirical distribution determined by $\{x_{i,n}\}_{i=1}^{N}$. Algorithmically, ridge regression is quite simple [32]: given training samples $\hat{\mathbf{z}}$, the prediction for a new t test distribution is

$$\boxed{\begin{aligned}(f_{\hat{\mathbf{z}}}^{\lambda}\circ\mu)(t) &= [y_1,\ldots,y_l](\mathbf{K} + l\lambda\mathbf{I}_l)^{-1}\mathbf{k} \in \mathbb{R},\\ \mathbf{K} &= [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{l\times l},\\ \mathbf{k} &= [K(\mu_{\hat{x}_1}, \mu_t);\ldots;K(\mu_{\hat{x}_l}, \mu_t)] \in \mathbb{R}^l.\end{aligned}} \quad (3)$$

Remarks:

1. It is important to note that the algorithm has access to the sample points *only via* their *mean embeddings* $\{\mu_{\hat{x}_i}\}_{i=1}^l$ in Eq. (2).
2. There is a *two-stage sampling difficulty* to tackle: The transition from $f_{\mathcal{H}}$ to $f_{\mathbf{z}}^{\lambda}$ represents the fact that we have only l distribution samples ($\mathbf{z}$); the transition from $f_{\mathbf{z}}^{\lambda}$ to $f_{\hat{\mathbf{z}}}^{\lambda}$ means that the x_i distributions can be accessed only via samples ($\hat{\mathbf{z}}$).
3. While ridge regression can be performed using the kernel $k_{\mathcal{G}}$, the two-stage sampling makes it difficult to work with arbitrary $k_{\mathcal{G}}$. By contrast, our choice of $k_{\mathcal{G}}(x,x') = K(\mu_x,\mu_{x'})$ enables us to handle the two-stage sampling by estimating μ_x with an empirical estimator and using it in the algorithm as shown above.

The main **goal** of this paper is to analyse the excess risk $\mathcal{E}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}]$, i.e., the regression performance compared to the best possible estimation from $\mathcal{H}$, and to establish consistency and rates of convergence as a function of the (l, N, λ) triplet, and of the difficulty of the problem in the sense of [5].

3 ASSUMPTIONS

In this section we detail our assumptions on the $(\mathcal{X}, k, K)$ triplet, and show that regressing with set kernels fit into the studied problem family. Our analysis will rely on existing ridge regression results [5] which focus on problem (1), where only a single-stage sampling is present; hence we have to verify the associated conditions. Though we make use of these results, the analysis still remains rather challenging; the available bounds can moderately shorten our proof. We must also take particular care in verifying that [5]'s conditions are met, since they must hold for the space of *mean embeddings of the distributions* ($X = \mu\left(\mathcal{M}_1^+(\mathcal{X})\right)$), whose properties as a function of $\mathcal{X}$ and H must themselves be established. Our assumptions:

- $\exists f_{\mathcal{H}}$ such that $\mathcal{E}[f_{\mathcal{H}}] = \inf_{f\in\mathcal{H}} \mathcal{E}(f)$.
- $(\mathcal{X}, \tau)$ is a separable, topological domain.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded ($\exists B_k < \infty$ such that $\sup_{u\in\mathcal{X}} k(u,u) \le B_k$) and continuous.
- $K : X \times X \to \mathbb{R}$ is bounded, i.e., $\exists B_K < \infty$ such that

$$K(\mu_a, \mu_a) \le B_K, \quad (\forall \mu_a \in X), \tag{4}$$

and $\Psi(\mu_c) := K(\cdot, \mu_c) : X \to \mathcal{H}$ is Hölder continuous, i.e., $\exists L > 0$, $h \in (0,1]$ such that for $\forall(\mu_a, \mu_b) \in X \times X$

$$\|\Psi(\mu_a) - \Psi(\mu_b)\|_{\mathcal{H}} \le L \|\mu_a - \mu_b\|_H^h. \tag{5}$$

- y is bounded: $\exists C < \infty$ such that $|y| \le C$ almost surely.
- $X = \mu(\mathcal{M}_1^+(\mathcal{X})) \in \mathcal{B}(H)$.

Discussion of the assumptions: We give a short insight into the consequences of our assumptions and present some concrete examples.

1. The boundedness and continuity of k imply the measurability of $\mu : (\mathcal{M}_1^+(\mathcal{X}), \mathcal{B}(\tau_w)) \to (H, \mathcal{B}(H))$, which using the $X \in \mathcal{B}(H)$ condition guarantees that the ρ, the measure induced by $\mathcal{M}$ on $X \times \mathbb{R}$ is well-defined (see the supplementary material).
2. For a linear kernel, $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H$, ($\mu_a, \mu_b \in X$), one can verify (see the supplementary material) that Hölder continuity holds with $L = 1$, $h = 1$. Also, since $K(\mu_a, \mu_b) \le B_k$ for any $a, b \in \mathcal{M}_1^+(\mathcal{X})$, we can choose $B_K = B_k$. Evaluating the kernel, K at the $\mu_{\hat{x}_i} = \int_{\mathcal{X}} k(\cdot, u)\mathrm{d}\hat{x}_i(u) = \frac{1}{N}\sum_{n=1}^N k(\cdot, x_{i,n})$ empirical embeddings yields the standard set kernel:

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \frac{1}{N^2}\sum_{n,m=1}^N k(x_{i,n}, x_{j,m}).$$

3. One can also prove (see the supplement) by using the properties of negative/positive definite functions [33] that many K functions on $X \times X$ are kernels and (in case of compact metric $\mathcal{X}$ domains) Hölder continuous.[4] Some examples are listed in Table 1; these kernels are the natural extensions to distributions of the Gaussian [29], exponential, Cauchy, generalized t-student and inverse multiquadratic kernels.

[4]To guarantee the Hölder property of K-s, we assume the continuity of μ. For example, if $\mathcal{X}$ is a compact metric space and k is universal, then μ metrizes the weak topology τ_w [34, Theorem 23, page 1552], hence μ is continuous. In this case $X = \mu(\mathcal{M}_1^+(\mathcal{X}))$ is compact metric (see the supplement), thus closed and hence $X \in \mathcal{B}(H)$ also holds.

4. $Y = \mathbb{R}$ is a separable Hilbert space hence Polish, and thus the $\rho(y|\mu_a)$ conditional distribution ($y \in \mathbb{R}$, $\mu_a \in X$) is well-defined; see [6, Lemma A.3.16, page 487].
5. The separability of $\mathcal{X}$ and the continuity of k implies the separability of H [6, Lemma 4.33, page 130]. Also, since $X \subseteq H$, X is separable; hence so is $\mathcal{H}$ due to the continuity of K.

Verification of [5]'s conditions: Below we prove that [5]'s conditions hold under our assumptions.

1. $Y = \mathbb{R}$ and $\mathcal{H}$ are separable Hilbert spaces – as we have seen.
2. By the bilinearity of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and the reproducing property of K, the measurability of $(\mu_x, \mu_t) \mapsto \langle K(\cdot,\mu_x)w, K(\cdot,\mu_t)v\rangle_{\mathcal{H}} = wK(\mu_x,\mu_t)v \quad (\forall w, v \in \mathbb{R})$ is equivalent to that of $(\mu_x, \mu_t) \mapsto K(\mu_x, \mu_t)$; the latter follows from the Hölder continuity of Ψ (see the supplement).
3. Due to the boundedness of y, we have $\int_{X\times\mathbb{R}} y^2 \mathrm{d}\rho(\mu_x, y) \le \int_{X\times\mathbb{R}} C^2 \mathrm{d}\rho(\mu_x, y) \le C^2 < \infty$, and $\exists\Sigma > 0, \exists M > 0$ such that

$$\int_{\mathbb{R}} e^{\frac{|y - f_{\mathcal{H}}(\mu_x)|}{M}} - \frac{|y - f_{\mathcal{H}}(\mu_x)|}{M} - 1 \mathrm{d}\rho(y|\mu_x) \le \frac{\Sigma^2}{2M^2} \tag{6}$$

for ρ_X-almost $\mu_x \in X$, where $\rho(\mu_x, y) = \rho(y|\mu_x)\rho_X(\mu_x)$ is factorized into conditional and marginal distributions. (6) is a model of the noise of the output y; it is satisfied, for example in case of bounded noise [5, page 9]. By the boundedness of y and that of kernel K this property holds: $|y - f_{\mathcal{H}}(\mu_x)| \le |y| + |f_{\mathcal{H}}(\mu_x)| \le C + \|f_{\mathcal{H}}\|_H \sqrt{B_K}$, where we used the triangle inequality and Lemma 4.23 (page 124) from [6].

4 ERROR BOUNDS, CONSISTENCY, CONVERGENCE RATE

In this section, we present our main result: we derive high probability upper bound for the excess risk $\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right]$ of the mean embedding based ridge regression (MERR) method, see our main theorem. We also illustrate the upper bound for particular classes of prior distributions, resulting in sufficient conditions for convergence and concrete convergence rates (see Consequences 1-2). We first give a high-level sketch of our convergence analysis and the results are stated with their intuitive interpretation. Then an outline of the main proof ideas follows; technical details of the proof steps may be found in the supplement.

At a high level, our convergence analysis takes the following form: Having explicit expressions for $f_{\mathbf{z}}^{\lambda}$, $f_{\hat{\mathbf{z}}}^{\lambda}$ [see Eq. (9)-(10)], we will decompose the excess risk

Table 1: Nonlinear kernels on mean embedded distributions: $K = K(\mu_a, \mu_b)$; $\theta > 0$. For the Hölder continuity, we assume that $\mathcal{X}$ is a compact metric space and μ is continuous (the latter is implied e.g., by a universal k).

K_G	K_e	K_C	K_t	K_i
$e^{-\frac{\|\mu_a-\mu_b\|_H^2}{2\theta^2}}$	$e^{-\frac{\|\mu_a-\mu_b\|_H}{2\theta^2}}$	$\left(1+\|\mu_a-\mu_b\|_H^2/\theta^2\right)^{-1}$	$\left(1+\|\mu_a-\mu_b\|_H^\theta\right)^{-1}$	$\left(\|\mu_a-\mu_b\|_H^2+\theta^2\right)^{-\frac{1}{2}}$
$h=1$	$h=\frac{1}{2}$	$h=1$	$h=\frac{\theta}{2}$ ($\theta \le 2$)	$h=1$

$\mathcal{E}[f_{\hat{\mathbf{z}}}^\lambda] - \mathcal{E}[f_{\mathcal{H}}]$ into five terms:

$$\begin{aligned}
\mathcal{E}\left[f_{\hat{\mathbf{z}}}^\lambda\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] &\le 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda) + S_1 + S_2\right],\\
S_{-1} &= \|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}})\|_{\mathcal{H}}^2,\\
S_0 &= \|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}(T_{\mathbf{x}} - T_{\hat{\mathbf{x}}})f_{\mathbf{z}}^\lambda\|_{\mathcal{H}}^2,\\
S_1 &= \|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}}f_{\mathcal{H}})\|_{\mathcal{H}}^2,\\
S_2 &= \|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}(T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}}^2,\\
\mathcal{A}(\lambda) &= \|\sqrt{T}(f^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}}^2,
\end{aligned}$$

where $f^\lambda = \arg\min_{f\in\mathcal{H}} \mathcal{E}[f] + \lambda\|f\|_{\mathcal{H}}^2$, $T_{\mu_a} = K(\cdot,\mu_a)\delta_{\mu_a}$ $[T_{\mu_a}(f) = K(\cdot,\mu_a)f(\mu_a),\ \mu_a \in X]$,

$$T = \int_X T_{\mu_a}\mathrm{d}\rho_X(\mu_a) \in \mathcal{L}(\mathcal{H}), \quad T_{\mu_a} \in \mathcal{L}(\mathcal{H}). \tag{7}$$

1. Three of the terms (S_1, S_2, $\mathcal{A}(\lambda)$) will be identical to terms in [5], hence their bounds can be applied.
2. The two new terms (S_{-1}, S_0), the result of two-stage sampling, will be upper bounded by making use of the convergence of the empirical mean embeddings, and the Hölder property of K.

These bounds will lead to the following results:

Main theorem (bound on the excess risk). *Let M, Σ and T be as in (6), (7). Let $\Psi(\mu_a) = K(\cdot,\mu_a) : X \to \mathcal{H}$ be Hölder continuous with constants L, h. Let $l \in \mathbb{N}$, $N \in \mathbb{N}$, $\lambda > 0$, $0 < \eta < 1$, $C > 0$, $\delta > 0$, $C_\eta = 32\log^2(6/\eta)$, $|y| \le C$ (a.s.), $\mathcal{A}(\lambda)$ the residual as above, and define $\mathcal{B}(\lambda) = \|f^\lambda - f_{\mathcal{H}}\|_{\mathcal{H}}^2$ the reconstruction error, $\mathcal{N}(\lambda) = Tr[(T+\lambda)^{-1}T]$ the effective dimension. Then with probability at least $1-\eta-e^{-\delta}$*

$$\begin{aligned}
&\mathcal{E}\left[f_{\hat{\mathbf{z}}}^\lambda\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] \le\\
&\le 5\left\{\frac{4L^2C^2\left(1+\sqrt{\log(l)+\delta}\right)^{2h}(2B_k)^h}{\lambda N^h}\left[1+\frac{4(B_K)^2}{\lambda^2}\right] + \right.\\
&\left.\mathcal{A}(\lambda) + C_\eta\left[\frac{B_K^2\mathcal{B}(\lambda)}{l^2\lambda} + \frac{B_K\mathcal{A}(\lambda)}{4l\lambda} + \frac{B_KM^2}{l^2\lambda} + \frac{\Sigma^2\mathcal{N}(\lambda)}{l}\right]\right\}
\end{aligned}$$

provided that $l \ge 2C_\eta B_K\mathcal{N}(\lambda)/\lambda$, $\lambda \le \|T\|_{\mathcal{L}(\mathcal{H})}$, $N \ge (1+\sqrt{\log(l)+\delta})^2 2^{\frac{h+6}{h}} B_k(B_K)^{\frac{1}{h}} L^{\frac{2}{h}}/\lambda^{\frac{2}{h}}$.

Below we specialize our bound on the excess risk for a general prior class, which captures the difficulty of the regression problem as defined in [5]. This $\mathcal{P}(b,c)$ class is described by two parameters b and c: intuitively, larger b means faster decay of the eigenvalues of the covariance operator T [(7)], hence smaller effective input dimension; larger c corresponds to smoother $f_{\mathcal{H}}$. Formally:

Definition of the $\mathcal{P}(b,c)$ class: Let us fix the positive constants M, Σ, R, α, β. Then given $1 < b$, $c \in [1,2]$, the $\mathcal{P}(b,c)$ class is the set of probability distributions ρ on $Z = X \times \mathbb{R}$ such that (i) the (μ_x, y) assumption holds with M, Σ in (6), (ii) there is a $g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{\frac{c-1}{2}}g$ with $\|g\|_{\mathcal{H}}^2 \le R$, (iii) in the $T = \sum_{n=1}^N t_n \langle\cdot, e_n\rangle_{\mathcal{H}} e_n$ spectral theorem based decomposition ($(e_n)_{n=1}^N$ is a basis of $ker(T)^\perp$), $N = +\infty$, and the eigenvalues of T satisfy $\alpha \le n^b t_n \le \beta$ $(\forall n \ge 1)$.

We can provide a simple example of when the source decay conditions hold, in the event that the distributions are normal with means m_i and identical variance ($x_i = N(m_i, \sigma^2 I)$). When Gaussian kernels (k) are used with linear K, then $K(\mu_{x_i}, \mu_{x_j}) = e^{-c\|m_i - m_j\|^2}$ [30, Table 1, line 2] (Gaussian, with arguments equal to the difference in means). Thus, this Gram matrix will correspond to the Gram matrix using a Gaussian kernel between points m_i. The spectral decay of the Gram matrix will correspond to that of the Gaussian kernel, with points drawn from the meta-distribution over the m_i. Thus, the source conditions are analysed in the same manner as for Gaussian Gram matrices, e.g. see [6] for a discussion of the spectral decay properties.

In the $\mathcal{P}(b,c)$ family, the behaviour of $\mathcal{A}(\lambda)$, $\mathcal{B}(\lambda)$ and $\mathcal{N}(\lambda)$ is known; specializing our theorem we get:[5]

Consequence 1 (Excess risk in the $\mathcal{P}(b,c)$ class)**.**

$$\begin{aligned}
&\mathcal{E}\left[f_{\hat{\mathbf{z}}}^\lambda\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] \le 5\left\{\frac{4L^2C^2\left(1+\sqrt{\log(l)+\delta}\right)^{2h}(2B_k)^h}{\lambda N^h}\right.\\
&\times\left[1+\frac{4(B_K)^2}{\lambda^2}\right] + R\lambda^c + C_\eta \times\\
&\left.\left[\frac{B_K^2R\lambda^{c-2}}{l^2} + \frac{B_KR\lambda^{c-1}}{4l} + \frac{B_KM^2}{l^2\lambda} + \frac{\Sigma^2\beta b}{(b-1)l\lambda^{\frac{1}{b}}}\right]\right\}.
\end{aligned}$$

[5] In what follows, we assume the conditions of the main theorem and $\rho \in \mathcal{P}(b,c)$.

By choosing λ appropriately as a function of l and N, the excess risk $\mathcal{E}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}]$ converges to 0, and we can use Consequence 1 to obtain convergence rates: the task reduces to the study of

$$r(l,N,\lambda) = \frac{\log^h(l)}{N^h\lambda^3} + \lambda^c + \frac{1}{l^2\lambda} + \frac{1}{l\lambda^{\frac{1}{b}}} \to 0, \qquad (8)$$

subject to $l \geq \lambda^{-\frac{1}{b}-1}$.[6] By matching two terms in (8), solving for λ and plugging the result back to the bound (see the supplementary material), we obtain:

Consequence 2 (Consistency and convergence rate in $\mathcal{P}(b,c)$)**.** *Let $l = N^a$ $(a>0)$. The excess risk can be upper bounded (constant multipliers are discarded) by the quantities given in the last column of Table 2.*

Note: in function r [Eq. (8)] (i) the first term comes from the error of the mean embedding estimation, (ii) the second term corresponds to $\mathcal{A}(\lambda)$, a complexity measure of $f_{\mathcal{H}}$, (iii) the third term is due to the S_1 bound, (iv) the fourth term expresses $\mathcal{N}(\lambda)$, a complexity index of the hypothesis space $\mathcal{H}$ according to the marginal measure ρ_X. As an example, let us take two rows from Table 2:

1. First row: In this case the first and second terms dominate $r(l,N,\lambda)$ in (8); in other words the error is determined by the mean embedding estimation process and the complexity of $f_{\mathcal{H}}$. Let us assume that b is large in the sense that $1/b \approx 0$, $(b+1)/b \approx 1$ (hence, the effective dimension of the input space is small); and assume that K is Lipschitz ($h=1$). Under these conditions the lower bound for a is approximately $\max(c/(c+3), 1/(c+3)) = c/(c+3) \leq a$ (since $c \geq 1$). Using such an a (i.e., the exponent in $l = N^a$ is not too small), then the convergence rate is $[\log(N)/N]^{\frac{c}{c+3}}$. Thus, for example, if $c = 2$ ($f_{\mathcal{H}} = T^{\frac{c-1}{2}}g$ is smoothed by T from a $g \in \mathcal{H}$), then $a = \frac{2}{2+3} = 0.4$ and the convergence rate is $[\log(N)/N]^{0.4}$; in other words the rate is approximately $1/N^{0.4}$. If c takes its minimal value ($c=1$; $f_{\mathcal{H}}$ is less smooth), then $a = \frac{1}{1+3} = \frac{1}{4}$ results in an approximate rate of $1/N^{0.25}$. Alternatively, if we keep the total number of samples processed $t = lN = N^{a+1}$ fixed, $r(t) \approx 1/N^a = 1/t^{a/(a+1)} = 1/t^{1-1/(a+1)}$, i.e., the convergence rate becomes larger for smoother regression problems (increasing c).

2. Last row: At this extreme, two terms dominate: the complexity of $\mathcal{H}$ according to ρ_X, and a term from the bound on S_1. Under this condition, although one can solve the matching criterion for λ, and it is possible to drive the individual terms of r to zero, l cannot be chosen large enough (within the analysed $l = N^a$ $(a>0)$ scheme) to satisfy the $l \geq \lambda^{-\frac{1}{b}-1}$ constraint; thus convergence fails.

[6] Note that the $N \geq \log(l)/\lambda^{\frac{2}{h}}$ constraint has been discarded; it is implied by the convergence of the first term in r [Eq. (8)] (see the supplementary material).

Proof of main theorem: We present the main steps of the proof of our theorem; detailed derivations can be found in the supplementary material. Let us define $\mathbf{x} = \{x_i\}_{i=1}^l$ and $\hat{\mathbf{x}} = \{\{x_{i,n}\}_{n=1}^N\}_{i=1}^l$ as the 'x-part' of $\mathbf{z}$ and $\hat{\mathbf{z}}$. One can express $f_{\mathbf{z}}^{\lambda}$ [5], and similarly $f_{\hat{\mathbf{z}}}^{\lambda}$ as

$$f_{\mathbf{z}}^{\lambda} = (T_{\mathbf{x}}+\lambda)^{-1}g_{\mathbf{z}}, \qquad T_{\mathbf{x}} = \frac{1}{l}\sum_{i=1}^{l} T_{\mu_{x_i}}, \qquad (9)$$

$$f_{\hat{\mathbf{z}}}^{\lambda} = (T_{\hat{\mathbf{x}}}+\lambda)^{-1}g_{\hat{\mathbf{z}}}, \qquad T_{\hat{\mathbf{x}}} = \frac{1}{l}\sum_{i=1}^{l} T_{\mu_{\hat{x}_i}}, \qquad (10)$$

$$g_{\mathbf{z}} = \frac{1}{l}\sum_{i=1}^{l} K(\cdot,\mu_{x_i})y_i, \quad g_{\hat{\mathbf{z}}} = \frac{1}{l}\sum_{i=1}^{l} K(\cdot,\mu_{\hat{x}_i})y_i. \quad (11)$$

In Eqs. (9), (10), (11), $T_{\mathbf{x}}, T_{\hat{\mathbf{x}}} : \mathcal{H} \to \mathcal{H}$, $g_{\mathbf{z}}, g_{\hat{\mathbf{z}}} \in \mathcal{H}$.

• **Decomposition of the excess risk:** We derive the upper bound for the excess risk

$$\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] \leq 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda) + S_1 + S_2\right]. \quad (12)$$

• **It is sufficient to upper bound S_{-1} and S_0:** [5] has shown that $\forall \eta > 0$ if $l \geq \frac{2C_\eta B_K \mathcal{N}(\lambda)}{\lambda}$ and $\lambda \leq \|T\|_{\mathcal{L}(\mathcal{H})}$, then $\mathbb{P}(\boldsymbol{\Theta}(\lambda,\mathbf{z}) \leq 1/2) \geq 1-\eta/3$, where

$$\boldsymbol{\Theta}(\lambda,\mathbf{z}) = \|(T-T_{\mathbf{x}})(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}$$

and one can obtain upper bounds on S_1 and S_2 which hold with probability $1-\eta$. For $\mathcal{A}(\lambda)$ no probabilistic argument was needed.

• **Probabilistic bounds on $\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2$, $\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$:** By using the $\|Mu\|_{\mathcal{H}} \leq \|M\|_{\mathcal{L}(\mathcal{H})}\|u\|_{\mathcal{H}}$ ($M \in \mathcal{L}(\mathcal{H}), u \in \mathcal{H}$) inequality, we bound S_{-1} and S_0 as

$$S_{-1} \leq \|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2 \|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2,$$
$$S_0 \leq \|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2 \|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2 \|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2.$$

For the terms on the r.h.s., we can derive the upper bounds [for α see Eq. (13)]:

$$\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2 \leq L^2C^2 \frac{(1+\sqrt{\alpha})^{2h}(2B_k)^h}{N^h},$$
$$\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \frac{2}{\sqrt{\lambda}},$$
$$\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2 \leq \frac{(1+\sqrt{\alpha})^{2h}\, 2^{h+2}(B_k)^h B_K L^2}{N^h},$$
$$\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2 \leq \frac{C^2 B_K}{\lambda^2}.$$

The bounds hold under the following conditions:

Table 2: Convergence conditions, convergence rates. Rows from top: $1-2$, $1-3$, $1-4$, $2-3$, $2-4$, $3-4^{th}$ terms are matched in $r(l,N,\lambda)$, the upper bound on the excess risk; see Eq. (8). First column: convergence condition. Second column: conditions for the dominance of the matched terms *while* they also converge to zero. Third column: convergence rate of the excess risk.

Convergence condition	Dominance + convergence condition	Convergence rate
$\max\left(\frac{h}{(c+3)\min(2,b)}, \frac{h(b+1)}{(c+3)b}\right) \le a$	$\max\left(\frac{h\left(\frac{1}{b}+c\right)}{c+3}, \frac{h(b+1)}{(c+3)b}\right) \le a$	$\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}$
$\max\left(\frac{h}{6}, \frac{h}{2(b+1)}, \frac{h(b+1)}{2(2b+1)}\right) \le a < \frac{h}{2}$	$\max\left(\frac{h}{6}, \frac{h(b+1)}{2(2b+1)}\right) \le a < \min\left(\frac{h}{2} - \frac{h}{c+3}, \frac{\frac{h}{2}\left(\frac{1}{b}-1\right)}{\frac{1}{b}-2}\right)$	$\frac{1}{N^{3a-\frac{h}{2}} \log^{\frac{h}{2}}(N)}$
$\max\left(\frac{hb}{7b-2}, \frac{h}{3b}, \frac{h(b+1)}{4b}\right) \le a < h$	$\max\left(\frac{h(b-1)}{4b-2}, \frac{h}{3b}, \frac{h(b+1)}{4b}\right) \le a < \frac{h(bc+1)}{3b+bc}$	$\frac{1}{N^{a+\frac{a-h}{3b-1}} \log^{\frac{h}{3b-1}}(N)}$
$a < \frac{h(c+1)}{6}$, $1 > \frac{2(b+1)}{(c+1)b}$	never	never
$a < \frac{h(bc+1)}{3b}$, $1 > \frac{b+1}{bc+1}$	$a < \frac{h(bc+1)}{3b+bc}$, $1 > \frac{b+1}{bc+1}$	$\frac{1}{N^{\frac{abc}{bc+1}}}$
never	never	never

1. $\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2$: if the empirical mean embeddings are close to their population counterparts, i.e.,

$$\|\mu_{x_i} - \mu_{\hat{x}_i}\|_H \le \frac{(1+\sqrt{\alpha})\sqrt{2B_k}}{\sqrt{N}} \tag{13}$$

for $\forall i = 1,\ldots,l$. This event has probability $1 - le^{-\alpha}$ over all l samples by a union bound.
2. $\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$: (13) is assumed.
3. $\|\sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$: $\frac{(1+\sqrt{\alpha})^2 2^{\frac{h+6}{h}} B_k (B_K)^{\frac{1}{h}} L^{\frac{2}{h}}}{(\lambda)^{\frac{2}{h}}} \le N$, (13), and $\mathbf{\Theta}(\lambda,\mathbf{z}) \le \frac{1}{2}$.
4. $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$: This upper bound always holds (under the model assumptions).

• **Union bound:** By applying an $\alpha = \log(l) + \delta$ reparameterization, and combining the received upper bounds with [5]'s results for S_1 and S_2, the theorem follows with a union bound.

Finally, we note that

• existing results were used at two points to simplify our analysis: bounding S_1, S_2, $\Theta(\lambda,\mathbf{z})$ [5] and $\|\mu_{x_i} - \mu_{\hat{x}_i}\|_H$ [25].
• although the primary focus of our paper is clearly theoretical, we have provided some illustrative experiments in the supplementary material. These include
1. a comparison with the only alternative, theoretically justified distribution regression method [1][3] on supervised entropy learning, where our approach gives better performance,
2. an experiment on aerosol prediction based on satellite images, where we perform as well as recent domain-specific, engineered methods [35] (which themselves beat state-of-the-art multiple instance learning alternatives).

5 CONCLUSION

In this paper we established the learning theory of distribution regression under mild conditions, for probability measures on separable, topological domains endowed with kernels. We analysed an algorithmically simple and parallelizable[7] ridge regression scheme defined on the embeddings of the input distributions to a RKHS. As a special case of our analysis, we proved the consistency of regression for set kernels [7, 23] in the distribution-to-real regression setting (which was a 15-year-old open problem), and for a recent kernel family [29], which we have expanded upon (Table 1). To keep the presentation simple we focused on the quadratic loss ($\mathcal{E}$), bounded kernels (k, K), real-valued labels (Y), and mean embedding (μ) based distribution regression with i.i.d. samples ($\{x_{i,n}\}_{n=1}^N$). In future work, we will relax these assumptions, and also consider deriving bounds with approximation error (capturing the richness of class $\mathcal{H}$ in the bounds).[8] Another exciting open question is whether (i) lower bounds on convergence can be proved, (ii) optimal convergence rates can be derived, (iii) one can obtain error bounds for non-point estimates.

Acknowledgements

This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.

[7] Recently, [36] has constructed theoretically sound parallelization algorithms for kernel ridge regression.

[8] The extension to separable Hilbert output spaces and the misspecicified case with approximation error are already available [37].

References

[1] Barnabás Póczos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. *AISTATS (JMLR W&CP)*, 31:507–515, 2013.

[2] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[3] Soumya Ray and David Page. Multiple instance regression. In *ICML*, pages 425–432, 2001.

[4] Daniel R. Dooly, Qi Zhang, Sally A. Goldman, and Robert A. Amar. Multiple-instance learning of real-valued data. *Journal of Machine Learning Research*, 3:651–678, 2002.

[5] Andrea Caponnetto and Ernesto De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[6] Ingo Steinwart and Andres Christmann. *Support Vector Machines*. Springer, 2008.

[7] David Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999. (`http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf`).

[8] Fei Wang, Tanveer Syeda-Mahmood, Baba C. Vemuri, David Beymer, and Anand Rangarajan. Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration. *Medical Image Computing and Computer-Assisted Intervention*, 12:648–655, 2009.

[9] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *ICML*, pages 361–368, 2003.

[10] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[11] Frank Nielsen and Richard Nock. A closed-form expression for the Sharma-Mittal entropy of exponential families. *Journal of Physics A: Mathematical and Theoretical*, 45:032003, 2012.

[12] Shaohua Kevin Zhou and Rama Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:917–929, 2006.

[13] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[14] Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:11691198, 2005.

[15] André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Nonextensive information theoretical kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

[16] Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pages 136–143, 2005.

[17] Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *UAI*, pages 599–608, 2011.

[18] Barnabás Póczos, Liang Xiong, Dougal Sutherland, and Jeff Schneider. Support distribution machines. Technical report, Carnegie Mellon University, 2012. (`http://arxiv.org/abs/1202.0302`).

[19] Junier B. Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. Fast distribution to real regression. *AISTATS (JMLR W&CP)*, 33:706–714, 2014.

[20] Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. *ICML (JMLR W&CP)*, 28:1049–1057, 2013.

[21] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New-york, 2002.

[22] Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. Technical report, Max Planck Institute for Intelligent Systems, 2011. (`http://arxiv.org/abs/1110.4300`).

[23] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[24] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[25] Yasemin Altun and Alexander Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, pages 139–153, 2006.

[26] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[27] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

[28] Kenji Fukumizu, Francis Bach, and Michael Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[29] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, pages 406–414, 2010.

[30] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, pages 10–18, 2012.

[31] Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

[32] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

[33] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.

[34] Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[35] Zhuang Wang, Liang Lan, and Slobodan Vucetic. Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50:2226–2237, 2012.

[36] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Technical report, University of California, Berkeley, 2014. (`http://arxiv.org/abs/1305.5029`).

[37] Zoltán Szabó, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. Technical report, Gatsby Unit, University College London, 2014. (`http://arxiv.org/abs/1411.2066`).

[38] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics – Functional Analysis*. Academic Press, 1980.

[39] Gabriel Nagy. Real analysis (lecture notes): Chapter III: Measure theory, Section 3: Measurable spaces and measurable maps. Technical report, Kansas State University. (`http://www.math.ksu.edu/~nagy/real-an/3-03-measbl.pdf`).

[40] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

[41] John L. Kelley. *General Topology*. Springer, 1975.

[42] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[43] Zoltán Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.

A SUPPLEMENTARY MATERIAL

This supplementary material contains (i) detailed proofs of the consistency of MERR (Section A.1), (ii) numerical illustrations (Section A.2).

A.1 Proofs

A.1.1 Proof of k: continuous, bounded $\Rightarrow$ μ: H-measurable; μ: H-measurable, $X=\mu(\mathcal{M}_1^+(\mathcal{X}))\in\mathcal{B}(H)$ $\Rightarrow$ μ: X-measurable $\Rightarrow \exists\rho$

Below we give sufficient conditions for the existence of probability measure ρ. We divide the proof into 3 steps:

- **k: continuous, bounded $\Rightarrow$ μ: H-measurable**: The mapping $\mu:(\mathcal{M}_1^+(\mathcal{X}),\mathcal{B}(\tau_w))\rightarrow(H,\mathcal{B}(H))$ is measurable, iff the $L_h:(\mathcal{M}_1^+(\mathcal{X}),\mathcal{B}(\tau_w))\rightarrow(\mathbb{R},\mathcal{B}(\mathbb{R}))$ map defined as $L_h(x)=\langle h,\mu_x\rangle_H\left(=\int_{\mathcal{X}}h(u)\mathrm{d}x(u)\right)$ is measurable for $\forall h\in H$ [38, Theorem IV. 22, page 116]. If k is assumed to be continuous and bounded, these properties also hold for $\forall h\in H$ [6, Lemma 4.23, page 124; Lemma 4.28, page 128], i.e. $H=H(k)\subseteq C_b(\mathcal{X})$. By the definition of the weak topology the L_h functions are continuous (for $\forall h\in H$), which implies the required Borel measurability [6, page 480] of L_h-s (for $\forall h\in H$).
- **μ : H-measurable, $X=\mu(\mathcal{M}_1^+(\mathcal{X})\in\mathcal{B}(H)$ $\Rightarrow$ μ : X-measurable**: Let τ denote the open sets on $H=H(k)$. Let $\tau|_X=\{A\cap X: A\in\tau\}$ be the subspace topology on X, and let $\mathcal{B}(H)|_X=\{A\cap X: A\in\mathcal{B}(H)\}$ be the subspace σ-algebra on X. Since $\mathcal{B}\left(\tau|_X\right)=\mathcal{B}(H)|_X\subseteq\mathcal{B}(H)$ (the containing relation follows from the $X\in\mathcal{B}(H)$ condition), and $\mathcal{B}(H)|_X=\{A\in\mathcal{B}(H):A\subseteq X\}$, the measurability of $\mu:(\mathcal{M}_1^+(\mathcal{X}),\mathcal{B}(\tau_w))\rightarrow(H,\mathcal{B}(H))$ implies the measurability of $\mu:(\mathcal{M}_1^+(\mathcal{X}),\mathcal{B}(\tau_w))\rightarrow(X,\mathcal{B}(H)|_X)$.
- **μ : X-measurable $\Rightarrow\exists\rho$**: Let us consider the

$$g:(\mathcal{M}_1^+(\mathcal{X})\times\mathbb{R},\mathcal{B}(\tau_w)\otimes\mathcal{B}(\mathbb{R}))\rightarrow(X\times\mathbb{R},\mathcal{B}(H)|_X\otimes\mathcal{B}(\mathbb{R}))\tag{14}$$

  $g(x,y)=[g_1(x,y);g_2(x,y)]=[\mu_x;y]$ mapping. If g is a *measurable* function, then it defines ρ, a probability measure on $(X\times\mathbb{R},\mathcal{B}(H)|_X\otimes\mathcal{B}(\mathbb{R}))$ by looking at g as a random variable taking values in $X\times\mathbb{R}$:

$$\rho(C):=\mathcal{M}\left(g^{-1}(C)\right),\quad(C\in\mathcal{B}(H)|_X\otimes\mathcal{B}(\mathbb{R})).\tag{15}$$

  Function g in Eq. (14) is measurable iff its coordinate functions, g_1 and g_2 are both measurable functions [39, Proposition 3.2, page 201]. Thus, we need for $\forall A\in\mathcal{B}(H)|_X$, $\forall B\in\mathcal{B}(\mathbb{R})$

$$\mathcal{B}(\tau_w)\otimes\mathcal{B}(\mathbb{R})\ni g_1^{-1}(A)=\{(x,y):g_1(x,y)=\mu_x\in A\}=\mu^{-1}(A)\times\mathbb{R},\tag{16}$$
$$\mathcal{B}(\tau_w)\otimes\mathcal{B}(\mathbb{R})\ni g_2^{-1}(B)=\{(x,y):g_2(x,y)=y\in B\}=\mathcal{M}_1^+(\mathcal{X})\times B.\tag{17}$$

  According to Eqs. (16)-(17), the measurability of g follows from the X-measurability of $\mu:(\mathcal{M}_1^+(\mathcal{X}),\mathcal{B}(\tau_w))\rightarrow(X,\mathcal{B}(H)|_X)$, which is guaranteed by our conditions.

A.1.2 Proof of Ψ: Hölder continuous $\Rightarrow$ K: measurable

[5]'s original assumption that $(\mu_a,\mu_b)\in X\times X\mapsto K(\mu_a,\mu_b)\in\mathbb{R}$ is measureable follows from the required Hölder continuity [see Eq. (5)] since (i) the continuity of Ψ is equivalent to that of K, (ii) a continuous map between topological spaces is Borel measurable [6, Lemma 4.29 on page 128; page 480].

A.1.3 Proof of K: linear $\Rightarrow$ Ψ: Hölder continuous with $L=1$, $h=1$

In case of a linear K kernel $K(\mu_a,\mu_b)=\langle\mu_a,\mu_b\rangle_H\quad(\mu_a,\mu_b\in X)$, by the bilinearity of $\langle\cdot,\cdot\rangle_H$ and $\|\langle\cdot,a\rangle_H\|_{\mathcal{H}}^2=\|a\|_H^2$, we get that $\|K(\cdot,\mu_a)-K(\cdot,\mu_b)\|_{\mathcal{H}}=\|\langle\cdot,\mu_a\rangle_H-\langle\cdot,\mu_b\rangle_H\|_{\mathcal{H}}=\|\langle\cdot,\mu_a-\mu_b\rangle_H\|_{\mathcal{H}}=\|\mu_a-\mu_b\|_H$. In other words, Hölder continuity holds with $L=1$, $h=1$; K is Lipschitz continuous ($h=1$).

A.1.4 Proof of $\mathcal{X}$: compact metric, μ: continuous $\Rightarrow$ $X=\mu\left(\mathcal{M}_1^+(\mathcal{X})\right)$: compact metric

Let us suppose that $\mathcal{X}=(\mathcal{X},d)$ is a compact metric space. This implies that $\mathcal{M}_1^+(\mathcal{X})$ is also a compact metric space by Theorem 6.4 in [40] (page 55). The continuous (μ) image of a compact set is compact (see page 478 in [6]), thus $X=\mu\left(\mathcal{M}_1^+(\mathcal{X})\right)\subseteq H$ is compact metric.

A.1.5 Proof of the Kernel Examples on $X = \mu(\mathcal{M}_1^+(\mathcal{X}))$

Below we prove for the $K : X \times X \to \mathbb{R}$ functions in Table 1 that they are kernels on mean embedded distributions.

We need some definitions and lemmas. $\mathbb{Z}, \mathbb{Z}^+, \mathbb{R}^+, \mathbb{R}^{\geq 0}$ denotes the set of integers, positive integers, positive real numbers and non-negative real numbers, respectively.

Definition 1. *Let X be a non-empty set. A $K : X \times X \to \mathbb{R}$ function is called*

- positive definite *(pd; also referred to as kernel) on X, if it is*
  1. *symmetric [$K(a,b) = K(b,a)$, $\forall a, b \in X$], and*
  2. *$\sum_{i,j=1}^n c_i c_j K(a_i, a_j) \geq 0$ for all $n \in \mathbb{Z}^+$, $\{a_1, \ldots, a_n\} \subseteq X^n$, $\mathbf{c} = [c_1; \ldots; c_n] \in \mathbb{R}^n$.*
- negative definite *(nd; sometimes $-K$ is called conditionally positive definite) on X, if it is*
  1. *symmetric, and*
  2. *$\sum_{i,j=1}^n c_i c_j K(a_i, a_j) \leq 0$ for all $n \in \mathbb{Z}^+$, $\{a_1, \ldots, a_n\} \subseteq X^n$, $\mathbf{c} = [c_1; \ldots; c_n] \in \mathbb{R}^n$, where $\sum_{j=1}^n c_j = 0$.*

We will use the following properties of positive/negative definite functions:

1. K is nd $\Leftrightarrow e^{-tK}$ is pd for all $t > 0$; see Chapter 3 in [33].
2. $K : X \times X \to \mathbb{R}^{\geq 0}$ is nd $\Leftrightarrow \frac{1}{t+K}$ is pd for all $t > 0$; see Chapter 3 in [33].
3. If K is nd and non-negative on the diagonal ($K(x,x) \geq 0$, $\forall x \in X$), then K^α is nd for all $\alpha \in [0,1]$; see Chapter 3 in [33].
4. $K(x,y) = \langle x, y \rangle_X$ is pd, where X is a Hilbert space (since the pd property is equivalent to being a kernel).
5. $K(x,y) = \|x - y\|_X^2$ is nd, where X is a Hilbert space; see Chapter 3 in [33].
6. If K is nd, $K + d$ ($d \in \mathbb{R}$) is also nd. Proof: (i) $K(x,y) + d = K(y,x) + d$ holds by the symmetry of K, (ii) $\sum_{i,j=1}^n c_i c_j [K(a_i,a_j) + d] = \sum_{i,j=1}^n c_i c_j K(a_i,a_j) + \sum_{i=1}^n c_i \sum_{j=1}^n c_j d = \sum_{i,j=1}^n c_i c_j K(a_i,a_j) + \sum_{i=1}^n c_i d \sum_{j=1}^n c_j = \sum_{i,j=1}^n c_i c_j K(a_i,a_j) + 0 \leq 0$, where we used that $\sum_{j=1}^n c_j = 0$ and K is nd.
7. If K is pd (nd) on X, then it is pd (nd) on $X' \subseteq X$ as well. Proof: less constraints have to be satisfied for $X' \subseteq X$.
8. If K is pd (nd) on X, then sK ($s \in \mathbb{R}^+$) is also pd (nd). Proof: multiplication by a positive constant does not affect the sign of $\sum_{i,j=1}^n c_i c_j K(a_i,a_j)$.
9. If K is nd on X and $K(x,y) > 0$ $\forall x, y \in X$, then $\frac{1}{K}$ is pd; see Chapter 3 in [33].
10. If K is pd on X, and $h(u) = \sum_{n=0}^\infty a_n u^n$ with $a_n \geq 0$, then $h \circ K$ is pd; see Chapter 3 in [33].

Making use of these properties one can prove the kernel property of the K-s in Table 1 (see also Table 3) as follows. All the K-s are functions of $\|\mu_a - \mu_b\|_H$, $\|\mu_a - \mu_b\|_H = \|\mu_b - \mu_a\|_H$, hence K-s are symmetric.

$K(x,y) = \|x-y\|_H^2$ is nd on $H = H(k)$ (Prop. 5), thus $K(x,y) = \|x-y\|_H^2$ is nd on $X = \mu\left(\mathcal{M}_1^+(\mathcal{X})\right) \subseteq H(k)$ (Prop. 7). Consequently, $K(x,y) = \|x-y\|_H^d$ is nd on X, where $d \in [0,2]$ ($K(x,x) = 0 \geq 0$, Prop. 3).

- Hence, $K(x,y) = e^{-t\|x-y\|_H^d}$ is pd, where $t > 0$, $d \in [0,2]$ (Prop. 1). By the $(t,d) = \left(\frac{1}{2\theta^2}, 2\right)$ and $(t,d) = \left(\frac{1}{2\theta^2}, 1\right)$ choices, we get that K_G and K_e are kernels.
- Using Prop. 2 ($\|x-y\|_H^d \geq 0$), one obtains that $K(x,y) = \frac{1}{t + \|x-y\|_H^d}$ is pd on X, where $t > 0$, $d \in [0,2]$. By the $(t,d) = (1, \leq 2)$ choice the kernel property of K_t follows.
- Thus, $K(x,y) = s\|x-y\|_H^d$ is nd on X, where $s > 0$, $d \in [0,2]$ (Prop. 8). Consequently, $K(x,y) = \frac{1}{t + s\|x-y\|_H^d}$ is pd on X, where $s > 0$, $d \in [0,2]$, $t > 0$ (Prop. 2). By the $(d,t,s) = (2, 1, \frac{1}{\theta^2})$, we have that K_C is kernel.
- Hence, $K(x,y) = \|x-y\|_H^d + e$ is nd on X, where $d \in [0,2]$, $e \in \mathbb{R}^+$ (Prop. 6). Thus, $K(x,y) = \left(\|x-y\|_H^d + e\right)^f$ is nd on X, where $d \in [0,2]$, $e \in \mathbb{R}^+$, $f \in (0,1]$ ($\|x-y\|_H^d + e \geq 0$, Prop. 3). Consequently, $K(x,y) = \frac{1}{\left(\|x-y\|_H^d + e\right)^f}$ is pd on X, where $d \in [0,2]$, $e \in \mathbb{R}^+$, $f \in (0,1]$ ($\left(\|x-y\|_H^d + e\right)^f > 0$; Prop. 9); with the $(d,e,f) = \left(2, \theta^2, \frac{1}{2}\right)$ choice, one obtains that K_i is a kernel.

A.1.6 Proof of "Conditions of Proof A.1.4 and Proof A.1.5" $\Rightarrow$ Ψ-s of K-s in Proof A.1.5: Hölder continuous

We tackle the problem more generally:

Table 3: Nonlinear kernels on mean embedded distributions.

Kernel (K)	Parameter(s)
$K(\mu_a,\mu_b) = e^{-t\|\mu_a-\mu_b\|_H^d}$	$t>0$, $d\in[0,2]$
$K(\mu_a,\mu_b) = \frac{1}{t+\|\mu_a-\mu_b\|_H^d}$	$t>0$, $d\in[0,2]$
$K(\mu_a,\mu_b) = \frac{1}{t+s\|\mu_a-\mu_b\|_H^d}$	$s>0$, $d\in[0,2]$, $t>0$
$K(\mu_a,\mu_b) = \frac{1}{\left(\|\mu_a-\mu_b\|_H^d+e\right)^f}$	$d\in[0,2]$, $e\in\mathbb{R}^+$

1. we give sufficient conditions for K kernels of the form

$$K(\mu_a,\mu_b) = \bar{K}\left(\|\mu_a-\mu_b\|_H\right), \tag{18}$$

i.e., for radial kernels to have Hölder continuous canonical feature map ($\Psi(\mu_c) = K(\cdot,\mu_c)$): $\exists L>0$, $h\in(0,1]$ such that $\|K(\cdot,\mu_a)-K(\cdot,\mu_b)\|_{\mathcal{H}} \le L\|\mu_a-\mu_b\|_H^h$.

2. Then we show that these sufficient conditions are satisfied for the K kernels listed in Table 1.

Let us first note that K is bounded. Indeed, since Ψ is Hölder continuous, specially it is continuous. Hence using Lemma 4.29 in [6] (page 128), the

$$K_0 : \mu_a \in X \to K(\mu_a,\mu_a) \in \mathbb{R}$$

mapping is continuous. As we have already seen (Section A.1.4) X is compact. The continuous (K_0) image of a compact set (X), i.e., the $\{K(\mu_a,\mu_a):\mu_a\in X\}\subseteq\mathbb{R}$ set is compact, specially it is bounded above.

1. Sufficient conditions: Now, we present sufficient conditions for the assumed Hölder continuity

$$\|K(\cdot,\mu_a)-K(\cdot,\mu_b)\|_{\mathcal{H}} \le L\|\mu_a-\mu_b\|_H^h. \tag{19}$$

Using $\|u\|_{\mathcal{H}}^2 = \langle u,u\rangle_{\mathcal{H}}$, the bilinearity of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, the reproducing property of K and Eq. (18), we get

$$\begin{aligned}\|K(\cdot,\mu_a)-K(\cdot,\mu_b)\|_{\mathcal{H}}^2 &= \langle K(\cdot,\mu_a)-K(\cdot,\mu_b), K(\cdot,\mu_a)-K(\cdot,\mu_b)\rangle_{\mathcal{H}}\\ &= K(\mu_a,\mu_a)+K(\mu_b,\mu_b)-2K(\mu_a,\mu_b) = 2\bar{K}(0)-2\bar{K}(\|\mu_a-\mu_b\|_H)\\ &= 2\left[\bar{K}(0)-\bar{K}\left(\|\mu_a-\mu_b\|_H\right)\right].\end{aligned}$$

Hence, the Hölder continuity of K is equivalent to the existence of an $L'\left(=\frac{L^2}{2}\right)>0$ such that

$$\bar{K}(0)-\bar{K}\left(\|\mu_a-\mu_b\|_H\right) \le L'\|\mu_a-\mu_b\|_H^{2h}.$$

Since for $\mu_a=\mu_b$ both sides are equal to 0, this requirement is equivalent to

$$u(\mu_a,\mu_b) := \frac{\bar{K}(0)-\bar{K}\left(\|\mu_a-\mu_b\|_H\right)}{\|\mu_a-\mu_b\|_H^{2h}} \le L', \quad (\mu_a\ne\mu_b)$$

i.e., that the $u: X\times X\to\mathbb{R}$ function is bounded above. Function u is the composition ($u=u_2\circ u_1$) of the mappings:

$$\begin{aligned}&u_1: X\times X\to\mathbb{R}^{\ge 0}, \quad u_1(\mu_a,\mu_b) = \|\mu_a-\mu_b\|_H,\\ &u_2:\mathbb{R}^{\ge 0}\to\mathbb{R}, \quad u_2(v) = \frac{\bar{K}(0)-\bar{K}(v)}{v^{2h}}.\end{aligned} \tag{20}$$

Here, u_1 is continuous. Let us *suppose* for u_2 that

(a) (i) $\exists h\in(0,1]$ such that $\lim_{v\to 0+} u_2(v)$ exists, and

(b) u_2 is continuous.

In this case, since the composition of continuous functions is continuous (see page 85 in [41]), u is continuous. As we have seen (Section A.1.4), X is compact. The product of compact sets ($X \times X$) is compact by the Tychonoff theorem (see page 143 in [41]). Finally, since the continuous (u) image of a compact set ($X \times X$), i.e. $\{u(\mu_a, \mu_b) : (\mu_a, \mu_b) \in X \times X\} \subseteq \mathbb{R}$ is compact (Theorem 8 in [41], page 141), we get that u is bounded, specially bounded above.

To sum up, we have proved that if

(a) K is radial [see Eq. (18)],

(b) u_2 [Eq. (20)] is (i) continuous and (ii) $\exists h \in (0,1]$ such that $\lim_{v\to 0+} u_2(v)$ exists,

then the Hölder property [Eq. (19)] holds for K with exponent h. In other words, the Hölder property of a kernel K on mean embedded distributions can be simply guaranteed by the appropriate behavior of $\bar{K}$ at zero.

2. Verification of the sufficient conditions: In the sequel we show that these conditions hold for the u_2 functions of the K kernels in Table 1. In the examples

$$\bar{K}_G(v) = e^{-\frac{v^2}{2\theta^2}}, \quad \bar{K}_e(v) = e^{-\frac{v}{2\theta^2}}, \quad \bar{K}_C(v) = \frac{1}{1+\frac{v^2}{\theta^2}}, \quad \bar{K}_t(v) = \frac{1}{1+v^\theta}, \quad \bar{K}_i(v) = \frac{1}{\sqrt{v^2+\theta^2}}.$$

The corresponding u_2 functions are

$$u_{2G}(v) = \frac{1-e^{-\frac{v^2}{2\theta^2}}}{v^{2h}}, \quad u_{2e}(v) = \frac{1-e^{-\frac{v}{2\theta^2}}}{v^{2h}}, \quad u_{2C}(v) = \frac{1-\frac{1}{1+\frac{v^2}{\theta^2}}}{v^{2h}}, \quad u_{2t}(v) = \frac{1-\frac{1}{1+v^\theta}}{v^{2h}}, \quad u_{2i}(v) = \frac{\frac{1}{\theta}-\frac{1}{\sqrt{v^2+\theta^2}}}{v^{2h}}.$$

The limit requirements at zero complementing the continuity of u_2-s are satisfied:

- u_{2G}: In this case
$$\lim_{v\to 0+} u_{2G}(v) = \lim_{v\to 0+} \frac{1-e^{-\frac{v^2}{2\theta^2}}}{v^2} = \lim_{v\to 0+} \frac{1-e^{-\frac{v}{2\theta^2}}}{v} = \lim_{v\to 0+} \frac{\frac{1}{2\theta^2}e^{-\frac{v}{2\theta^2}}}{1} = \frac{1}{2\theta^2},$$
where we applied a v^2 substitution and the L'Hopital rule; $h = 1$.
- u_{2e}:
$$\lim_{v\to 0+} u_{2e}(v) = \lim_{v\to 0+} \frac{1-e^{-\frac{v}{2\theta^2}}}{v^{2h}} = \lim_{v\to 0+} \frac{\frac{1}{2\theta^2}e^{-\frac{v}{2\theta^2}}}{2hv^{2h-1}} = \frac{1}{2\theta^2},$$
where we applied the L'Hopital rule and chose $h = \frac{1}{2}$, the largest h from the $2h-1 \le 0$ convergence domain.
- u_{2C}:
$$u_{2C}(v) = \frac{1-\frac{1}{1+\frac{v^2}{\theta^2}}}{v^{2h}} = \frac{1-\frac{\theta^2}{\theta^2+v^2}}{v^{2h}} = \frac{\frac{v^2}{\theta^2+v^2}}{v^{2h}} = \frac{v^{2-2h}}{\theta^2+v^2 \xrightarrow{v\to 0+} \theta^2},$$
we chose $h = 1$, the largest value from the convergence domain ($2-2h \ge 0 \Rightarrow 1 \ge h$).
- u_{2t}:
$$u_{2t}(v) = \frac{1-\frac{1}{1+v^\theta}}{v^{2h}} = \frac{v^{\theta-2h}}{1+v^\theta \xrightarrow{v\to 0} 1},$$
thus we can have $h = \frac{\theta}{2}$, the largest element of the convergence domain ($\theta - 2h \ge 0 \Leftrightarrow \frac{\theta}{2} \ge h$). Here we require $\theta \le 2$ in order to guarantee that $h = \frac{\theta}{2} \le 1$.
- u_{2i}: Let g denote the nominator of u_{2i}
$$g(v) = \frac{1}{\theta} - \frac{1}{\sqrt{v^2+\theta^2}} = \frac{1}{\theta} - \left[g(0) + g'(0)v + \frac{g''(0)}{2}v^2 + \dots\right]$$
$$= \frac{1}{\theta} - \left[\frac{1}{\theta} + \left(-\frac{1}{2}\frac{1}{(v^2+\theta)^{\frac{3}{2}}}2v\right)\Big|_{v=0} v + \frac{g''(0)}{2}v^2 + \dots\right] = -v^2\left[\frac{g''(0)}{2!} + \frac{g^{(3)}(0)}{3!}v + \dots\right].$$
Hence,
$$\lim_{v\to 0+} u_{2i}(v) = \lim_{v\to 0+} \frac{\frac{1}{\theta}-\frac{1}{\sqrt{v^2+\theta^2}}}{v^2} = \lim_{v\to 0+} \frac{-v^2\left[\frac{g''(0)}{2!} + \frac{g^{(3)}(0)}{3!}v + \dots\right]}{v^2} = -\frac{g''(0)}{2},$$
i.e., h can be chosen to be 1 ($h = 1$).

A.1.7 Proof of $\left\|\sum_{i=1}^{n} f_i\right\|^2 \le n \sum_{i=1}^{n} \|f_i\|^2$

In a normed space $(N, \|\cdot\|)$

$$\left\|\sum_{i=1}^{n} f_i\right\|^2 \le n \sum_{i=1}^{n} \|f_i\|^2 , \tag{21}$$

where $f_i \in N$ $(i = 1, \ldots, n)$.

Indeed the statement holds since $\left\|\sum_{i=1}^{n} f_i\right\|^2 \le \left(\sum_{i=1}^{n} \|f_i\|\right)^2 \le n \sum_{i=1}^{n} \|f_i\|^2$, where we applied the triangle inequality, and a consequence that the arithmetic mean is smaller or equal than the squared mean (special case of the generalized mean inequality) with $a_i = \|f_i\| \ge 0$. Particularly, $\frac{\sum_{i=1}^{n} a_i}{n} \le \sqrt{\frac{\sum_{i=1}^{n} (a_i)^2}{n}} \Rightarrow \left(\sum_{i=1}^{n} a_i\right)^2 \le n \sum_{i=1}^{n} (a_i)^2$.

A.1.8 Proof of the Decomposition of the Excess Risk

It is known [5] that $\mathcal{E}[f] - \mathcal{E}[f_{\mathcal{H}}] = \|\sqrt{T}(f - f_{\mathcal{H}})\|_{\mathcal{H}}^2 \quad (\forall f \in \mathcal{H})$. Applying this identity with $f = f_{\hat{\mathbf{z}}}^{\lambda} \in \mathcal{H}$ and a telescopic trick, we get

$$\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] = \left\|\sqrt{T}\left(f_{\hat{\mathbf{z}}}^{\lambda} - f_{\mathcal{H}}\right)\right\|_{\mathcal{H}}^2 = \left\|\sqrt{T}\left[\left(f_{\hat{\mathbf{z}}}^{\lambda} - f_{\mathbf{z}}^{\lambda}\right) + \left(f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right) + \left(f^{\lambda} - f_{\mathcal{H}}\right)\right]\right\|_{\mathcal{H}}^2 . \tag{22}$$

By Eqs. (9), (10), and the operator identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ one obtains for the first term in Eq. (22)

$$\begin{aligned}
f_{\hat{\mathbf{z}}}^{\lambda} - f_{\mathbf{z}}^{\lambda} &= (T_{\hat{\mathbf{x}}} + \lambda)^{-1} g_{\hat{\mathbf{z}}} - (T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}} = (T_{\hat{\mathbf{x}}} + \lambda)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}) + (T_{\hat{\mathbf{x}}} + \lambda)^{-1} g_{\mathbf{z}} - (T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}} \\
&= (T_{\hat{\mathbf{x}}} + \lambda)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}) + \left[(T_{\hat{\mathbf{x}}} + \lambda)^{-1} - (T_{\mathbf{x}} + \lambda)^{-1}\right] g_{\mathbf{z}} \\
&= (T_{\hat{\mathbf{x}}} + \lambda)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}) + \left[(T_{\hat{\mathbf{x}}} + \lambda)^{-1}(T_{\mathbf{x}} - T_{\hat{\mathbf{x}}})(T_{\mathbf{x}} + \lambda)^{-1}\right] g_{\mathbf{z}} \\
&= (T_{\hat{\mathbf{x}}} + \lambda)^{-1}\left[(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}) + (T_{\mathbf{x}} - T_{\hat{\mathbf{x}}})(T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}}\right] = (T_{\hat{\mathbf{x}}} + \lambda)^{-1}\left[(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}) + (T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}) f_{\mathbf{z}}^{\lambda}\right].
\end{aligned}$$

Thus, we can rewrite the first term in (22) as

$$\sqrt{T}\left(f_{\hat{\mathbf{z}}}^{\lambda} - f_{\mathbf{z}}^{\lambda}\right) =: f_{-1} + f_0, \qquad f_{-1} = \sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}), \qquad f_0 = \sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda)^{-1}(T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}) f_{\mathbf{z}}^{\lambda}.$$

The second term in (22) can be decomposed [5] as

$$\begin{aligned}
\sqrt{T}\left[\left(f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right) + \left(f^{\lambda} - f_{\mathcal{H}}\right)\right] &= \sqrt{T}\left[(T_{\mathbf{x}} + \lambda)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}}) + (T_{\mathbf{x}} + \lambda)^{-1}(T - T_{\mathbf{x}})(f^{\lambda} - f_{\mathcal{H}}) + (f^{\lambda} - f_{\mathcal{H}})\right] \\
&=: f_1 + f_2 + f_3,
\end{aligned}$$

where

$$f_1 = \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}}), \qquad f_2 = \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T - T_{\mathbf{x}})(f^{\lambda} - f_{\mathcal{H}}), \qquad f_3 = \sqrt{T}(f^{\lambda} - f_{\mathcal{H}}).$$

Using these f_i notations, (22) can be upper bounded as

$$\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] = \left\|\sum_{i=-1}^{3} f_i\right\|_{\mathcal{H}}^2 \le 5 \sum_{i=-1}^{3} \|f_i\|_{\mathcal{H}}^2 , \tag{23}$$

exploiting Section A.1.7 ($\|\cdot\|^2 = \|\cdot\|_{\mathcal{H}}^2$, $n = 5$). Consequently, introducing the

$$\begin{aligned}
&S_{-1} = S_{-1}(\lambda, \mathbf{z}, \hat{\mathbf{z}}) = \|f_{-1}\|_{\mathcal{H}}^2, && S_0 = S_0(\lambda, \mathbf{z}, \hat{\mathbf{z}}) = \|f_0\|_{\mathcal{H}}^2, \\
&S_1 = S_1(\lambda, \mathbf{z}) = \|f_1\|_{\mathcal{H}}^2, && S_2 = S_2(\lambda, \mathbf{z}) = \|f_2\|_{\mathcal{H}}^2, && \mathcal{A}(\lambda) = \|f_3\|_{\mathcal{H}}^2,
\end{aligned}$$

notations (for $\mathcal{A}(\lambda)$ see also Theorem 4), (23) can be rewritten as

$$\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] \le 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda) + S_1 + S_2\right]. \tag{24}$$

A.1.9 Proof of the Upper Bounding Terms of S_{-1} and S_0

Using the

$$\|Mu\|_{\mathcal{H}} \leq \|M\|_{\mathcal{L}(\mathcal{H})}\|u\|_{\mathcal{H}} \quad (M \in \mathcal{L}(\mathcal{H}), u \in \mathcal{H}), \tag{25}$$

relation, we get

$$S_{-1} \leq \left\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\right\|^2_{\mathcal{L}(\mathcal{H})} \|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|^2_{\mathcal{H}},$$

$$S_0 \leq \left\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\right\|^2_{\mathcal{L}(\mathcal{H})} \left\|(T_{\mathbf{x}} - T_{\hat{\mathbf{x}}})f^{\lambda}_{\mathbf{z}}\right\|^2_{\mathcal{H}} \leq \left\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\right\|^2_{\mathcal{L}(\mathcal{H})} \|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})} \left\|f^{\lambda}_{\mathbf{z}}\right\|^2_{\mathcal{H}}.$$

A.1.10 Proof of the Convergence Rate of the Empirical Mean Embedding

The statement we prove is as follows.[25][9]

Let $\mu_x = \int_{\mathcal{X}} k(\cdot,u)\mathrm{d}x(u)$ denote the mean embedding of distribution $x \in \mathcal{M}_1^+(\mathcal{X})$ to the $H = H(k)$ RKHS determined by kernel k ($\mu_x \in H$), which is assumed to be bounded $k(u,u) \leq B_k$ ($\forall u \in \mathcal{X}$). Let us given N i.i.d. samples from distribution x: x_1, ..., x_N. Let $\mu_{\hat{x}} = \frac{1}{N}\sum_{n=1}^N k(\cdot,x_n) \in H$ be the empirical mean embedding. Then $\mathbb{P}\left(\|\mu_{\hat{x}} - \mu_x\|_H \leq \frac{\sqrt{2B_k}}{\sqrt{N}} + \epsilon\right) \geq 1 - e^{-\frac{\epsilon^2 N}{2B_k}}$, or

$$\|\mu_{\hat{x}_i} - \mu_{x_i}\|_H \leq \frac{\sqrt{2B_k}}{\sqrt{N}} + \frac{\sqrt{2\alpha B_k}}{\sqrt{N}} = \frac{(1+\sqrt{\alpha})\sqrt{2B_k}}{\sqrt{N}}$$

with probability at least $1 - e^{-\alpha}$, where $\alpha = \frac{\epsilon^2 N}{2B_k}$.

The proof will make use of the McDiarmid's inequality.

Lemma 1 (**McDiarmid's inequality** [42]). *Let $x_1, \ldots, x_N \in \mathcal{X}$ be independent random variables and function $g \in \mathcal{X}^n \to \mathbb{R}$ be such that $\sup_{u_1,\ldots,u_N,u'_j \in \mathcal{X}} \left|g(u_1,\ldots,u_N) - g(u_1,\ldots,u_{j-1},u'_j,u_{j+1},\ldots,u_N)\right| \leq c_j$ $\forall j = 1,\ldots,N$. Then for all $\epsilon > 0$ $\mathbb{P}\left(g(x_1,\ldots,x_N) - \mathbb{E}\left[g(x_1,\ldots,x_N)\right] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{n=1}^N c_n^2}}$.*

Namely, let $\phi(u) = k(\cdot,u)$, and thus $k(u,u) = \|\phi(u)\|^2_H$. Let us define

$$g(S) = \|\mu_{\hat{x}} - \mu_x\|_H = \left\|\frac{1}{N}\sum_{n=1}^N \phi(x_n) - \mu_x\right\|_H,$$

where $S = \{x_1,\ldots,x_N\}$ be the sample set. Define $S' = \{x_1,\ldots,x_{j-1},x'_j,x_{j+1},\ldots,x_N\}$, i.e., let us replace in the sample set x_j with x'_j. Then

$$\begin{aligned}|g(S) - g(S')| &= \left|\left\|\frac{1}{N}\sum_{n=1}^N \phi(x_n) - \mu_x\right\|_H - \left\|\frac{1}{N}\sum_{n=1;n\neq j}^N \phi(x_n) + \frac{1}{N}\phi(x'_j) - \mu_x\right\|_H\right| \\ &\leq \frac{1}{N}\left\|\phi(x_j) - \phi(x'_j)\right\|_H \leq \frac{1}{N}\left(\|\phi(x_j)\|_H + \left\|\phi(x'_j)\right\|_H\right) \leq \frac{1}{N}\left[\sqrt{k(x_j,x_j)} + \sqrt{k\left(x'_j,x'_j\right)}\right] \leq \frac{2\sqrt{B_k}}{N}\end{aligned}$$

based on (i) the reverse and the standard triangle inequality, and (ii) the boundedness of kernel k. By using the McDiarmid's inequality (Lemma 1), we get

$$\mathbb{P}\left(g(S) - \mathbb{E}[g(S)] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{n=1}^N\left(\frac{2\sqrt{B_k}}{N}\right)^2}} = e^{-\frac{2\epsilon^2}{N\frac{4B_k}{N^2}}} = e^{-\frac{\epsilon^2 N}{2B_k}},$$

or, in other words

$$1 - e^{-\frac{\epsilon^2 N}{2B_k}} \leq \mathbb{P}\left(g(S) < \mathbb{E}[g(S)] + \epsilon\right) \leq \mathbb{P}\left(g(S) \leq \mathbb{E}[g(S)] + \epsilon\right).$$

[9] In the original result a factor of 2 is missing from the denominator in the exponential function; we correct the proof here.

Considering the $\mathbb{E}[g(S)]$ term: since for a non-negative random variable (a) the $\mathbb{E}(a)=\mathbb{E}(a1)\leq\sqrt{\mathbb{E}(a^2)}\sqrt{\mathbb{E}(1^2)}=\sqrt{\mathbb{E}(a^2)}$ inequality holds due to the CBS, we obtain

$$\mathbb{E}[g(S)]=\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\phi(x_n)-\mu_x\right\|_H\right]\leq\sqrt{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\phi(x_n)-\mu_x\right\|_H^2\right]}$$
$$=\sqrt{\mathbb{E}\left[\left\langle\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)-\mu_x,\frac{1}{N}\sum_{j=1}^{N}\phi(x_j)-\mu_x\right\rangle_H\right]}=\sqrt{b+c+d}$$

using that $\|a\|_H^2=\sqrt{\langle a,a\rangle_H}$. Here,

$$b=\mathbb{E}\left[\frac{1}{N^2}\left(\sum_{i,j=1;i\neq j}^{N}k(x_i,x_j)+\sum_{i=1}^{N}k(x_i,x_i)\right)\right]=\frac{N(N-1)}{N^2}\mathbb{E}_{t\sim x,t'\sim x}k(t,t')+\frac{N}{N^2}\mathbb{E}_{t\sim x}\left[k(t,t)\right],$$
$$c=-\frac{2}{N}\mathbb{E}\left[\left\langle\sum_{i=1}^{N}\phi(x_i),\mu_x\right\rangle_H\right]=-\frac{2N}{N}\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right],$$
$$d=\mathbb{E}\left[\|\mu_x\|_H^2\right]=\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]$$

applying the bilinearity of $\langle\cdot,\cdot\rangle_H$, and the representation property of μ_x. Thus,

$$\sqrt{b+c+d}=\sqrt{\left[\frac{N-1}{N}-2+1\right]\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]+\frac{1}{N}\mathbb{E}_{t\sim x}\left[k(t,t)\right]}$$
$$=\sqrt{\frac{1}{N}\left(\mathbb{E}_{t\sim x}\left[k(t,t)\right]-\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]\right)}=\frac{\sqrt{\mathbb{E}_{t\sim x}\left[k(t,t)\right]-\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]}}{\sqrt{N}}.$$

Since

$$\sqrt{\mathbb{E}_{t\sim x}\left[k(t,t)\right]-\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]}\leq\sqrt{\left|\mathbb{E}_{t\sim x}\left[k(t,t)\right]\right|+\left|\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]\right|}\leq\sqrt{\mathbb{E}_{t\sim x}\left|k(t,t)\right|+\mathbb{E}_{t\sim x,t'\sim x}\left|k(t,t')\right|},$$

where we applied the triangle inequality, $|k(t,t)|=k(t,t)\leq B_k$ and $|k(t,t')|\leq\sqrt{k(t,t)}\sqrt{k(t',t')}$ (which holds to the CBS), we get $\sqrt{\mathbb{E}_{t\sim x}\left[k(t,t)\right]-\mathbb{E}_{t\sim x,t'\sim x}\left[k(t,t')\right]}\leq\sqrt{B_k+\sqrt{B_k}\sqrt{B_k}}=\sqrt{2B_k}$.

To sum up, we obtained that $\|\mu_x-\mu_{\hat{x}}\|_H\leq\frac{\sqrt{2B_k}}{\sqrt{N}}+\epsilon$ holds with probability at least $1-e^{-\frac{\epsilon^2N}{2B_k}}$. This is what we wanted to prove.

A.1.11 Proof of the Bound on $\|g_{\hat{\mathbf{z}}}-g_{\mathbf{z}}\|_{\mathcal{H}}^2$, $\|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$

Below, we present the detailed derivations of the upper bounds on $\|g_{\hat{\mathbf{z}}}-g_{\mathbf{z}}\|_{\mathcal{H}}^2$, $\|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$ and $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$.

- **Bound on** $\|g_{\hat{\mathbf{z}}}-g_{\mathbf{z}}\|_{\mathcal{H}}^2$: By (11), we have $g_{\hat{\mathbf{z}}}-g_{\mathbf{z}}=\frac{1}{l}\sum_{i=1}^{l}\left[K(\cdot,\mu_{\hat{x}_i})-K(\cdot,\mu_{x_i})\right]y_i$. Applying Eq. (21), the Hölder property of K, the homogenity of norms $\|av\|=|a|\,\|v\|$ $(a\in\mathbb{R})$, assuming that y_i is bounded $(|y_i|\leq C)$, and using (13), we obtain

$$\|g_{\hat{\mathbf{z}}}-g_{\mathbf{z}}\|_{\mathcal{H}}^2\leq\frac{1}{l^2}l\sum_{i=1}^{l}\|K(\cdot,\mu_{\hat{x}_i})-K(\cdot,\mu_{x_i})y_i\|_{\mathcal{H}}^2\leq\frac{L^2}{l}\sum_{i=1}^{l}y_i^2\,\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H^{2h}\leq\frac{L^2C^2}{l}\sum_{i=1}^{l}\left[\frac{(1+\sqrt{\alpha})\sqrt{2B_k}}{\sqrt{N}}\right]^{2h}$$
$$=L^2C^2\frac{(1+\sqrt{\alpha})^{2h}(2B_k)^h}{N^h}$$

  with probability at least $1-le^{-\alpha}$, based on a union bound.

- **Bound on** $\|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})}$: Using the definition of $T_{\mathbf{x}}$ and $T_{\hat{\mathbf{x}}}$, and (21) with the $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ operator norm, we get

$$\|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})}\le\frac{1}{l^2}l\sum_{i=1}^{l}\left\|T_{\mu_{x_i}}-T_{\mu_{\hat{x}_i}}\right\|^2_{\mathcal{L}(\mathcal{H})}. \tag{26}$$

To upper bound the quantities $\|T_{\mu_{x_i}}-T_{\mu_{\hat{x}_i}}\|^2_{\mathcal{L}(\mathcal{H})}$, let us see how T_{μ_u} acts

$$T_{\mu_u}(f)=K(\cdot,\mu_u)\delta_{\mu_u}(f)=K(\cdot,\mu_u)f(\mu_u). \tag{27}$$

If we can prove that

$$\|(T_{\mu_u}-T_{\mu_v})(f)\|_{\mathcal{H}}\le E\,\|f\|_{\mathcal{H}}, \tag{28}$$

then this implies $\|T_{\mu_u}-T_{\mu_v}\|_{\mathcal{L}(\mathcal{H})}\le E$. We continue with the l.h.s. of (28) using (27), (21) with $n=2$, the homogenity of norms, the reproducing and Hölder property of K:

$$\begin{aligned}\|(T_{\mu_u}-T_{\mu_v})(f)\|^2_{\mathcal{H}}&=\|K(\cdot,\mu_u)\delta_{\mu_u}(f)-K(\cdot,\mu_v)\delta_{\mu_v}(f)\|^2_{\mathcal{H}}\\&=\|K(\cdot,\mu_u)\left[\delta_{\mu_u}(f)-\delta_{\mu_v}(f)\right]+\left[K(\cdot,\mu_u)-K(\cdot,\mu_v)\right]\delta_{\mu_v}(f)\|^2_{\mathcal{H}}\\&\le 2\left[\|K(\cdot,\mu_u)\left[\delta_{\mu_u}(f)-\delta_{\mu_v}(f)\right]\|^2_{\mathcal{H}}+\|K(\cdot,\mu_u)-K(\cdot,\mu_v)\delta_{\mu_v}(f)\|^2_{\mathcal{H}}\right]\\&=2\left[\left[\delta_{\mu_u}(f)-\delta_{\mu_v}(f)\right]^2\|K(\cdot,\mu_u)\|^2_{\mathcal{H}}+\left[\delta_{\mu_v}(f)\right]^2\|\left[K(\cdot,\mu_u)-K(\cdot,\mu_v)\right]\|^2_{\mathcal{H}}\right]\\&\le 2\left[\left[\delta_{\mu_u}(f)-\delta_{\mu_v}(f)\right]^2K(\mu_u,\mu_u)+L^2\left[\delta_{\mu_v}(f)\right]^2\|\mu_u-\mu_v\|^{2h}_H\right].\end{aligned}$$

By rewriting the first terms, we arrive at

$$\begin{aligned}\delta_{\mu_u}(f)-\delta_{\mu_v}(f)&=\langle f,K(\cdot,\mu_u)\rangle_{\mathcal{H}}-\langle f,K(\cdot,\mu_v)\rangle_{\mathcal{H}}\le|\langle f,K(\cdot,\mu_u)-K(\cdot,\mu_v)\rangle_{\mathcal{H}}|\\&\le\|f\|_{\mathcal{H}}\|K(\cdot,\mu_u)-K(\cdot,\mu_v)\|_{\mathcal{H}}\le\|f\|_{\mathcal{H}}\,L\,\|\mu_u-\mu_v\|^h_H,\\\delta_{\mu_v}(f)&=\langle f,K(\cdot,\mu_v)\rangle_{\mathcal{H}}\le|\langle f,K(\cdot,\mu_v)\rangle_{\mathcal{H}}|\le\|f\|_{\mathcal{H}}\|K(\cdot,\mu_v)\|_{\mathcal{H}}=\|f\|_{\mathcal{H}}\sqrt{K(\mu_v,\mu_v)},\end{aligned}$$

where we applied the reproducing and Hölder property of K, the bilinearity of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and the CBS inequality. Hence

$$\begin{aligned}\|(T_{\mu_u}-T_{\mu_v})(f)\|^2_{\mathcal{H}}&\le 2\left[\|f\|^2_{\mathcal{H}}L^2\|\mu_u-\mu_v\|^{2h}_H K(\mu_u,\mu_u)+L^2\|f\|^2_{\mathcal{H}}K(\mu_v,\mu_v)\|\mu_u-\mu_v\|^{2h}_H\right]\\&=2L^2\|f\|^2_{\mathcal{H}}\|\mu_u-\mu_v\|^{2h}_H\left[K(\mu_u,\mu_u)+K(\mu_v,\mu_v)\right].\end{aligned}$$

Thus

$$E^2=2L^2\|\mu_u-\mu_v\|^{2h}_H\left[K(\mu_u,\mu_u)+K(\mu_v,\mu_v)\right].$$

Exploiting this property in (26), (4), and (13)

$$\begin{aligned}\|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})}&\le\frac{2L^2}{l}\sum_{i=1}^{l}\|\mu_{x_i}-\mu_{\hat{x}_i}\|^{2h}_H\left[K(\mu_{x_i},\mu_{x_i})+K(\mu_{\hat{x}_i},\mu_{\hat{x}_i})\right]\le\frac{4B_KL^2}{l}\sum_{i=1}^{l}\frac{\left(1+\sqrt{\alpha}\right)^{2h}(2B_k)^h}{N^h}\\&=\frac{\left(1+\sqrt{\alpha}\right)^{2h}2^{h+2}(B_k)^hB_KL^2}{N^h}.\end{aligned} \tag{29}$$

- **Bound on** $\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\|^2_{\mathcal{L}(\mathcal{H})}$: First we rewrite $T_{\hat{\mathbf{x}}}+\lambda$,

$$T_{\hat{\mathbf{x}}}+\lambda=(T+\lambda)-(T-T_{\hat{\mathbf{x}}})=\left[I-(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right](T+\lambda).$$

Let us now use the Neumann series of $I-(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}$

$$\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}=\sqrt{T}(T+\lambda)^{-1}\sum_{n=0}^{\infty}\left[(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right]^n$$

to have

$$\begin{aligned}\left\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} &= \left\|\sqrt{T}(T+\lambda)^{-1}\sum_{n=0}^{\infty}\left[(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right]^n\right\|_{\mathcal{L}(\mathcal{H})}\\ &\le \left\|\sqrt{T}(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\left\|\sum_{n=0}^{\infty}\left[(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right]^n\right\|_{\mathcal{L}(\mathcal{H})}\\ &\le \left\|\sqrt{T}(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\sum_{n=0}^{\infty}\left\|\left[(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right]^n\right\|_{\mathcal{L}(\mathcal{H})}\\ &\le \left\|\sqrt{T}(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\sum_{n=0}^{\infty}\left\|(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right\|^n_{\mathcal{L}(\mathcal{H})},\end{aligned}$$

where $\|AB\|_{\mathcal{L}(\mathcal{H})} \le \|A\|_{\mathcal{L}(\mathcal{H})}\|B\|_{\mathcal{L}(\mathcal{H})}$ and the triangle inequality was applied. By the spectral theorem, the first term can be bounded as $\|\sqrt{T}(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \le \frac{1}{2\sqrt{\lambda}}$, whereas for the second term, applying a telescopic trick and a triangle inequality, we get

$$\begin{aligned}\left\|(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} &= \left\|\left[(T-T_{\mathbf{x}})+(T_{\mathbf{x}}-T_{\hat{\mathbf{x}}})\right](T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\\ &\le \left\|(T-T_{\mathbf{x}})(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}+\left\|(T_{\mathbf{x}}-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}.\end{aligned}$$

We know that

$$\mathbf{\Theta}(\lambda,\mathbf{z}) := \|(T-T_{\mathbf{x}})(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \le \frac{1}{2} \tag{30}$$

with probability at least $1-\frac{\eta}{3}$ [5]. Considering the second term, using (29) and $\|(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \le \frac{1}{\lambda}$ (by the spectral theorem),

$$\left\|(T_{\mathbf{x}}-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \le \|T_{\mathbf{x}}-T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}\left\|(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \le \frac{\left(1+\sqrt{\alpha}\right)^h 2^{\frac{h}{2}+1}(B_k)^{\frac{h}{2}}(B_K)^{\frac{1}{2}}L}{N^{\frac{h}{2}}}\frac{1}{\lambda}.$$

For fixed λ, the value of N can be chosen such that

$$\begin{aligned}&\frac{\left(1+\sqrt{\alpha}\right)^h 2^{\frac{h}{2}+1}(B_k)^{\frac{h}{2}}(B_K)^{\frac{1}{2}}L}{N^{\frac{h}{2}}}\frac{1}{\lambda} \le \frac{1}{4} \Leftrightarrow \frac{\left(1+\sqrt{\alpha}\right)^h 2^{\frac{h}{2}+3}(B_k)^{\frac{h}{2}}(B_K)^{\frac{1}{2}}L}{\lambda} \le N^{\frac{h}{2}} \Leftrightarrow\\ &\frac{\left(1+\sqrt{\alpha}\right)^2 2^{\frac{h+6}{h}}B_k(B_K)^{\frac{1}{h}}L^{\frac{2}{h}}}{\lambda^{\frac{2}{h}}} \le N.\end{aligned} \tag{31}$$

In this case $\left\|(T-T_{\hat{\mathbf{x}}})(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \le \frac{3}{4}$ (the Neumann series trick is legitimate) and

$$\left\|\sqrt{T}(T_{\hat{\mathbf{x}}}+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \le \frac{1}{2\sqrt{\lambda}}\frac{1}{1-\frac{3}{4}} \le \frac{2}{\sqrt{\lambda}}. \tag{32}$$

- **Bound on** $\|f_{\mathbf{z}}^{\lambda}\|^2_{\mathcal{H}}$: Using the explicit form of $f_{\mathbf{z}}^{\lambda}$ [(9)], (25), the positivity of $T_{\mathbf{x}}$ $\left[\Rightarrow \left\|(T_{\mathbf{x}}+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \le \frac{1}{\lambda}\right]$, the homogenity of norms, Eq. (21), the boundedness assumption on y_i ($|y_i| \le C$), the reproducing property and the boundedness of K [Eq. (4)], we get

$$\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}} \le \left\|(T_{\mathbf{x}}+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\|g_{\mathbf{z}}\|_{\mathcal{H}} \le \frac{1}{\lambda}\|g_{\mathbf{z}}\|_{\mathcal{H}},$$

where

$$\|g_{\mathbf{z}}\|^2_{\mathcal{H}} \le \frac{1}{l^2}l\sum_{i=1}^{l}\|K(\cdot,\mu_{x_i})y_i\|^2_{\mathcal{H}} \le \frac{1}{l}\sum_{i=1}^{l}C^2\|K(\cdot,\mu_{x_i})\|^2_{\mathcal{H}} = \frac{1}{l}\sum_{i=1}^{l}C^2K(\mu_{x_i},\mu_{x_i}) \le \frac{1}{l}\sum_{i=1}^{l}C^2B_K = C^2B_K.$$

Thus, we have obtained that

$$\left\|f_{\mathbf{z}}^{\lambda}\right\|^2_{\mathcal{H}} \le \frac{1}{\lambda^2}C^2B_K. \tag{33}$$

A.1.12 Final Step of the Proof (Union Bound)

Until now, we obtained that if

1. the sample number N satisfies Eq. (31),
2. (13) holds for $\forall i = 1, \ldots, l$ (which has probability at least $1 - le^{-\alpha} = 1 - e^{-[\alpha - \log(l)]} = 1 - e^{-\delta}$ applying a union bound argument; $\alpha = \log(l) + \delta$), and
3. $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ is fulfilled [see Eq. (30)], then

$$
\begin{aligned}
S_{-1} + S_0 &\leq \frac{4}{\lambda} \left[L^2 C^2 \frac{\left(1+\sqrt{\alpha}\right)^{2h} (2B_k)^h}{N^h} + \frac{\left(1+\sqrt{\alpha}\right)^{2h} 2^{h+2} (B_k)^h B_K L^2}{N^h} \frac{C^2 B_K}{\lambda^2} \right] \\
&= \frac{4L^2C^2\left(1+\sqrt{\alpha}\right)^{2h}(2B_k)^h}{\lambda N^h} \left[1 + \frac{4(B_K)^2}{\lambda^2} \right].
\end{aligned}
$$

By taking into account [5]'s bounds for S_1 and S_2

$$
S_1 \leq 32 \log^2\left(\frac{6}{\eta}\right) \left[\frac{B_K M^2}{l^2 \lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l} \right], \qquad S_2 \leq 8 \log^2\left(\frac{6}{\eta}\right) \left[\frac{4B_K^2 \mathcal{B}(\lambda)}{l^2 \lambda} + \frac{B_K \mathcal{A}(\lambda)}{l \lambda} \right],
$$

plugging all the expressions to (24), we obtain Theorem 4 via a union bound.

A.1.13 Proof of Consequence 2

Since constant multipliers do not matter in the orders of rates, we discard them in the (in)equalities below. Our goal is to choose $\lambda = \lambda_{l,N}$ such that

- $\lim_{l,N\to\infty} \lambda_{l,N} = 0$, and
- in Theorem 4: (i') $\frac{\log(l)}{\lambda^{\frac{2}{h}}} \leq N$, (i) $l\lambda^{\frac{b+1}{b}} \geq 1$,[10] and (ii) $r(l, N, \lambda) = \frac{\log^h(l)}{N^h\lambda^3} + \lambda^c + \frac{\lambda^{c-2}}{l^2} + \frac{\lambda^{c-1}}{l} + \frac{1}{l^2\lambda} + \frac{1}{l\lambda^{\frac{1}{b}}} \to 0$.

In $r(l, N, \lambda)$ we will require that the first term goes to zero $\left[\frac{\log^h(l)}{N^h\lambda^3} \to 0\right]$, which implies $\frac{\log(l)}{N\lambda^{\frac{3}{h}}} \to 0$ and hence $\frac{\log(l)}{N\lambda^{\frac{2}{h}}} \to 0$. Thus constraint (i') can be discarded, and our goal is to fulfill (i)-(ii). Since

1. $2 - c \leq 1$ ($\Leftrightarrow 1 \leq c$), $\frac{1}{l^2\lambda^{2-c}} = \frac{\lambda^{c-2}}{l^2} \leq \frac{1}{l^2\lambda}$ (in order), and
2. $c - 1 \geq 0$ ($\Leftrightarrow 1 \leq c$), $\frac{\lambda^{c-1}}{l} \leq \frac{1}{l\lambda^{\frac{1}{b}}}$ (in order)

condition (i)-(ii) reduces to

$$
r(l, N, \lambda) = \frac{\log^h(l)}{N^h\lambda^3} + \lambda^c + \frac{1}{l^2\lambda} + \frac{1}{l\lambda^{\frac{1}{b}}} \to 0, \text{ subject to } l\lambda^{\frac{b+1}{b}} \geq 1. \tag{34}
$$

Our goal is to study the behavior of this quantity in terms of the (l, N, λ) triplet; $1 < b$, $c \in [1, 2]$, $h \in (0, 1]$. To do so, we

1. choose λ such a way that two terms match in order (and $\lambda = \lambda_{l,N} \to 0$);
2. setting $l = N^a$ ($a > 0$) we examine under what conditions (i)-(ii) the convergence of r to 0 holds with the constraint $l\lambda^{\frac{b+1}{b}} \geq 1$ satisfied, (iii) are the matched terms also dominant, i.e., give the convergence rate.

We carry out the computation for all the $\binom{4}{2} = 6$ pairs in Eq. (34). Below we give the derivation of the results summarized in Table 2.

- $\boxed{1} = \boxed{2}$ in Eq. (34) [i.e., the first and second terms are equal in Eq. (34)]:

[10] $\mathcal{N}(\lambda)$ can be upper bounded by (constant multipliers are discarded) $\lambda^{-\frac{1}{b}}$ [5]. Using this upper bound in the l constraint of Theorem 4 we get $l \geq \frac{\lambda^{-\frac{1}{b}}}{\lambda} \Leftrightarrow l\lambda^{\frac{1}{b}+1=\frac{b+1}{b}} \geq 1$.

– (i)-(ii): Exploiting $\frac{h}{c+3}>0$ in the λ choice, we get

$$\frac{\log^h(l)}{N^h\lambda^3}=\lambda^c \Leftrightarrow \left[\frac{\log(l)}{N}\right]^h=\lambda^{c+3}\Leftrightarrow \left[\frac{\log(l)}{N}\right]^{\frac{h}{c+3}}=\lambda\to 0,\ \text{if } \frac{\log(l)}{N}\to 0.$$

$$r(l,N)=\left[\frac{\log(l)}{N}\right]^{\frac{hc}{c+3}}+\frac{1}{l^2\left[\frac{\log(l)}{N}\right]^{\frac{h}{c+3}}}+\frac{1}{l\left[\frac{\log(l)}{N}\right]^{\frac{h}{b(c+3)}}}.$$

$$\begin{aligned} r(N)&=\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}+\frac{1}{N^{2a}\left[\frac{\log(N)}{N}\right]^{\frac{h}{c+3}}}+\frac{1}{N^a\left[\frac{\log(N)}{N}\right]^{\frac{h}{b(c+3)}}}\\ &=\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}+\frac{N^{\frac{h}{c+3}}}{N^{2a}\log^{\frac{h}{c+3}}(N)}+\frac{N^{\frac{h}{b(c+3)}}}{N^a\log^{\frac{h}{b(c+3)}}(N)}. \end{aligned} \tag{35}$$

Here,

- (ii): $r(N)\to 0$ if
  - $\boxed{1}\to 0$: [i.e., the first term goes to zero in Eq. (35)]; no constraint using that $\frac{hc}{c+3}>0$.
  - $\boxed{2}\to 0$: $2a\geq\frac{h}{c+3}$ [$\Leftarrow \frac{h}{c+3}>0$].
  - $\boxed{3}\to 0$: $a\geq\frac{h}{b(c+3)}$ [$\Leftarrow \frac{h}{b(c+3)}>0$],

  i.e., $a\geq\max\left(\frac{h}{2(c+3)},\frac{h}{b(c+3)}\right)=\frac{h}{(c+3)\min(2,b)}$.
- (i): We require $N^a\left(\left[\frac{\log(N)}{N}\right]^{\frac{h}{c+3}}\right)^{\frac{b+1}{b}}\geq 1\Leftrightarrow \frac{\log^{\frac{h}{c+3}\frac{b+1}{b}}(N)}{N^{\frac{h}{c+3}\frac{b+1}{b}-a}}\geq 1$. Since $\frac{h}{c+3}\frac{b+1}{b}>0$, it is sufficient to have $\frac{h}{c+3}\frac{b+1}{b}-a\leq 0\Leftrightarrow\frac{h(b+1)}{(c+3)b}\leq a$.

To sum up, for (i)-(ii) we got $a\geq\max\left(\frac{h}{(c+3)\min(2,b)},\frac{h(b+1)}{(c+3)b}\right)$.

– (iii):
- (i): $\frac{h(b+1)}{(c+3)b}\leq a$.
- $\boxed{1}\to 0$: no constraint.
- $\boxed{1}\geq\boxed{2}$ [i.e., the first term dominates the second one in Eq. (35)]: $\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}\geq\frac{N^{\frac{h}{c+3}}}{N^{2a}\log^{\frac{h}{c+3}}(N)}\Leftrightarrow$ $\log^{\frac{hc}{c+3}+\frac{h}{c+3}}(N)\geq N^{\frac{hc}{c+3}+\frac{h}{c+3}-2a}$. Thus, since $\frac{h(c+1)}{c+3}>0$ we need $\frac{h(c+1)}{c+3}-2a\leq 0$, i.e., $\frac{h(c+1)}{2(c+3)}\leq a$.
- $\boxed{1}\geq\boxed{3}$ [i.e., the first term dominates the third one in Eq. (35)]: $\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}\geq\frac{N^{\frac{h}{b(c+3)}}}{N^a\log^{\frac{h}{b(c+3)}}(N)}\Leftrightarrow$ $\log^{\frac{hc}{c+3}+\frac{h}{b(c+3)}}(N)\geq N^{\frac{h}{b(c+3)}+\frac{hc}{c+3}-a}$. Since $\frac{hc}{c+3}+\frac{h}{b(c+3)}>0$ we require $\frac{h}{b(c+3)}+\frac{hc}{c+3}-a\leq 0$, i.e., $\frac{h}{b(c+3)}+\frac{hc}{c+3}\leq a$.

To sum up, the obtained condition for a is $\max\left(\frac{h}{b(c+3)}+\frac{hc}{c+3},\frac{h(c+1)}{2(c+3)}\right)=\frac{h\max\left(\frac{1}{b}+c,\frac{c+1}{2}\right)}{c+3}\leq a$. Since $\frac{1}{b}+c\geq\frac{c+1}{2}\Leftrightarrow\frac{1}{b}+\frac{c}{2}\geq\frac{1}{2}$[$\Leftarrow c\geq 1, b>0$], we got

$$\max\left(\frac{h\left(\frac{1}{b}+c\right)}{c+3},\frac{h(b+1)}{(c+3)b}\right)\leq a,\qquad r(N)=\left[\frac{\log(N)}{N}\right]^{\frac{hc}{c+3}}\to 0.$$

- $\boxed{1}=\boxed{3}$ in Eq. (34):

– (i)-(ii): Using in the λ choice that $\frac{h}{2}>0$, we obtain that

$$\frac{\log^h(l)}{N^h\lambda^3}=\frac{1}{l^2\lambda}\Leftrightarrow\frac{l^2\log^h(l)}{N^h}=\lambda^2\Leftrightarrow\frac{l\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}}=\lambda\to 0,\text{ if } a<\frac{h}{2}\text{ in } l=N^a.$$

$$r(l,N)=\left[\frac{l\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}}\right]^c+\frac{1}{l^2\frac{l\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}}}+\frac{1}{l\left[\frac{l\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}}\right]^{\frac{1}{b}}}.$$

$$r(N)=N^{ac-\frac{hc}{2}}\log^{\frac{hc}{2}}(N)+\frac{1}{N^{3a-\frac{h}{2}}\log^{\frac{h}{2}}(N)}+\frac{1}{N^{a+\frac{a}{b}-\frac{h}{2b}}\log^{\frac{h}{2b}}(N)}.$$

Here,

* (ii): $r(N)\to 0$ if
  · $\boxed{1}\to 0$: $ac-\frac{hc}{2}=c\left(a-\frac{h}{2}\right)<0$ [$\Leftarrow\frac{hc}{2}>0$], i.e., $a<\frac{h}{2}$ using that $c>0$.
  · $\boxed{2}\to 0$: $3a-\frac{h}{2}\geq 0$ [$\Leftarrow\frac{h}{2}>0$], i.e., $\frac{h}{6}\leq a$.
  · $\boxed{3}\to 0$: $a+\frac{a}{b}-\frac{h}{2b}\geq 0$ [$\Leftarrow\frac{h}{2b}>0$], i.e., $\frac{h}{2b\left(1+\frac{1}{b}\right)}=\frac{h}{2b\frac{b+1}{b}}=\frac{h}{2(b+1)}\leq a$ exploiting that $1+\frac{1}{b}>0$.

  In other words, the requirement is $\max\left(\frac{h}{6},\frac{h}{2(b+1)}\right)\leq a<\frac{h}{2}$.
* (i): $N^a\left[\frac{N^a\log^{\frac{h}{2}}(N)}{N^{\frac{h}{2}}}\right]^{\frac{b+1}{b}}\geq 1\Leftrightarrow\frac{\log^{\frac{h}{2}\frac{b+1}{b}}(N)}{N^{\frac{h}{2}\frac{b+1}{b}-a-a\frac{b+1}{b}}}\geq 1$. Since $\frac{h}{2}\frac{b+1}{b}>0$ it is enough to have $\frac{h}{2}\frac{b+1}{b}-a-a\frac{b+1}{b}\leq 0\Leftrightarrow\frac{h}{2}\frac{b+1}{b}\leq a\left(1+\frac{b+1}{b}\right)=a\frac{2b+1}{b}\Leftrightarrow\frac{h}{2}\frac{b+1}{2b+1}\leq a$ using that $2b+1>0$, $b>0$ [$\Leftarrow b>1$].

To sum up, for (i)-(ii) we obtained $\max\left(\frac{h}{6},\frac{h}{2(b+1)},\frac{h}{2}\frac{b+1}{2b+1}\right)\leq a<\frac{h}{2}$.

– (iii):
* (i): $\frac{h}{2}\frac{b+1}{2b+1}\leq a$
* $\boxed{2}\to 0$: $\frac{h}{6}\leq a$.
* $\boxed{2}\geq\boxed{1}$: $\frac{1}{N^{3a-\frac{h}{2}}\log^{\frac{h}{2}}(N)}\geq N^{ac-\frac{hc}{2}}\log^{\frac{hc}{2}}(N)\Leftrightarrow N^{\frac{h}{2}-3a+\frac{hc}{2}-ac}\geq\log^{\frac{h(c+1)}{2}}(N)$. Thus, since $\frac{h(c+1)}{2}>0$ we need $\frac{h}{2}-3a+\frac{hc}{2}-ac>0$, i.e., $\frac{h(c+1)}{2(c+3)}=\frac{h(c+3-2)}{2(c+3)}=\frac{h}{2}-\frac{h}{c+3}>a$, using that $c+3>0$.
* $\boxed{2}\geq\boxed{3}$: $\frac{1}{N^{3a-\frac{h}{2}}\log^{\frac{h}{2}}(N)}\geq\frac{1}{N^{a+\frac{a}{b}-\frac{h}{2b}}\log^{\frac{h}{2b}}(N)}\Leftrightarrow N^{a+\frac{a}{b}-\frac{h}{2b}+\frac{h}{2}-3a}\geq\log^{\frac{h}{2}-\frac{h}{2b}}(N)$. Since $\frac{h}{2}-\frac{h}{2b}=\frac{h}{2}\left(1-\frac{1}{b}\right)>0$ using that $h>0$ and $b>1$, we need $a+\frac{a}{b}-\frac{h}{2b}+\frac{h}{2}-3a>0$, i.e., $a\left(1+\frac{1}{b}-3\right)>\frac{h}{2}\left(\frac{1}{b}-1\right)\Leftrightarrow a\left(\frac{1}{b}-2\right)>\frac{h}{2}\left(\frac{1}{b}-1\right)$. Using that $b>1$, $0>\frac{1}{b}-1>\frac{1}{b}-2$; hence $a<\frac{\frac{h}{2}\left(\frac{1}{b}-1\right)}{\frac{1}{b}-2}$.

To sum up, we got

$$\max\left(\frac{h}{6},\frac{h}{2}\frac{b+1}{2b+1}\right)\leq a<\min\left(\frac{h}{2}-\frac{h}{c+3},\frac{\frac{h}{2}\left(\frac{1}{b}-1\right)}{\frac{1}{b}-2}\right)\qquad r(N)=\frac{1}{N^{3a-\frac{h}{2}}\log^{\frac{h}{2}}(N)}\to 0.$$

• $\boxed{1}=\boxed{4}$ in Eq. (34):

– (i)-(ii): Using in the λ choice that $\frac{b}{3b-1}>0$, we get

$$\frac{\log^h(l)}{N^h\lambda^3}=\frac{1}{l\lambda^{\frac{1}{b}}}\Leftrightarrow\frac{l\log^h(l)}{N^h}=\lambda^{3-\frac{1}{b}=\frac{3b-1}{b}}\Leftrightarrow\left[\frac{l\log^h(l)}{N^h}\right]^{\frac{b}{3b-1}}=\lambda\to 0,\text{ if } h>a\text{ in } l=N^a.$$

$$r(l,N)=\left[\frac{l\log^h(l)}{N^h}\right]^{\frac{bc}{3b-1}}+\frac{1}{l^2\left[\frac{l\log^h(l)}{N^h}\right]^{\frac{b}{3b-1}}}+\frac{1}{l\left[\frac{l\log^h(l)}{N^h}\right]^{\frac{1}{3b-1}}}.$$

$$r(N)=\left[\frac{\log^h(N)}{N^{h-a}}\right]^{\frac{bc}{3b-1}}+\frac{1}{N^{2a+\frac{ab}{3b-1}-\frac{hb}{3b-1}}\log^{\frac{hb}{3b-1}}(N)}+\frac{1}{N^{a+\frac{a}{3b-1}-\frac{h}{3b-1}}\log^{\frac{h}{3b-1}}(N)}.$$

Here,

* (ii): $r(N) \to 0$, if
  · $\boxed{1} \to 0$: $h - a > 0$ using that $h > 0$ and $\frac{bc}{3b-1} > 0$, i.e., $a < h$,
  · $\boxed{2} \to 0$: $2a + \frac{ab}{3b-1} - \frac{hb}{3b-1} \geq 0$ [using that $\frac{hb}{3b-1} > 0$]. In other words, $a\left(2 + \frac{b}{3b-1}\right) \geq \frac{hb}{3b-1} \Leftrightarrow a \geq \frac{\frac{hb}{3b-1}}{\left(2+\frac{b}{3b-1}\right)} = \frac{hb}{3b-1}\frac{3b-1}{6b-2+b} = \frac{hb}{7b-2}$ using that $\left(2 + \frac{b}{3b-1}\right) > 0$.
  · $\boxed{3} \to 0$: $a + \frac{a}{3b-1} - \frac{h}{3b-1} \geq 0$ [using that $\frac{h}{3b-1} > 0$], i.e., $a\left(1 + \frac{1}{3b-1}\right) \geq \frac{h}{3b-1} \Leftrightarrow a \geq \frac{\frac{h}{3b-1}}{1+\frac{1}{3b-1}} = \frac{h}{3b-1}\frac{3b-1}{3b-1+1} = \frac{h}{3b}$ making use of $\left(1 + \frac{1}{3b-1}\right) > 0$.

  Thus, we need $\max\left(\frac{hb}{7b-2}, \frac{h}{3b}\right) \leq a < h$.
* (i): $N^a\left(\left[\frac{N^a \log^h(N)}{N^h}\right]^{\frac{b}{3b-1}}\right)^{\frac{b+1}{b}} \geq 1 \Leftrightarrow \frac{\log^{\frac{h(b+1)}{3b-1}}(N)}{N^{\frac{h(b+1)}{3b-1} - a - a\frac{b+1}{3b-1}}} \leq 1$. Since $\frac{h(b+1)}{3b-1} > 0$, it is sufficient $\frac{h(b+1)}{3b-1} - a - a\frac{b+1}{3b-1} \leq 0 \Leftrightarrow \frac{h(b+1)}{3b-1} \leq a\left(1 + \frac{b+1}{3b-1}\right) = a\frac{3b-1+b+1}{3b-1} = a\frac{4b}{3b-1} \Leftrightarrow \frac{h(b+1)}{4b} \leq a$, where we used that $4b > 0$, $3b - 1 > 0$ [$\Leftarrow b > 1$].

To sum up, for (i)-(ii) we received $\max\left(\frac{hb}{7b-2}, \frac{h}{3b}, \frac{h(b+1)}{4b}\right) \leq a < h$.

– (iii):
* (i): $\frac{h(b+1)}{4b} \leq a$.
* $\boxed{3} \to 0$: $a \geq \frac{h}{3b}$.
* $\boxed{3} \geq \boxed{1}$: $\frac{1}{N^{a+\frac{a}{3b-1}-\frac{h}{3b-1}} \log^{\frac{h}{3b-1}}(N)} \geq \left[\frac{\log^h(N)}{N^{h-a}}\right]^{\frac{bc}{3b-1}} \Leftrightarrow N^{\frac{(h-a)bc}{3b-1} - a - \frac{a}{3b-1} + \frac{h}{3b-1}} \geq \log^{\frac{h(bc+1)}{3b-1}}(N)$.
  Since $\frac{h(bc+1)}{3b-1} > 0$, we need $\frac{(h-a)bc}{3b-1} - a - \frac{a}{3b-1} + \frac{h}{3b-1} > 0 \Leftrightarrow \frac{h(bc+1)}{3b-1} > a\left(\frac{bc}{3b-1} + 1 + \frac{1}{3b-1}\right) \Leftrightarrow \frac{h(bc+1)}{3b-1} > a\left(1 + \frac{bc+1}{3b-1}\right) \Leftrightarrow \frac{h(bc+1)}{3b-1} > a\frac{3b-1+bc+1}{3b-1} \Leftrightarrow \frac{h(bc+1)}{3b-1} > a\frac{3b+bc}{3b-1} \Leftrightarrow \frac{h(bc+1)}{3b+bc} > a$
  using at the last step that $3b - 1 > 0$ and $3b + bc > 0$.
* $\boxed{3} \geq \boxed{2}$: $\frac{1}{N^{a+\frac{a}{3b-1}-\frac{h}{3b-1}} \log^{\frac{h}{3b-1}}(N)} \geq \frac{1}{N^{2a+\frac{ab}{3b-1}-\frac{hb}{3b-1}} \log^{\frac{hb}{3b-1}}(N)} \Leftrightarrow$
  $\log^{\frac{h(b-1)}{3b-1}}(N) \geq N^{-2a-\frac{ab}{3b-1}+\frac{hb}{3b-1}+a+\frac{a}{3b-1}-\frac{h}{3b-1}}$. Since $\frac{h(b-1)}{3b-1} > 0$, we require that $-2a - \frac{ab}{3b-1} + \frac{hb}{3b-1} + a + \frac{a}{3b-1} - \frac{h}{3b-1} \leq 0 \Leftrightarrow \frac{h(b-1)}{3b-1} \leq a\left(1 + \frac{b-1}{3b-1}\right) \Leftrightarrow \frac{h(b-1)}{3b-1} \leq a\frac{3b-1+b-1}{3b-1} \Leftrightarrow \frac{h(b-1)}{4b-2} \leq a$ using that $3b - 1 > 0$ and $4b - 2 > 0$.

To sum up, we obtained that

$$\max\left(\frac{h(b-1)}{4b-2}, \frac{h}{3b}, \frac{h(b+1)}{4b}\right) \leq a < \frac{h(bc+1)}{3b+bc}, \qquad r(N) = \frac{1}{N^{a+\frac{a}{3b-1}-\frac{h}{3b-1}} \log^{\frac{h}{3b-1}}(N)} \to 0.$$

• $\boxed{2} = \boxed{3}$ in Eq. (34):
– (i)-(ii):

$$\lambda^c = \frac{1}{l^2\lambda} \Leftrightarrow \lambda^{c+1} = \frac{1}{l^2} \Leftrightarrow \lambda = \frac{1}{l^{\frac{2}{c+1}}} \to 0, \text{ if } l \to \infty. \quad [\Leftarrow \frac{2}{c+1} > 0]$$

$$r(l, N) = \frac{l^{\frac{6}{c+1}} \log^h(l)}{N^h} + \frac{1}{l^{\frac{2c}{c+1}}} + \frac{l^{\frac{2}{b(c+1)}}}{l} \Rightarrow r(N) = \frac{\log^h(N)}{N^{h-\frac{6a}{c+1}}} + \frac{1}{N^{\frac{2ac}{c+1}}} + \frac{1}{N^{a\left(1-\frac{2}{b(c+1)}\right)}}.$$

Here,
* (ii): $r(N) \to 0$ if
  · $\boxed{1} \to 0$: $h - \frac{6a}{c+1} > 0$ since $h > 0$, i.e., $a < \frac{h(c+1)}{6}$ using that $c + 1 > 0$.
  · $\boxed{2} \to 0$: $\frac{2ac}{c+1} > 0$ – this condition is satisfied by our assumptions ($a > 0$, $c > 0$).
  · $\boxed{3} \to 0$: $a\left(1 - \frac{2}{b(c+1)}\right) > 0$. Using that $a > 0$, $b > 0$, $c + 1 > 0$ this requirement is $1 > \frac{2}{b(c+1)} \Leftrightarrow b(c+1) > 2$[$\Leftarrow b > 1, c \geq 1$].

  Thus, we need $a < \frac{h(c+1)}{6}$.
* (i): $N^a\left(\frac{1}{N^{\frac{2a}{c+1}}}\right)^{\frac{b+1}{b}} \geq 1 \Leftrightarrow N^{a-\frac{2a(b+1)}{(c+1)b}} \geq 1$. Thus it is enough to satisfy $a - \frac{2a(b+1)}{(c+1)b} > 0 \Leftrightarrow 1 > \frac{2(b+1)}{(c+1)b}$, where we used that $a > 0$.

To sum up, for (i)-(ii) we obtained $a < \frac{h(c+1)}{6}$, $1 > \frac{2(b+1)}{(c+1)b}$.

- (iii):
  * (i): $1 > \frac{2(b+1)}{(c+1)b}$.
  * $\boxed{2} \to 0$: no constraint.
  * $\boxed{2} \geq \boxed{1}$: $\frac{1}{N^{\frac{2ac}{c+1}}} \geq \frac{\log^h(N)}{N^{h-\frac{6a}{c+1}}} \Leftrightarrow N^{h-\frac{6a}{c+1}-\frac{2ac}{c+1}} \geq \log^h(N)$. Thus, since $h > 0$ we require that $h - \frac{6a}{c+1} - \frac{2ac}{c+1} > 0 \Leftrightarrow h > a\frac{6+2c}{c+1} \Leftrightarrow \frac{h(c+1)}{6+2c} > a$, where the $6+2c > 0$, $c+1 > 0$ relations were exploited [$\Leftarrow c \geq 1$].
  * $\boxed{2} \geq \boxed{3}$: $\frac{1}{N^{\frac{2ac}{c+1}}} \geq \frac{1}{N^{a\left(1-\frac{2}{b(c+1)}\right)}} \Leftrightarrow N^{a\left(1-\frac{2}{b(c+1)}\right)-\frac{2ac}{c+1}} \geq 1$. Hence, by $a > 0$ and $c+1 > 0$ we need $a\left(1-\frac{2}{b(c+1)}\right) - \frac{2ac}{c+1} > 0 \Leftrightarrow a\frac{b(c+1)-2}{b(c+1)} > \frac{2ac}{c+1} \Leftrightarrow b(c+1)-2 > 2bc \Leftrightarrow b-2 > bc \Leftrightarrow -2 > b(c-1)$. Since $b > 0$ and $c \geq 1$, $b(c-1) \geq 0$; thus, this condition is never satisfied.

- $\boxed{2} = \boxed{4}$ in Eq. (34):
  - (i)-(ii):

$$\lambda^c = \frac{1}{l\lambda^{\frac{1}{b}}} \Leftrightarrow \lambda^{c+\frac{1}{b}=\frac{cb+1}{b}} = \frac{1}{l} \Leftrightarrow \lambda = \frac{1}{l^{\frac{b}{bc+1}}} \to 0, \text{ if } l \to \infty \quad [\Leftarrow \frac{b}{bc+1} > 0].$$

$$r(l,N) = \frac{l^{\frac{3b}{bc+1}}\log^h(l)}{N^h} + \frac{1}{l^{\frac{bc}{bc+1}}} + \frac{l^{\frac{b}{bc+1}}}{l^2} \Rightarrow r(N) = \frac{\log^h(N)}{N^{h-\frac{3ab}{bc+1}}} + \frac{1}{N^{\frac{abc}{bc+1}}} + \frac{1}{N^{2a-\frac{ab}{bc+1}}}.$$

  Here,
  * (ii): $r(N) \to 0$, if
    · $\boxed{1} \to 0$: Since $h > 0$ we get $h - \frac{3ab}{bc+1} > 0$, i.e., $\frac{h(bc+1)}{3b} > a$ using that $b > 0$, $bc+1 > 0$.
    · $\boxed{2} \to 0$: $\frac{abc}{bc+1} > 0$ – the second condition is satisfied by our assumptions ($a > 0$, $b > 0$, $c > 0$).
    · $\boxed{3} \to 0$: $2a - \frac{ab}{bc+1} > 0$. Making use of the positivity of a and $bc+1$, this requirement is equivalent to $2 > \frac{b}{bc+1} \Leftrightarrow 2bc+2 > b \Leftrightarrow 2 > b(1-2c)$, which holds since $b(1-2c) < 0$.

    Thus, we need $\frac{h(bc+1)}{3b} > a$.
  * (i): $N^a\left(\frac{1}{N^{\frac{ab}{bc+1}}}\right)^{\frac{b+1}{b}} \geq 1 \Leftrightarrow N^{a-\frac{a(b+1)}{bc+1}} \geq 1$. Thus it is sufficient to have $a - \frac{a(b+1)}{bc+1} > 0 \Leftrightarrow 1 > \frac{b+1}{bc+1}$, using $a > 0$.

  To sum up, for (i)-(ii) we got $\frac{h(bc+1)}{3b} > a$, $1 > \frac{b+1}{bc+1}$.
  - (iii):
  * (i): $1 > \frac{b+1}{bc+1}$.
  * $\boxed{2} \to 0$: no constraint.
  * $\boxed{2} \geq \boxed{1}$: $\frac{1}{N^{\frac{abc}{bc+1}}} \geq \frac{\log^h(N)}{N^{h-\frac{3ab}{bc+1}}} \Leftrightarrow N^{h-\frac{3ab}{bc+1}-\frac{abc}{bc+1}} \geq \log^h(N)$. Since $h > 0$, this holds if $h - \frac{3ab}{bc+1} - \frac{abc}{bc+1} > 0 \Leftrightarrow h > a\frac{3b+bc}{bc+1} \Leftrightarrow \frac{h(bc+1)}{3b+bc} > a$, exploiting that $3b+bc > 0$, $bc+1 > 0$.
  * $\boxed{2} \geq \boxed{3}$: $\frac{1}{N^{\frac{abc}{bc+1}}} \geq \frac{1}{N^{2a-\frac{ab}{bc+1}}} \Leftrightarrow N^{2a-\frac{ab}{bc+1}-\frac{abc}{bc+1}} \geq 1$. Hence, since $a > 0$ and $bc+1 > 0$ we have $2a - \frac{ab}{bc+1} - \frac{abc}{bc+1} > 0 \Leftrightarrow 2 > \frac{b+bc}{bc+1} \Leftrightarrow 2bc+2 > b+bc \Leftrightarrow bc+2 > b \Leftrightarrow 2 > b(1-c)$. This holds since $b(1-c) \leq 0$.

  Thus, we got

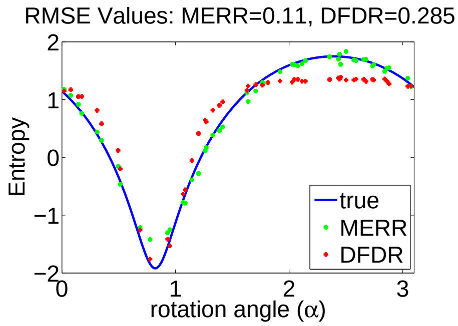
$$\frac{h(bc+1)}{3b+bc} > a, \qquad 1 > \frac{b+1}{bc+1} \qquad r(N) = \frac{1}{N^{\frac{abc}{bc+1}}} \to 0.$$

- $\boxed{3} = \boxed{4}$ in Eq. (34):
  - (i)-(ii):

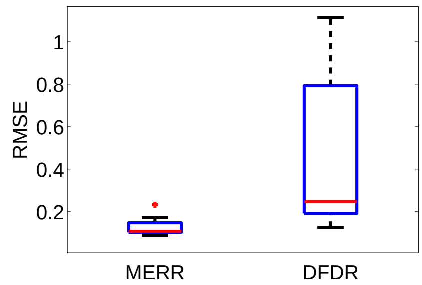
$$\frac{1}{l^2\lambda} = \frac{1}{l\lambda^{\frac{1}{b}}} \Leftrightarrow \frac{1}{l} = \lambda^{1-\frac{1}{b}=\frac{b-1}{b}} \Leftrightarrow \frac{1}{l^{\frac{b}{b-1}}} = \lambda \to 0, \text{ if } l \to \infty \quad [\Leftarrow \frac{b}{b-1} > 0].$$

$$r(l,N) = \frac{l^{\frac{3b}{b-1}}\log^h(l)}{N^h} + \frac{1}{l^{\frac{bc}{b-1}}} + \frac{l^{\frac{b}{b-1}}}{l^2} \Rightarrow r(N) = \frac{\log^h(N)}{N^{h-\frac{3ab}{b-1}}} + \frac{1}{N^{\frac{abc}{b-1}}} + \frac{1}{N^{2a-\frac{ab}{b-1}}}.$$

  Here,



(a) Entropy of Gaussian



(b) Boxplot of RMSE

Figure 1: (a) Learned entropy of a one-dimensional marginal distribution of a rotated 2d Gaussian. Axes x: rotation angle in $[0, \pi]$. Axis y: entropy. (b) RMSE values of the MERR and DFDR algorithms. Boxplots are calculated from 25 experiments.

* (ii): $r(N) \to 0$ if
  - $\boxed{1} \to 0$: Since $h > 0$ we get $h - \frac{3ab}{b-1} > 0$, i.e., $\frac{h(b-1)}{3b} > a$ using that $3b > 0$ and $b - 1 > 0$.
  - $\boxed{2} \to 0$: $\frac{abc}{b-1} > 0$. This requirement holds by our assumptions $[a > 0,\ b > 1,\ c > 0]$.
  - $\boxed{3} \to 0$: $2a - \frac{ab}{b-1} > 0$. By $a > 0$ and $b-1 > 0$, this constraint is $2 > \frac{b}{b-1} \Leftrightarrow 2b-2 > b \Leftrightarrow b > 2$.

  Hence, we need $\frac{h(b-1)}{3b} > a$, $b > 2$.
* (i): $N^a \left(\frac{1}{N^{\frac{ab}{b-1}}}\right)^{\frac{b+1}{b}} \geq 1 \Leftrightarrow N^{a - a\frac{b+1}{b-1}} \geq 1$. Thus we need $a - a\frac{b+1}{b-1} > 0 \Leftrightarrow 1 - 1\frac{b+1}{b-1} > 0 \Leftrightarrow 1 > \frac{b+1}{b-1}$, where we used that $a > 0$. The $1 > \frac{b+1}{b-1}$ is never satisfied since $\frac{b+1}{b-1} > 1$.

A.2 Numerical Experiments: Aerosol Prediction

In this section we provide numerical results to demonstrate the efficiency of the analysed ridge regression technique. The experiments serve to illustrate that the MERR approach compares favourably to

1. the only alternative, theoretically justified distribution regression method (since it avoids density estimation);[3] see Section A.2.1,
2. modern domain-specific, engineered methods (which beat state-of-the-art multiple instance learning alternatives); see Section A.2.2.

In our experiments we used the ITE toolbox (Information Theoretical Estimators; [43]).[11]

A.2.1 Supervised entropy learning

We compare our MERR (RKHS based mean embedding ridge regression) algorithm with [1]'s DFDR (kernel smoothing based distribution free distribution regression) method, on a benchmark problem taken from the latter paper. The goal is to learn the entropy of Gaussian distributions in a supervised way. We chose an $A \in \mathbb{R}^{2\times 2}$ matrix, whose A_{ij} entries were uniformly distributed on $[0, 1]$ ($A_{ij} \sim U[0, 1]$). We constructed 100 sample sets from $\{N(0, \Sigma_u)\}_{u=1}^{100}$, where $\Sigma_u = R(\beta_u)AA^T R(\beta_u)^T$ and $R(\beta_u)$ was a 2d rotation matrix with angle $\beta_u \sim U[0, \pi]$. From each $N(0, \Sigma_u)$ distribution we sampled 500 2-dimensional i.i.d. points. From the 100 sample sets, 25 were used for training, 25 for validation (i.e., selecting appropriate parameters), and 50 points for testing. Our goal is to learn the entropy of the first marginal distribution: $H = \frac{1}{2}\ln(2\pi e\sigma^2)$, where $\sigma^2 = M_{1,1}$, $M = \Sigma_u \in \mathbb{R}^{2\times 2}$. Figure 1(a) displays the learned entropies of the 50 test sample sets in a typical experiment. We compare the results of DFDR and MERR. One can see that the true and the estimated values are close to each other for both algorithms, but MERR performs better. The boxplot diagrams of the RMSE (root mean square error)

[11] The ITE toolbox contains the MERR method and its numerical demonstrations (among others); see `https://bitbucket.org/szzoli/ite/`.

Table 4: Prediction accuracy of the MERR method in AOD prediction using different kernels: $100 \times RMSE(\pm std)$. K: linear. The best single and ensemble predictions are written in bold.

k_G	k_e	k_C	k_t	$k_p(p=2)$	$k_p(p=3)$
7.97 ($\pm$1.81)	8.25 ($\pm$1.92)	7.92 ($\pm$1.69)	8.73 ($\pm$2.18)	12.5 ($\pm$2.63)	171.24 ($\pm$56.66)
k_r	k_i	$k_{M,\frac{3}{2}}$	$k_{M,\frac{5}{2}}$	ensemble	
9.66 ($\pm$2.68)	**7.91** (**$\pm$1.61**)	8.05 ($\pm$1.83)	7.98 ($\pm$1.75)	**7.86** (**$\pm$1.71**)	

values calculated from 25 experiments confirm this performance advantage (Figure 1(b)). A reason why MERR achieves better performance is that DFDR needs to do many density estimations, which can be very challenging when the sample sizes are small. By contrast, the MERR algorithm does not require density estimation.

A.2.2 Aerosol prediction

Aerosol prediction is one of the largest challenges of current climate research; we chose this problem as a further testbed of our method. [35] pose the AOD (aerosol optical depth) prediction problem as a MIL task: (i) a given pixel of a multispectral image corresponds to a small area of $200 \times 200m^2$, (ii) spatial variability of AOD can be considered to be small over distances up to $100km$, (iii) ground-based instruments provide AOD labels ($y_i \in \mathbb{R}$), (iv) a bag consists of randomly selected pixels within a $20km$ radius around an AOD sensor. The MIL task can be tackled using our MERR approach, assuming that (i) bags correspond to distributions (x_i), (ii) instances in the bag ($\{x_{i,n}\}_{n=1}^N$) are samples from the distribution.

We selected the MISR1 dataset [35], where (i) cloudy pixels are also included, (ii) there are 800 bags with (iii) 100 instances in each bag, (iv) the instances are 16-dimensional ($x_{i,n} \in \mathbb{R}^{16}$). Our baselines are the reported state-of-the-art EM (expectation-maximization) methods achieving average $100 \times RMSE = 7.5 - 8.5$ $(\pm 0.1 - 0.6)$ accuracy. The experimental protocol followed the original work, where 5-fold cross-validation (4×160 (160) samples for training (testing)) was repeated 10 times; the only difference is that we made the problem a bit harder, as we used only 3×160 samples for training, 160 for validation (i.e., setting the λ regularization and the θ kernel parameter), and 160 for testing.

- Linear K: In the first set of experiments, K was linear. To study the robustness of our method, we picked 10 different kernels (k) and their ensembles: the Gaussian, exponential, Cauchy, generalized t-student, polynomial kernel of order 2 and 3 ($p = 2$ and 3), rational quadratic, inverse multiquadratic kernel, Matérn kernel (with $\frac{3}{2}$ and $\frac{5}{2}$ smoothness parameters). The expressions for these kernels are

$$k_G(a,b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \qquad k_e(a,b) = e^{-\frac{\|a-b\|_2}{2\theta^2}}, \qquad k_C(a,b) = \frac{1}{1+\frac{\|a-b\|_2^2}{\theta^2}},$$

$$k_t(a,b) = \frac{1}{1+\|a-b\|_2^\theta}, \qquad k_p(a,b) = (\langle a,b\rangle + \theta)^p, \qquad k_r(a,b) = 1 - \frac{\|a-b\|_2^2}{\|a-b\|_2^2+\theta},$$

$$k_i(a,b) = \frac{1}{\sqrt{\|a-b\|_2^2+\theta^2}}, \quad k_{M,\frac{3}{2}}(a,b) = \left(1+\frac{\sqrt{3}\,\|a-b\|_2}{\theta}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\theta}}$$

$$k_{M,\frac{5}{2}}(a,b) = \left(1+\frac{\sqrt{5}\,\|a-b\|_2}{\theta}+\frac{5\,\|a-b\|_2^2}{3\theta^2}\right) e^{-\frac{\sqrt{5}\|a-b\|_2}{\theta}},$$

  where $p = 2,3$ and $\theta > 0$. The explored parameter domain was $(\lambda,\theta) \in \{2^{-65}, 2^{-64}, \ldots, 2^{-3}\} \times \{2^{-15}, 2^{-14}, \ldots, 2^{10}\}$; increasing the domain further did not improve the results.

  Our results are summarized in Table 4. According to the table, we achieve $100 \times RMSE = 7.91$ (± 1.61) using a single kernel, or 7.86 ($\pm$1.71) with ensemble of kernels (further performance improvements might be obtained by learning the weights).

- Nonlinear K: We also studied the efficiency of nonlinear K-s. In this case, the argument of K was $\|\mu_a - \mu_b\|_H$ instead of $\|a-b\|_2$ (see the definition of k-s); for K examples, see Table 1. Our obtained

Table 5: Prediction accuracy of the MERR method in AOD prediction using different kernels: $100 \times RMSE(\pm std)$; single prediction case. K: nonlinear. Rows: kernel k. Columns: kernel K. For each row (k), the smallest RMSE value is written in bold.

$k\backslash K$	K_G	K_e	K_C	K_t	$K_{M,\frac{3}{2}}$
k_e	8.14 (±1.80)	8.10 (±1.81)	8.14 (±1.81)	**8.07 (±1.77)**	8.09 (±1.88)
k_C	7.97 (±1.58)	8.13 (±1.79)	7.96 (±1.62)	8.09 (±1.69)	**7.90 (±1.63)**
$k_{M,\frac{3}{2}}$	8.00 (±1.66)	8.14 (±1.80)	8.00 (±1.69)	8.08 (±1.72)	**7.96 (±1.69)**
k_i	8.01 (±1.53)	8.17 (±1.74)	8.03 (±1.63)	7.93 (±1.57)	8.04 (±1.67)

$k\backslash K$	$K_{M,\frac{5}{2}}$	K_r	K_i	linear
k_e	8.14 (±1.78)	8.12 (±1.81)	8.12 (±1.80)	8.25 (±1.92)
k_C	7.95 (±1.60)	7.92 (±1.61)	7.93 (±1.61)	7.92 (±1.69)
$k_{M,\frac{3}{2}}$	8.02 (±1.71)	8.04 (±1.69)	7.98 (±1.72)	8.05 (±1.83)
k_i	8.05 (±1.61)	8.05 (±1.63)	8.06 (±1.65)	**7.91(±1.61)**

results are summarized in Table 5. One can see that using nonlinear K kernels, the RMSE error drops to 7.90 (±1.63) in the single prediction case, and decreases further to 7.81 (±1.64) in the ensemble setting.

Despite the fact that MERR has no domain-specific knowledge wired in, the results fall within the same range as [35]'s algorithms. The prediction is fairly precise and robust to the choice of the kernel, however polynomial kernels perform poorly (they violate our boundedness assumption).