Journal of Machine Learning Research 17 (2016) 1-40 Submitted 12/14; Revised 1/16; Published 9/16

Learning Theory for Distribution Regression

Zoltán Szabó* ZOLTAN.SZABO@GATSBY.UCL.AC.UK
ORCID 0000-0001-6183-7603
Gatsby Unit, University College London
Sainsbury Wellcome Centre, 25 Howland Street
London - W1T 4JG, UK

Bharath K. Sriperumbudur BKS18@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Machine Learning Department
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue Pittsburgh PA 15213 USA

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
ORCID 0000-0003-3169-7624
Gatsby Unit, University College London
Sainsbury Wellcome Centre, 25 Howland Street
London - W1T 4JG, UK

Editor: Ingo Steinwart

Abstract

We focus on the distribution regression problem: regressing to vector-valued outputs from probability measures. Many important machine learning and statistical tasks fit into this framework, including multi-instance learning and point estimation problems without analytical solution (such as hyperparameter or entropy estimation). Despite the large number of available heuristics in the literature, the inherent two-stage sampled nature of the problem makes the theoretical analysis quite challenging, since in practice only samples from sampled distributions are observable, and the estimates have to rely on similarities computed between sets of points. To the best of our knowledge, the only existing technique with consistency guarantees for distribution regression requires kernel density estimation as an intermediate step (which often performs poorly in practice), and the domain of the distributions to be compact Euclidean. In this paper, we study a simple, analytically computable, ridge regression-based alternative to distribution regression, where we embed the distributions to a reproducing kernel Hilbert space, and learn the regressor from the embeddings to the outputs. Our main contribution is to prove that this scheme is consistent in the two-stage sampled setup under mild conditions (on separable topological domains enriched with kernels): we present an exact computational-statistical efficiency trade-off analysis showing that our estimator is able to match the *one-stage* sampled minimax op-

∗. Now at Applied Mathematics Department, Center for Applied Mathematics, École Polytechnique, University of Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France.

timal rate (Caponnetto and De Vito, 2007; Steinwart et al., 2009). This result answers a 17-year-old open question, establishing the consistency of the classical set kernel (Haussler, 1999; Gärtner et al., 2002) in regression. We also cover consistency for more recent kernels on distributions, including those due to Christmann and Steinwart (2010).

Keywords: Two-Stage Sampled Distribution Regression, Kernel Ridge Regression, Mean Embedding, Multi-Instance Learning, Minimax Optimality

1. Introduction

We address the learning problem of *distribution regression* in the two-stage sampled setting, where we only have bags of samples from the probability distributions: we regress from probability measures to real-valued (Póczos et al., 2013) responses, or more generally to vector-valued outputs (belonging to an arbitrary separable Hilbert space). Many classical problems in machine learning and statistics can be analysed in this framework. On the machine learning side, multiple instance learning (Dietterich et al., 1997; Ray and Page, 2001; Dooly et al., 2002) can be thought of in this way, 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 without closed form analytical expressions (e.g., its entropy or a hyperparameter).

Intuitive description of our goal: Let us start with a somewhat informal definition of the distribution regression problem and an intuitive phrasing of our goals. Suppose that our data consist of $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l}$, where x_i is a probability distribution, y_i is its label (in the simplest case $y_i \in \mathbb{R}$ or $y_i \in \mathbb{R}^d$) 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 $x_{i,1}, \dots, x_{i,N_i} \overset{i.i.d.}{\sim} x_i$. Thus the observed data are $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^{N_i}, y_i)\}_{i=1}^{l}$. Since $x_{i,j}$ is sampled from x_i, and x_i is sampled from $\mathcal{M}$, we call this process two-stage sampling. Our goal is to predict a new y_{l+1} from a new batch of samples $x_{l+1,1}, \dots, x_{l+1,N_{l+1}}$ drawn from a new distribution $x_{l+1} \sim \mathcal{M}$. For example, in a medical application, the i^{th} patient might be identified with a probability distribution (x_i), which can be periodically assessed 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 $(\hat{\mathbf{z}})$, we might wish 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". Briefly, we consider the following questions:

Can the distribution regression problem be solved consistently under mild conditions? What is the exact computational-statistical efficiency trade-off implied by the two-stage sampling?

In our work the estimated mapping $(\hat{f})$ is the analytical solution of a kernel ridge regression (KRR) problem.[1] The performance of $\hat{f}$ depends on the assumed function class $(\mathcal{H})$, the

1. Beyond its simple analytical formula, kernel ridge regression also allows efficient distributed (Zhang et al., 2015; Richtárik and Takáč, 2016), sketch (Alaoui and Mahoney, 2015; Yang et al., 2016) and Nyström based approximations (Rudi et al., 2015).

family of $\hat{f}$ candidates used in the ridge formulation. We shall focus on the analysis of two settings:

1. **Well-specified case ($f_* \in \mathcal{H}$):** In this case we assume that the regression function f_* belongs to $\mathcal{H}$. We focus on bounding the goodness of $\hat{f}$ compared to f_*. In other words, if $\mathcal{R}[f_*]$ denotes the prediction error (expected risk) of f_*, then our goal is to derive a finite-sample bound for the excess risk, $\mathcal{E}(\hat{f}, f_*) = \mathcal{R}[\hat{f}] - \mathcal{R}[f_*]$ that holds with high probability. We make use of this bound to establish the consistency of the estimator (i.e., drive the excess risk to zero) and to derive the exact computational-statistical efficiency trade-off of the estimator as a function of the sample number (l, $N = N_i$, $\forall\, i$) and the problem difficulty (see Theorem 5 and its corresponding remarks for more details).

2. **Misspecified case ($f_* \in L^2 \backslash \mathcal{H}$):** Since in practise it might be hard to check whether $f_* \in \mathcal{H}$, we also study the misspecified setting of $f_* \in L^2$; the relevant case is when $f_* \in L^2 \backslash \mathcal{H}$. In the misspecified setting the 'richness' of $\mathcal{H}$ has crucial importance, in other words the size of $D^2_{\mathcal{H}} = \inf_{f \in \mathcal{H}} \|f_* - f\|_2^2$, the approximation error from $\mathcal{H}$. Our main contributions consist of proving a finite-sample excess risk bound, using which we show that the proposed estimator can attain the ideal performance, i.e., $\mathcal{E}(\hat{f}, f_*) - D^2_{\mathcal{H}}$ can be driven to zero. Moreover, on smooth classes of f_*-s, we give a simple and explicit description for the computational-statistical efficiency trade-off of our estimator (see Theorem 9 and its corresponding remarks for more details).

There exist a vast number of heuristics to tackle learning problems on distributions; we will review them in Section 5. However, to the best of our knowledge, the only prior work addressing the *consistency* of regression on distributions requires kernel density estimation (Póczos et al., 2013; Oliva et al., 2014; Sutherland et al., 2016), which assumes that the response variable is scalar-valued,[2] 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 have a density) and then constructed a kernel regressor that acts on these kernel density estimates.[3] Using the classical bias-variance decomposition analysis for kernel regressors, they showed the consistency of the constructed kernel regressor, and provided 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 (Kpotufe, 2011).[4]

One can define kernel learning algorithms on bags based on set kernels (Gärtner et al., 2002) by computing the similarity of the sets/bags of samples representing the input distributions; set kernels are also called called multi-instance kernels or ensemble kernels, and are examples of convolution kernels (Haussler, 1999). In this case, the similarity of two sets

2. Oliva et al. (2013, 2015) consider the case where the responses are also distributions or functions.
3. We would like to clarify that the kernels used in their work are classical smoothing kernels—extensively studied in non-parametric statistics (Györfi et al., 2002)—and not the reproducing kernels that appear throughout our paper.
4. Using a random kitchen sinks approach, with orthonormal basis projection estimators Oliva et al. (2014); Sutherland et al. (2016) propose distribution regression algorithms that can computationally handle large scale datasets; as with Póczos et al. (2013), these approaches are based on density estimation in $\mathbb{R}^d$.

is measured by the average pairwise point similarities between the sets. From a theoretical perspective, nothing is known about the consistency of set kernel based learning method since their introduction in 1999 (Haussler, 1999; Gärtner et al., 2002): 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, are 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; Berlinet and Thomas-Agnan, 2004). The population limits are well-defined as being dot products between the embeddings of the generating distributions (Altun and Smola, 2006), and for characteristic kernels the distance between embeddings defines a metric on probability measures (Sriperumbudur et al., 2011; Gretton et al., 2012). When bounded kernels are used, mean embeddings exist for all probability measures (Fukumizu et al., 2004). 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 Christmann and Steinwart (2010) 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 was also applied by Muandet et al. (2012) for classification on distribution-valued inputs, however consistency was not studied in that work. We also note that motivated by the current paper, Lopez-Paz et al. (2015) have recently presented the first theoretical results about surrogate risk guarantees on a class (relying on uniformly bounded Lipschitz functionals) of soft distribution-classification problems.

Our **contribution** in this paper is to establish the learning theory of a simple, mean embedding based ridge regression (MERR) method for the distribution regression problem. This result applies both to the basic set kernels of Haussler (1999); Gärtner et al. (2002), the distribution kernels of Christmann and Steinwart (2010), and additional related kernels. We provide finite-sample excess risk bounds, prove consistency, and show how the two-stage sampled nature of the problem (bag size) governs the computational-statistical efficiency of the MERR estimator. More specifically, in the

1. **well-specified case:** We
   (a) derive finite-sample bounds on the excess risk: We construct $\mathcal{R}[\hat{f}] - \mathcal{R}[f_*] \leq r(l, N, \lambda)$ bounds holding with high probability, where λ is the regularization parameter in the ridge problem ($\lambda \to 0$, $l \to \infty$, $N = N_i \to \infty$).
   (b) establish consistency and computational-statistical efficiency trade-off of the MERR estimator on a general prior family $\mathcal{P}(b, c)$ as defined by Caponnetto and De Vito (2007), where b captures the effective input dimension, and larger c means smoother f_* ($1 < b$, $c \in (1, 2]$). In particular, when the number of samples per bag is chosen as $N = l^a \log(l)$ and $a \geq \frac{b(c+1)}{bc+1}$, then the learning rate saturates at $l^{-\frac{bc}{bc+1}}$, which is known to be one-stage sampled minimax optimal (Caponnetto and De Vito, 2007). In other words, by choosing $a = \frac{b(c+1)}{bc+1} < 2$, we suffer *no loss in statistical performance* compared with the *best possible one-stage sampled estimator*.

   Note: the advantage of considering the $\mathcal{P}(b, c)$ family is two-fold. It does not assume parametric distributions, yet certain complexity terms can be explicitly upper bounded in the family. This property will be exploited in our analysis. Moreover, (for special

input distributions) the parameter b can be related to the spectral decay of Gaussian Gram matrices, and existing analysis techniques (Steinwart and Christmann, 2008) may be used in interpreting these decay conditions.

2. **misspecified case:** We establish consistency and convergence rates even if $f_* \notin \mathcal{H}$. Particularly, by deriving finite-sample bounds on the excess risk we
   (a) prove that the MERR estimator can achieve the best possible approximation accuracy from $\mathcal{H}$, i.e. the $\mathcal{R}[\hat{f}] - \mathcal{R}[f_*] - D_{\mathcal{H}}^2$ quantity can be driven to zero (recall that $D_{\mathcal{H}} = \inf_{f \in \mathcal{H}} \|f_* - f\|_2$). Specifically, this result implies that if $\mathcal{H}$ is dense in L^2 ($D_{\mathcal{H}} = 0$), then the excess risk $\mathcal{R}[\hat{f}] - \mathcal{R}[f_*]$ converges to zero.
   (b) analyse the computational-statistical efficiency trade-off: We show that by choosing the bag size as $N = l^{2a} \log(l)$ ($a > 0$) one can get rate $l^{-\frac{2sa}{s+1}}$ for $a \leq \frac{s+1}{s+2}$, and the rate saturates for $a \geq \frac{s+1}{s+2}$ at $l^{-\frac{2s}{s+2}}$, where the difficulty of the regression problem is captured by $s \in (0, 1]$ (a larger s means an easier problem). This means that easier tasks give rise to faster convergence (for $s = 1$, the rate is $l^{-\frac{2}{3}}$), the bag size N can again be *sub-quadratic* in l ($2a \leq \frac{2(s+1)}{s+2} \leq \frac{4}{3} < 2$), and the rate at saturation is close to $\tilde{r}(l) = l^{-\frac{2s}{2s+1}}$, which is the asymptotically optimal rate in the one-stage sampled setup, with real-valued output and stricter eigenvalue decay conditions (Steinwart et al., 2009).

Due to the differences in the assumptions made and the loss function used, a direct comparison of our theoretical result and that of Póczos et al. (2013) remains an open question, however we make three observations. First, our approach is more general, since we may regress from any probability measure defined on separable, topological domains endowed with kernels. 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, graphs, and other structured objects are disallowed. Second, in our analysis we will allow separable Hilbert space valued outputs, in contrast to the real-valued output considered by Póczos et al. (2013). Third, density estimates in high dimensional spaces suffer from slow convergence rates (Wasserman, 2006, Section 6.5). Our approach mitigates this problem, as it works directly on distribution embeddings, and does not make use of density estimation as an intermediate step.

The principal challenge in proving theoretical guarantees arises from the two-stage sampled nature of the inputs. In our analysis of the well-specified case, we make use of Caponnetto and De Vito (2007)'s results, which focus (only) on 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 objective terms. Even the verification of these conditions requires care since the inputs in the ridge regression are themselves distribution embeddings (i.e., functions in a reproducing kernel Hilbert space).

In the misspecified case, RKHS methods alone are not sufficient to obtain excess risk bounds: one has to take into account the "richness" of the modelling RKHS class ($\mathcal{H}$) in the embedding L^2 space. The fundamental challenge is whether it is possible to achieve the best possible performance dictated by $\mathcal{H}$; or in the special case when further smoothness conditions hold on f_*, what convergence rates can yet be attained, and what computational-statistical efficiency trade-off realized. The second smoothness property could be modelled

for example by range spaces of (fractional) powers of integral operators associated to $\mathcal{H}$. Indeed, there exist several results along these lines with KRR for the case of real-valued outputs: see for example (Sun and Wu, 2009a, Theorem 1.1), (Sun and Wu, 2009b, Corollary 3.2), (Mendelson and Neeman, 2010, Theorem 3.7 with Assumption 3.2). The question of optimal rates has also been addressed for the semi-supervised KRR setting (Caponnetto, 2006, Theorem 1), and for clipped KRR estimators (Steinwart et al., 2009) with integral operators of rapidly decaying spectrum. Our results apply more generally to the two-stage sampled setting and to vector valued outputs belonging to separable Hilbert spaces. Moreover, we obtain a general consistency result without range space assumptions, showing that the modelling power of $\mathcal{H}$ can be fully exploited, and convergence to the best approximation available from $\mathcal{H}$ can be realized.[5]

There are numerous areas in machine learning and statistics, where estimating vector-valued functions has crucial importance. Often in statistics, one is not only confronted with the estimation of a scalar parameter, but with a vector of parameters. On the machine learning side, multi-task learning (Evgeniou et al., 2005), functional response regression (Kadri et al., 2016), or structured output prediction (Brouard et al., 2011; Kadri et al., 2013) fall under the same umbrella: they can be naturally phrased as learning vector-valued functions (Micchelli and Pontil, 2005). The idea underlying all these tasks is simple and intuitive: if multiple prediction problems have to be solved simultaneously, it might be beneficial to exploit their dependencies. Imagine for example that the task is to predict the motion of a dancer: taking into account the interrelation of the actor's body parts is likely to lead to more accurate estimation, as opposed to predicting the individual parts one by one, independently. Successful real-world applications of a multi-task approach include for example preference modelling of users with similar demographics (Evgeniou et al., 2005), prediction of the daily precipitation profiles of weather stations (Kadri et al., 2010), acoustic-to-articulatory speech inversion (Kadri et al., 2016), identifying biomarkers capable of tracking the progress of Alzheimer's disease (Zhou et al., 2013), personalized human activity recognition based on iPod/iPhone accelerometer data (Sun et al., 2013), finger trajectory prediction in brain-computer interfaces (Kadri et al., 2012) or ecological inference (Flaxman et al., 2015); for a recent review on multi-output prediction methods see (Álvarez et al., 2011; Borchani et al., 2015). A mathematically sound way of encoding prior information about the relation of the outputs can be realized by operator-valued kernels and the associated vector-valued RKHS-s (Pedrick, 1957; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010); this is the tool we use to allow vector-valued learning tasks.

Finally, we note that the current work extends our earlier conference paper (Szabó et al., 2015) in several important respects: we now show that the MERR method can attain the one-stage sampled minimax optimal rate; we generalize the analysis in the well-specified setting to allow outputs belonging to an arbitrary separable Hilbert spaces (in contrast to the original scalar-valued output domain); and we tackle the misspecified setting, obtaining finite sample guarantees, consistency, and computational-statistical efficiency trade-offs.

The paper is structured as follows: The distribution regression problem and the MERR technique are introduced in Section 2. Our assumptions are detailed in Section 3. We present our theoretical guarantees (finite-sample bounds on the excess risk, consistency,

5. Specializing our result, we get explicit rates and an exact computational-statistical efficiency description for MERR as a function of sample numbers and problem difficulty, for smooth regression functions.

computational-statistical efficiency trade-offs) in Section 4: the well-specified case is considered in Section 4.1, and the misspecified setting is the focus of Section 4.2. Section 5 is devoted to an overview of existing heuristics for learning on distributions. Conclusions are drawn in Section 6. Section 7 contains proof details. In Section 8 we discuss our assumptions with concrete examples.

2. The Distribution Regression Problem

Below we first introduce our notation (Section 2.1), then formally define the distribution regression task (Section 2.2).

2.1 Notation

We use the following notations throughout the paper:

- **Sets, topology, measure theory:** Let $\mathcal{K}$ be a Hilbert space; $cl\left[V\right]$ is the closure of a set $V \subseteq \mathcal{K}$. $\times_{i\in I} S_i$ is the direct product of sets S_i. $f \circ g$ is the composition of function f and g. 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. If $(\mathcal{X}, d)$ is a metric space, then $\mathcal{B} = \mathcal{B}(d)$ is the Borel σ-algebra generated by the open sets induced by metric d. $\mathcal{M}_1^+(\mathcal{X})$ denotes the set of Borel probability measures on the $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ measurable space. Given measurable spaces $(U_1, \mathcal{S}_1)$ and $(U_2, \mathcal{S}_2)$, the $\mathcal{S}_1 \otimes \mathcal{S}_2$ product σ-algebra (Steinwart and Christmann, 2008, page 480) on the product space $U_1 \times U_2$ is the σ-algebra generated by the cylinder sets $U_1 \times S_2$, $S_1 \times U_2$ ($S_1 \in \mathcal{S}_1$, $S_2 \in \mathcal{S}_2$). 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 $h \in C_b(\mathcal{X}) = \{(\mathcal{X}, \tau) \to \mathbb{R} \text{ bounded, continuous functions}\}$.
- **Functional analysis:** Let $(N_1, \|\cdot\|_{N_1})$ and $(N_2, \|\cdot\|_{N_2})$ denote two normed spaces, then $\mathcal{L}(N_1, N_2)$ stands for the space of $N_1 \to N_2$ bounded linear operators; if $N_1 = N_2$, we will use the $\mathcal{L}(N_1) = \mathcal{L}(N_1, N_2)$ shorthand. For $M \in \mathcal{L}(N_1, N_2)$ the operator norm is defined as $\|M\|_{\mathcal{L}(N_1,N_2)} = \sup_{0\neq h\in N_1} \|Mh\|_{N_2} / \|h\|_{N_1}$, $Im(M) = \{Mn_1\}_{n_1\in N_1}$ denotes the range of M, $Ker(M) = \{n_1 \in N_1 : Mn_1 = 0\}$ is the null space of M. Let $\mathcal{K}$ be a Hilbert space. The adjoint operator $M^* \in \mathcal{L}(\mathcal{K})$ of an operator $M \in \mathcal{L}(\mathcal{K})$ is the operator such that $\langle Ma, b\rangle_{\mathcal{K}} = \langle a, M^*b\rangle_{\mathcal{K}}$ for all a and b in $\mathcal{K}$. $M \in \mathcal{L}(\mathcal{K})$ is called positive if $\langle Ma, a\rangle_{\mathcal{K}} \geq 0$ ($\forall a \in \mathcal{K}$), self-adjoint if $M = M^*$, and trace class if $\sum_{j\in J} \langle |M|e_j, e_j\rangle_{\mathcal{K}} < \infty$ for an $(e_j)_{j\in J}$ ONB (orthonormal basis) of $\mathcal{K}$ ($|M| := (M^*M)^{\frac{1}{2}}$), in which case $Tr(M) := \sum_{j\in J} \langle Me_j, e_j\rangle_{\mathcal{K}} < \infty$; compact if $cl\left[Ma : a \in \mathcal{K}, \|a\|_{\mathcal{K}} \leq 1\right]$ is a compact set. Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be Hilbert spaces. $M \in \mathcal{L}(\mathcal{K}_1, \mathcal{K}_2)$ is called Hilbert-Schmidt if $\|M\|^2_{\mathcal{L}_2(\mathcal{K}_1,\mathcal{K}_2)} = Tr(M^*M) = \sum_{j\in J} \langle Me_j, Me_j\rangle_{\mathcal{K}_2} < \infty$ for some $(e_j)_{j\in J}$ ONB of $\mathcal{K}_1$. The space of Hilbert-Schmidt operators is denoted by $\mathcal{L}_2(\mathcal{K}_1, \mathcal{K}_2) = \{M \in \mathcal{L}(\mathcal{K}_1, \mathcal{K}_2) : \|M\|_{\mathcal{L}_2(\mathcal{K}_1,\mathcal{K}_2)} < \infty\}$. We use the shorthand notation $\mathcal{L}_2(\mathcal{K}) = \mathcal{L}_2(\mathcal{K}, \mathcal{K})$ if $\mathcal{K} := \mathcal{K}_1 = \mathcal{K}_2$; $\mathcal{L}_2(\mathcal{K})$ is separable if and only if $\mathcal{K}$ is separable (Steinwart and Christmann, 2008, page 506). Trace class and Hilbert-Schmidt operators over a $\mathcal{K}$ Hilbert space are compact operators (Steinwart and Christmann, 2008, page 505-506); moreover,

$$\|A\|_{\mathcal{L}(\mathcal{K})} \leq \|A\|_{\mathcal{L}_2(\mathcal{K})}, \quad \forall A \in \mathcal{L}_2(\mathcal{K}), \tag{1}$$

$$\|AB\|_{\mathcal{L}_2(\mathcal{K})} \leq \|A\|_{\mathcal{L}_2(\mathcal{K})} \|B\|_{\mathcal{L}(\mathcal{K})}, \quad \forall A, B \in \mathcal{L}_2(\mathcal{K}). \tag{2}$$

I is the identity operator; $I_l \in \mathbb{R}^{l\times l}$ is the identity matrix.

- **RKHS, mean embedding:** Let $H = H(k)$ be an RKHS (Steinwart and Christmann, 2008) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel. Denote by

$$X = \mu\left(\mathcal{M}_1^+(\mathcal{X})\right) = \{\mu_x : x \in \mathcal{M}_1^+(\mathcal{X})\} \subseteq H, \quad \mu_x = \int_{\mathcal{X}} k(\cdot, u)\mathrm{d}x(u) = \mathbb{E}_{u\sim x}[k(\cdot, u)] \in H$$

the set of mean embeddings (Berlinet and Thomas-Agnan, 2004) of the distributions to the space H.[6] Let Y be a separable Hilbert space, where the inner product is denoted by $\langle\cdot,\cdot\rangle_Y$; the associated norm is $\|\cdot\|_Y$. $\mathcal{H} = \mathcal{H}(K)$ is the Y-valued RKHS (Pedrick, 1957; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010) of $X \to Y$ functions with $K : X \times X \to \mathcal{L}(Y)$ as the reproducing kernel (we will present some concrete examples of K in Section 3; see Table 1); $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$ is defined as

$$K(\mu_x, \mu_t)(y) = (K_{\mu_t} y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ or } K(\cdot, \mu_t)(y) = K_{\mu_t} y. \tag{3}$$

Further, $f(\mu_x) = K_{\mu_x}^* f$ $(\forall \mu_x \in X, f \in \mathcal{H})$.

- **Regression function:** Let ρ be the μ-induced probability measure on the $Z = X \times Y$ product space, and let $\rho(\mu_x, y) = \rho(y|\mu_x)\rho_X(\mu_x)$ be the factorization of ρ into conditional and marginal distributions.[7] The regression function of ρ with respect to the (μ_x, y) pair is denoted by

$$f_\rho(\mu_a) = \int_Y y\,\mathrm{d}\rho(y|\mu_a) \quad (\mu_a \in X) \tag{4}$$

and for $f \in L^2_{\rho_X}$ let $\|f\|_\rho = \sqrt{\langle f, f\rangle_\rho} := \|f\|_{L^2_{\rho_X}} = \left[\int_X \|f(\mu_a)\|_Y^2\,\mathrm{d}\rho_X(\mu_a)\right]^{\frac{1}{2}}$. Let us assume that the operator-valued kernel $K : X \times X \to \mathcal{L}(Y)$ is a Mercer kernel (that is $\mathcal{H} = \mathcal{H}(K) \subseteq C(X, Y) = \{X \to Y \text{ continuous functions}\}$), is bounded ($\exists B_K < \infty$ such that $\|K(x, x)\|_{\mathcal{L}(Y)} \le B_K$), and is a compact operator for all $x \in X$. These requirements will be guaranteed by our assumptions, see Section 7.2.6. In this case, the inclusion S_K^*: $\mathcal{H} \hookrightarrow L^2_{\rho_X}$ is bounded, and its adjoint $S_K : L^2_{\rho_X} \to \mathcal{H}$ is given by

$$(S_K g)(\mu_u) = \int_X K(\mu_u, \mu_t) g(\mu_t)\mathrm{d}\rho_X(\mu_t). \tag{5}$$

We further define $\tilde{T}$ as

$$\tilde{T} = S_K^* S_K : L^2_{\rho_X} \to L^2_{\rho_X}; \tag{6}$$

in other words, the result of operation (5) belongs to $\mathcal{H}$, which is embedded in $L^2_{\rho_X}$. $\tilde{T}$ is a compact, positive, self-adjoint operator (Carmeli et al., 2010, Proposition 3), thus by the spectral theorem $\tilde{T}^s$ exists, where $s \ge 0$.

6. The $x \mapsto \mu_x$ mapping is defined for *all* $x \in \mathcal{M}_1^+(\mathcal{X})$ if k is bounded, i.e., $\sup_{u\in\mathcal{X}} k(u, u) < \infty$.
7. Our assumptions will guarantee the existence of ρ (see Section 3). Since Y is a Polish space (because it is separable Hilbert) the $\rho(y|\mu_a)$ conditional distribution ($y \in Y$, $\mu_a \in X$) is also well-defined (Steinwart and Christmann, 2008, Lemma A.3.16, page 487).

2.2 Distribution Regression

We now formally define the distribution regression task. Let us assume that $\mathcal{M}_1^+(\mathcal{X})$ is endowed with $\mathcal{S}_1 = \mathcal{B}(\tau_w)$, the weak-topology generated σ-algebra; thus $(\mathcal{M}_1^+(\mathcal{X}), \mathcal{S}_1)$ is a measurable space. 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 $(\mathcal{M}_1^+(\mathcal{X}) \times Y, \mathcal{S}_1 \otimes \mathcal{B}(Y))$, the product space enriched with the product σ-algebra. 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 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 problem ($\mathbb{R}^d \to \mathbb{R}$), distribution regression can be tackled via kernel ridge regression (using a 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 Y. 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 $f \circ \mu_x$, i.e.

$$\mathcal{M}_1^+(\mathcal{X}) \xrightarrow{\mu} X (\subseteq H = H(k)) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} Y. \tag{7}$$

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 $\mathcal{H}$.

Let the expected risk for a $\tilde{f} : X \to Y$ (measurable) function be defined as

$$\mathcal{R}[\tilde{f}] = \mathbb{E}_{(x,y)\sim\mathcal{M}} \left\| \tilde{f}(\mu_x) - y \right\|_Y^2,$$

which is minimized by the f_ρ regression function. The classical regularization approach is to optimize

$$f_\mathbf{z}^\lambda = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{x_i}) - y_i\|_Y^2 + \lambda \|f\|_\mathcal{H}^2 \tag{8}$$

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

$$f_{\hat{\mathbf{z}}}^\lambda = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{\hat{x}_i}) - y_i\|_Y^2 + \lambda \|f\|_\mathcal{H}^2, \tag{9}$$

where $\hat{x}_i = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_{i,n}}$ is the empirical distribution determined by $\{x_{i,n}\}_{i=1}^N$. The ridge regression objective function has an analytical solution: given training samples $\hat{\mathbf{z}}$, the prediction for a new t test distribution is

$$(f_{\hat{\mathbf{z}}}^\lambda \circ \mu)(t) = \mathbf{k}(\mathbf{K} + l\lambda I)^{-1}[y_1; \ldots; y_l], \tag{10}$$

where $\mathbf{k} = [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_l}, \mu_t)] \in \mathcal{L}(Y)^{1\times l}$, $\mathbf{K} = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l\times l}$, $[y_1; \ldots; y_l] \in Y^l$.

Remark 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.* (9)*.*
- *There is a* two-stage sampling difficulty *to tackle: The transition from f_ρ 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 ($\hat{\mathbf{z}}$).*
- *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.*
- *In case of scalar output ($Y = \mathbb{R}$), $\mathcal{L}(Y) = \mathcal{L}(\mathbb{R}) = \mathbb{R}$ and* (10) *is a standard linear equation with $\mathbf{K} \in \mathbb{R}^{l\times l}$, $\mathbf{k} \in \mathbb{R}^{1\times l}$. More generally, if $Y = \mathbb{R}^d$, then $\mathcal{L}(Y) = \mathcal{L}(\mathbb{R}^d) = \mathbb{R}^{d\times d}$ and* (10) *is still a finite-dimensional linear equation with $\mathbf{K} \in \mathbb{R}^{(dl)\times(dl)}$ and $\mathbf{k} \in \mathbb{R}^{d\times(dl)}$.*
- *One could also formulate the problem (and get guarantees) for more abstract $X \subseteq H \to Y$ regression tasks [see Eq.* (7)*] on a convex set X with H and Y being general, separable Hilbert spaces. Since distribution regression is probably the most accessible example where two-stage sampling appears, and in order to keep the presentation simple, we do not consider such extended formulations in this work.*

Our main goals in this paper are as follows: first, to analyse the excess risk

$$\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) := \mathcal{R}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{R}[f_\rho],$$

both when $f_\rho \in \mathcal{H}$ (the well-specified case) and $f_\rho \in L^2_{\rho_X}\backslash\mathcal{H}$ (the misspecified case); second, to establish consistency ($\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) \to 0$, or in the misspecified case $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) - D_{\mathcal{H}}^2 \to 0$, where $D_{\mathcal{H}}^2 := \inf_{q\in\mathcal{H}} \|f_\rho - S_K^* q\|_\rho^2$ is the approximation error of f_ρ by a function in $\mathcal{H}$); and third, to derive an exact computational-statistical efficiency trade-off as a function of the (l, N, λ) triplet, and of the difficulty of the problem.

3. Assumptions

In this section, we detail our assumptions on the $(\mathcal{X}, Y, k, K)$ quartet. Our analysis for the well-specified case uses existing ridge regression results (Caponnetto and De Vito, 2007) focusing on problem (8) 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 challenging; the available bounds can moderately shorten our proof. We must take particular care in verifying that Caponnetto and De Vito (2007)'s conditions are met, since they need to 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** are as follows:

1. $(\mathcal{X}, \tau)$ is a separable, topological space.

2. Y is a separable Hilbert space.
3. k is bounded, in other words $\exists B_k < \infty$ such that $\sup_{u \in \mathcal{X}} k(u,u) \leq B_k$, and continuous.
4. The $\{K_{\mu_a}\}_{\mu_a \in X}$ operator family is uniformly bounded in Hilbert-Schmidt norm and Hölder continuous in operator norm. Formally, $\exists B_K < \infty$ such that

$$\|K_{\mu_a}\|^2_{\mathcal{L}_2(Y,\mathcal{H})} = Tr\left(K^*_{\mu_a} K_{\mu_a}\right) \leq B_K, \quad (\forall \mu_a \in X), \tag{11}$$

and $\exists L > 0$, $h \in (0,1]$ such that the mapping $K_{(\cdot)} : X \to \mathcal{L}(Y,\mathcal{H})$ is Hölder continuous:

$$\|K_{\mu_a} - K_{\mu_b}\|_{\mathcal{L}(Y,\mathcal{H})} \leq L \|\mu_a - \mu_b\|^h_H, \quad \forall (\mu_a, \mu_b) \in X \times X. \tag{12}$$

5. y is bounded: $\exists C < \infty$ such that $\|y\|_Y \leq C$ almost surely.

These requirements hold under mild conditions: in Section 8, we provide insight into the consequences of our assumptions, with several concrete illustrations (e.g. regression with set- and RBF-type kernels).

4. Error Bounds, Consistency & Computational-Statistical Efficiency Trade-off

In this section, we present our analysis of the consistency of the mean embedding based ridge regression (MERR) method.

Given the estimator $(f^{\lambda}_{\hat{\mathbf{z}}})$ in Eq. (9), we derive finite-sample high probability upper bounds (see Theorems 2 and 7) for the excess risk $\mathcal{E}\left(f^{\lambda}_{\hat{\mathbf{z}}}, f_\rho\right)$, and in the misspecified setting, for the excess risk compared to the best attainable value from $\mathcal{H}$, i.e., $\mathcal{E}\left(f^{\lambda}_{\hat{\mathbf{z}}}, f_\rho\right) - D^2_{\mathcal{H}}$. We illustrate the bounds for particular classes of prior distributions, and work through special cases to obtain consistency conditions and computational-statistical efficiency trade-offs (see Theorems 4, 9 and the 3rd bullet of Remark 8). The main challenge is how to turn the convergence rates of the mean embeddings into those for an error $\mathcal{E}$ of the predictor. Although the main ideas of the proofs can be summarized relatively briefly, the full details are more demanding. High-level ideas with the sketches of the proofs and the obtained results are presented in Section 4.1 (well-specified case) and Section 4.2 (misspecified case). The derivations of some technical details of Theorems 2 and 7 are available in Section 7.

4.1 Results for the Well-specified Case

We first focus on the well-specified case ($f_\rho \in \mathcal{H}$) and present our first main result. We derive a high probability upper bound for the excess risk $\mathcal{E}\left(f^{\lambda}_{\hat{\mathbf{z}}}, f_\rho\right)$ of the MERR method (Theorem 2). The upper bound is instantiated for a general class of prior distributions (Theorem 4), which leads to a simple computational-statistical efficiency description (Theorem 5); this shows (among others) conditions when the MERR technique is able to achieve the *one-stage* sampled minimax optimal rate. We first give a high-level sketch of our convergence analysis and an intuitive interpretation of the results. An outline of the main proof ideas is given below, with technical details in Section 7.

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 $\hat{\mathbf{z}}$, respectively. One can express $f^{\lambda}_{\mathbf{z}}$ [Eq. (8)] (Caponnetto and De Vito, 2007), and similarly $f^{\lambda}_{\hat{\mathbf{z}}}$ [Eq. (9)],

as

$$f_{\mathbf{z}}^{\lambda} = (T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}}, \qquad T_{\mathbf{x}} = \frac{1}{l}\sum_{i=1}^{l} T_{\mu_{x_i}}, \qquad g_{\mathbf{z}} = \frac{1}{l}\sum_{i=1}^{l} K_{\mu_{x_i}} y_i, \tag{13}$$

$$f_{\hat{\mathbf{z}}}^{\lambda} = (T_{\hat{\mathbf{x}}} + \lambda)^{-1} g_{\hat{\mathbf{z}}}, \qquad T_{\hat{\mathbf{x}}} = \frac{1}{l}\sum_{i=1}^{l} T_{\mu_{\hat{x}_i}}, \qquad g_{\hat{\mathbf{z}}} = \frac{1}{l}\sum_{i=1}^{l} K_{\mu_{\hat{x}_i}} y_i, \tag{14}$$

where $T_{\mu_a} = K_{\mu_a} K_{\mu_a}^* \in \mathcal{L}(\mathcal{H})$ ($\mu_a \in X$), $T_{\mathbf{x}}, T_{\hat{\mathbf{x}}} : \mathcal{H} \to \mathcal{H}$, $g_{\mathbf{z}}, g_{\hat{\mathbf{z}}} \in \mathcal{H}$. By these explicit expressions, one can decompose the excess risk into 5 terms (Szabó et al., 2015, Section A.1.8):

$$\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) = \mathcal{R}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{R}\left[f_\rho\right] \le 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda) + S_1 + S_2\right],$$

where

$$S_{-1} = S_{-1}(\lambda, \mathbf{z}, \hat{\mathbf{z}}) = \|\sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda I)^{-1}(g_{\hat{\mathbf{z}}} - g_{\mathbf{z}})\|_{\mathcal{H}}^2, \tag{15}$$

$$S_0 = S_0(\lambda, \mathbf{z}, \hat{\mathbf{z}}) = \|\sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda I)^{-1}(T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}) f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2, \tag{16}$$

$$\mathcal{A}(\lambda) = \|\sqrt{T}(f^{\lambda} - f_\rho)\|_{\mathcal{H}}^2, \quad S_1 = S_1(\lambda, \mathbf{z}) = \|\sqrt{T}(T_{\mathbf{x}} + \lambda I)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}} f_\rho)\|_{\mathcal{H}}^2,$$

$$S_2 = S_2(\lambda, \mathbf{z}) = \|\sqrt{T}(T_{\mathbf{x}} + \lambda I)^{-1}(T - T_{\mathbf{x}})(f^{\lambda} - f_\rho)\|_{\mathcal{H}}^2,$$

$$f^{\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{H}} (\mathcal{R}[f] + \lambda \|f\|_{\mathcal{H}}^2), \;\; T = \int_X T_{\mu_a} \mathrm{d}\rho_X(\mu_a) = S_K S_K^* : \mathcal{H} \to \mathcal{H}. \tag{17}$$

Three of the terms (S_1, S_2, $\mathcal{A}(\lambda)$) are identical to the terms in Caponnetto and De Vito (2007), hence the earlier bounds can be applied. The two new terms (S_{-1}, S_0) resulting from two-stage sampling will be upper bounded by making use of the convergence of the empirical mean embeddings. These bounds will lead to the following results:

Theorem 2 (Finite-sample excess risk bounds; well-specified case) *Let $K_{(\cdot)} : X \to \mathcal{L}(Y, \mathcal{H})$ be Hölder continuous with constants L, h. Let $l \in \mathbb{Z}^+$, $N \in \mathbb{Z}^+$, $0 < \lambda$, $0 < \eta < 1$, $0 < \delta$, $C_\eta = 32 \log^2(6/\eta)$, $\|y\|_Y \le C$ (a.s.) and $\mathcal{A}(\lambda)$ be the residual as defined above. Define $M = 2(C + \|f_\rho\|_{\mathcal{H}} \sqrt{B_K})$, $\Sigma = \frac{M}{2}$, T as in (17), $\mathcal{B}(\lambda) = \|f^{\lambda} - f_\rho\|_{\mathcal{H}}^2$ as the reconstruction error, and $\mathcal{N}(\lambda) = Tr[(T + \lambda I)^{-1} T]$ as the effective dimension. Then with probability at least $1 - \eta - e^{-\delta}$, the excess risk can be upper bounded as*

$$\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) \le 5 \Bigg\{ \frac{4L^2 \left(1 + \sqrt{\log(l) + \delta}\right)^{2h} (2B_k)^h}{\lambda N^h} \Bigg[C^2 + 4B_K \times$$

$$\times \left(\log^2\left(\frac{6}{\eta}\right) \left\{ \frac{64}{\lambda}\left[\frac{M^2 B_K}{l^2 \lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l}\right] + \frac{24}{\lambda^2}\left[\frac{4B_K^2 \mathcal{B}(\lambda)}{l^2} + \frac{B_K \mathcal{A}(\lambda)}{l}\right] \right\} + \mathcal{B}(\lambda) + \|f_\rho\|_{\mathcal{H}}^2 \right) \Bigg]$$

$$+ \mathcal{A}(\lambda) + C_\eta \left[\frac{B_K^2 \mathcal{B}(\lambda)}{l^2 \lambda} + \frac{B_K \mathcal{A}(\lambda)}{4l\lambda} + \frac{B_K M^2}{l^2 \lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l} \right] \Bigg\}$$

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

Below we specialize our excess risk bound for a general prior class, which captures the difficulty of the regression problem as defined in Caponnetto and De Vito (2007). This $\mathcal{P}(b,c)$ class is described by two parameters b and c: larger b means faster decay of the eigenvalues of the covariance operator T [in Eq. (17)], hence smaller effective input dimension; larger c corresponds to a smoother regression function. Formally:

Definition of the $\mathcal{P}(b,c)$ class: Let us fix the positive constants 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 $Z = X \times Y$ such that

1. a range space assumption is satisfied: $\exists g \in \mathcal{H}$ s.t. $f_\rho = T^{\frac{c-1}{2}} g$ with $\|g\|^2_{\mathcal{H}} \leq R$,
2. in the spectral decomposition of $T = \sum_{n=1}^{\infty} \lambda_n \langle \cdot, e_n \rangle_{\mathcal{H}} e_n$, where $(e_n)_{n=1}^{\infty}$ is a basis of $Ker(T)^{\perp}$, the eigenvalues of T satisfy $\alpha \leq n^b \lambda_n \leq \beta \quad (\forall n \geq 1)$.

Remark 3 *We make few remarks about the $\mathcal{P}(b,c)$ class:*

- Range space assumption on f_ρ: *The smoothness of f_ρ is expressed as a range space assumption, which is slightly different from the standard smoothness conditions appearing in non-parametric function estimation. By the spectral decomposition of T given above [$\lambda_1 \geq \lambda_2 \geq \ldots > 0, \lim_{n\to\infty} \lambda_n = 0$], $T^r u = \sum_{n=1}^{\infty} (\lambda_n)^r \langle u, e_n \rangle_{\mathcal{H}} e_n \quad (r = \frac{c-1}{2} \geq 0, u \in \mathcal{H})$ and*

$$Im(T^r) = \Big\{ \sum\nolimits_{n=1}^{\infty} c_n e_n : \sum\nolimits_{n=1}^{\infty} c_n^2 \lambda_n^{-2r} < \infty \Big\}. \tag{18}$$

  Specifically, in the limit as $r \to 0$, we obtain $f_\rho \in Im(T^0) = Im(I) = \mathcal{H}$ (no constraint); larger values of r give rise to faster decay of the $(c_n)_{n=1}^{\infty}$ Fourier coefficients. This is the concrete meaning of $f_\rho \in Im(T^r)$.

- Spectral decay condition: *We can provide a simple illustration of when the spectral 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 $K(\mu_{x_i}, \mu_{x_j}) = e^{-c\|m_i - m_j\|^2}$ (Muandet et al., 2012, 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: see e.g. Steinwart and Christmann (2008) for a discussion of these 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: $\mathcal{A}(\lambda) \leq R\lambda^c$, $\mathcal{B}(\lambda) \leq R\lambda^{c-1}$, $\mathcal{N}(\lambda) \leq \beta \frac{b}{b-1} \lambda^{-\frac{1}{b}}$. Specializing Theorem 2 and retaining its assumptions, we get:

Theorem 4 (Finite-sample excess risk bound for $\rho \in \mathcal{P}(b,c)$)
Suppose the conditions in Theorem 2 hold. Let $\rho \in \mathcal{P}(b,c)$, where $1 < b$ and $c \in (1,2]$.

Then

$$\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) \leq 5\Bigg\{\frac{4L^2\left(1+\sqrt{\log(l)+\delta}\right)^{2h}(2B_k)^h}{\lambda N^h}\Bigg[C^2+4B_K\times$$
$$\times\left(C_\eta\left\{\frac{2}{\lambda}\left[\frac{M^2B_K}{l^2\lambda}+\frac{\Sigma^2\beta b}{(b-1)l\lambda^{\frac{1}{b}}}\right]+\frac{3}{4\lambda^2}\left[\frac{4B_K^2R\lambda^{c-1}}{l^2}+\frac{B_KR\lambda^c}{l}\right]\right\}+R\lambda^{c-1}+\|f_\rho\|_{\mathcal{H}}^2\right)\Bigg]$$
$$+R\lambda^c+C_\eta\left[\frac{B_K^2R\lambda^{c-2}}{l^2}+\frac{B_KR\lambda^{c-1}}{4l}+\frac{B_KM^2}{l^2\lambda}+\frac{\Sigma^2\beta b}{(b-1)l\lambda^{\frac{1}{b}}}\right]\Bigg\}.$$

Discarding the constants in Theorem 4, the study of convergence of the excess risk $\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho)$ to 0 boils down to finding N and λ (as a function of l) where $N\to\infty$, $\lambda\to 0$ and

$$r(l,N,\lambda)=\frac{\log^h(l)}{N^h\lambda}\left(\frac{1}{\lambda^2l^2}+1+\frac{1}{l\lambda^{1+\frac{1}{b}}}\right)+\lambda^c+\frac{1}{l^2\lambda}+\frac{1}{l\lambda^{\frac{1}{b}}}\to 0,\ \text{s.t. } l\lambda^{\frac{b+1}{b}}\geq 1, \frac{\log(l)}{\lambda^{\frac{2}{h}}}\leq N \tag{19}$$

as $l\to\infty$. Let us choose $N=l^{\frac{a}{h}}\log(l)$; in this case Eq. (19) reduces to

$$r(l,\lambda)=\frac{1}{l^{2+a}\lambda^3}+\frac{1}{l^a\lambda}+\frac{1}{l^{a+1}\lambda^{2+\frac{1}{b}}}+\lambda^c+\frac{1}{l^2\lambda}+\frac{1}{l\lambda^{\frac{1}{b}}}\to 0,\ \text{s.t. } l\lambda^{\frac{b+1}{b}}\geq 1,\ l^a\lambda^2\geq 1. \tag{20}$$

One can assume that $a>0$, otherwise $r(l,\lambda)\to 0$ fails to hold; in other words, N should grow faster than $\log(l)$. Matching the 'bias' (λ^s) and 'variance' (other) terms in $r(l,\lambda)$ to choose λ, and guaranteeing that the matched terms dominate and the constraints in Eq. (20) hold, one gets the following simple description for the computational-statistical efficiency trade-off:[8]

Theorem 5 (Computational-statistical efficiency trade-off; well-specified case; $\rho\in\mathcal{P}(b,c)$) *Suppose the conditions in Theorem 2 hold. Let $\rho\in\mathcal{P}(b,c)$ and $N=l^{\frac{a}{h}}\log(l)$, where $0<a$, $1<b$, $c\in(1,2]$. If*

- $a\leq\frac{b(c+1)}{bc+1}$, *then* $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right)=\mathcal{O}_p\left(l^{-\frac{ac}{c+1}}\right)$ *with* $\lambda=l^{-\frac{a}{c+1}}$,
- $a\geq\frac{b(c+1)}{bc+1}$ *then* $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right)=\mathcal{O}_p\left(l^{-\frac{bc}{bc+1}}\right)$ *with* $\lambda=l^{-\frac{b}{bc+1}}$.

Remark 6 *Theorem 5 formulates an exact computational-statistical efficiency trade-off for the choice of the bag size (N) as a function of the number of distributions (l) and problem difficulty (b, c).*

- a-dependence: *A smaller bag size (smaller a; $N=l^{\frac{a}{h}}\log(l)$) means computational savings, but reduced statistical efficiency. It is not worth increasing a above $\frac{b(c+1)}{bc+1}$ since from that point the rate becomes $r(l)=l^{-\frac{bc}{bc+1}}$; remarkably, this rate is minimax in the one-stage sampled setup (Caponnetto and De Vito, 2007). The sensible choice $a=\frac{b(c+1)}{bc+1}<2$ means that the one-stage sampled minimax rate can be achieved in the two-stage sampled setting with bag size N sub-quadratic in l.*

8. The derivations are available in the supplement of `http://arxiv.org/pdf/1411.2066`.

- h-dependence: *In accord with our 'smoothness' assumptions it is rewarding to use smoother K kernels (larger $h \in (0,1]$) since this reduces the bag size [$N = l^{\frac{a}{h}} \log(l)$].*
- c-dependence: *The strictly decreasing property of $c \mapsto \frac{b(c+1)}{bc+1}$ implies that for 'smoother' problems (larger c) fewer samples (N) are sufficient.*

Below we elaborate on the sketched high-level idea and prove Theorem 2.

Proof of Theorem 2 (detailed derivations of each step can be found in Section 7.1)

1. **Decomposition of the excess risk**: We have the following upper bound for the excess risk

$$\mathcal{E}\big(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\big) = \mathcal{R}\big[f_{\hat{\mathbf{z}}}^{\lambda}\big] - \mathcal{R}\left[f_\rho\right] \leq 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda) + S_1 + S_2\right]. \tag{21}$$

2. **It is sufficient to upper bound S_{-1} and S_0**: Caponnetto and De Vito (2007) have shown that for $\forall \eta > 0$ if $l \geq 2C_\eta B_K \mathcal{N}(\lambda)/\lambda$, $\lambda \leq \|T\|_{\mathcal{L}(\mathcal{H})}$, then $\mathbb{P}(\mathbf{\Theta}(\lambda, \mathbf{z}) \leq 1/2) \geq 1 - \eta/3$, where

$$\mathbf{\Theta}(\lambda, \mathbf{z}) = \left\|(T - T_{\mathbf{x}})(T + \lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}, \tag{22}$$

using which upper bounds on S_1 and S_2 that hold with probability $1 - \eta$ are obtained. It is known that $\mathcal{A}(\lambda) \leq R\lambda^c$.

3. **Probabilistic bounds on $\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2$, $\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|\sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$, $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$**: One can bound S_{-1} and S_0 as

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

and

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

For the terms on the r.h.s., we derive upper bounds [for the definition of α, see Eq. (24)]

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

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

and

$$\begin{aligned}\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2 \leq 6 \Bigg(& \frac{16}{\lambda} \log^2\left(\frac{6}{\eta}\right) \left[\frac{M^2 B_K}{l^2 \lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l}\right] \\ & + \frac{4}{\lambda^2} \log^2\left(\frac{6}{\eta}\right) \left[\frac{4B_K^2 \mathcal{B}(\lambda)}{l^2} + \frac{B_K \mathcal{A}(\lambda)}{l}\right] + \mathcal{B}(\lambda) + \|f_\rho\|_{\mathcal{H}}^2 \Bigg).\end{aligned} \tag{23}$$

The bounds hold under the following conditions:

- $\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2$ (see Section 7.1.1): if the empirical mean embeddings are close to their population counterparts, i.e.,

$$\|\mu_{x_i} - \mu_{\hat{x}_i}\|_H \leq \frac{(1+\sqrt{\alpha})\sqrt{2B_k}}{\sqrt{N}}, \quad (\forall i = 1, \ldots, l). \tag{24}$$

  This event has probability $1 - le^{-\alpha}$ over all $i = 1, \ldots, l$ samples; see (Altun and Smola, 2006) and (Szabó et al., 2015, Section A.1.10).

- $\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|_{\mathcal{L}(\mathcal{H})}^2$ (see Section 7.1.2): (24) is assumed.

- $\|\sqrt{T}(T_{\hat{\mathbf{x}}} + \lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})}^2$ (Szabó et al., 2015, Section A.1.11): (24), $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$, and

$$\frac{(1+\sqrt{\alpha})^2 \, 2^{\frac{h+6}{h}} B_k (B_K)^{\frac{1}{h}} L^{\frac{2}{h}}}{\lambda^{\frac{2}{h}}} \leq N. \tag{25}$$

- $\|f_{\mathbf{z}}^{\lambda}\|_{\mathcal{H}}^2$: The bound is guaranteed to hold under the conditions of the bounds of S_1 and S_2.[8]

4. **Union bound**: By applying an $\alpha = \log(l) + \delta$ reparameterization, and combining the received upper bounds with Caponnetto and De Vito (2007)'s results for S_1 and S_2, Theorem 2 follows (Section 7.1.3) with a union bound.

Finally, we note that existing results/ideas were used at two points to simplify our analysis: bounding S_1, S_2, $\Theta(\lambda, \mathbf{z})$, $\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2$ (Caponnetto and De Vito, 2007) and $\|\mu_{x_i} - \mu_{\hat{x}_i}\|_H$ (Altun and Smola, 2006).[9]

4.2 Results for the Misspecified Case

In this section, we focus on the misspecified case ($f_\rho \in L_{\rho_X}^2 \backslash \mathcal{H}$) and present our second main result, which was inspired by the proof technique of Sriperumbudur et al. (2014, Theorem 12). We derive a high probability upper bound for $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right)$, i.e., the excess risk of the MERR method (Theorem 7) which gives rise to consistency results (3rd bullet of Remark 8) and precise computational-statistical efficiency trade-off (Theorem 9). Theorem 7 consists of two finite-sample bounds:

1. The first, more general bound [Eq. (27)] will be used to show consistency in the misspecified case (see the 3rd bullet of Remark 8), in other words that $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right)$ can be driven to its smallest possible value determined by the "richness" of $\mathcal{H}$:

$$D_{\mathcal{H}}^2 := \inf_{q \in \mathcal{H}} \|f_\rho - S_K^* q\|_\rho^2. \tag{26}$$

   The value of $D_{\mathcal{H}}$ equals the approximation error of f_ρ by a function from $\mathcal{H}$. Specifically, if $\mathcal{H}$ [precisely $S_K^*(\mathcal{H}) = \{S_K^* q : q \in \mathcal{H}\} \subseteq L_{\rho_X}^2$] is *dense* in $L_{\rho_X}^2$, then $D_{\mathcal{H}} = 0$.

9. We also corrected some constants in the previous works (Altun and Smola, 2006; Caponnetto and De Vito, 2007).

2. The second, specialized result [Eq. (28)] under additional smoothness assumptions on f_ρ will give rise to a precise computational-statistical efficiency trade-off in terms of the problem difficulty (s) and sample numbers (l, N); this result can be seen as the misspecified analogue of Theorem 5.

After stating our results, the main ideas of the proof follow; further technical details are available in Section 7.2. Our main theorem for bounding the excess risk is as follows:

Theorem 7 (Finite-sample excess risk bounds; misspecified case) *Let* $l \in \mathbb{Z}^+$, $N \in \mathbb{Z}^+$, $0 < \lambda$, $0 < \eta < 1$, $0 < \delta$ *and* $C_\eta = \log\left(\frac{6}{\eta}\right)$. *Assume that* $\left(\frac{12B_K}{\lambda}C_\eta\right)^2 \le l$ *and* $\left(1+\sqrt{\log(l)+\delta}\right)^2 2^{\frac{h+6}{h}} B_k (B_K)^{\frac{1}{h}} L^{\frac{2}{h}}/\lambda^{\frac{2}{h}} \le N$.

1. *Then for arbitrary* $q \in \mathcal{H}$ *with probability at least* $1-\eta-e^{-\delta}$

$$\sqrt{\mathcal{E}\left(f_{\hat{\mathbf{z}}}^\lambda, f_\rho\right)} \le \frac{2LC\left(1+\sqrt{\log(l)+\delta}\right)^h (2B_k)^{\frac{h}{2}}}{\sqrt{\lambda}N^{\frac{h}{2}}}\left(1+\frac{2\sqrt{B_K}}{\sqrt{\lambda}}\right)+ \tag{27}$$
$$\frac{2C_\eta}{\sqrt{\lambda}}\left\{\left(\frac{2C\sqrt{B}_K}{l}+\frac{C\sqrt{B_K}}{\sqrt{l}}\right)+\left(\frac{2B_K}{l}+\frac{\sigma}{\sqrt{l}}\right)\frac{1}{\lambda}\sqrt{\lambda\left\|f_\rho\right\|_\rho D_a(\lambda,q)}\right\}+D_a(\lambda,q),$$

where $D_a(\lambda,q) = \|f_\rho - S_K^* q\|_\rho + \max(1, \|T\|_{\mathcal{L}(\mathcal{H})})\lambda^{\frac{1}{2}}\|q\|_{\mathcal{H}}$.

2. *In addition, suppose* $f_\rho \in Im(\tilde{T}^s)$ *for some* $s > 0$, *where* $\tilde{T}$ *is defined in Eq.* (6). *Then with probability at least* $1-\eta-e^{-\delta}$, *we have*

$$\sqrt{\mathcal{E}\left(f_{\hat{\mathbf{z}}}^\lambda, f_\rho\right)} \le \frac{2LC\left(1+\sqrt{\log(l)+\delta}\right)^h (2B_k)^{\frac{h}{2}}}{\sqrt{\lambda}N^{\frac{h}{2}}}\left(1+\frac{2\sqrt{B_K}}{\sqrt{\lambda}}\right)+$$
$$\frac{2C_\eta}{\sqrt{\lambda}}\left\{\left(\frac{2C\sqrt{B}_K}{l}+\frac{C\sqrt{B_K}}{\sqrt{l}}\right)+\left(\frac{2B_K}{l}+\frac{\sigma}{\sqrt{l}}\right)\frac{1}{\lambda}\times\right.$$
$$\left.\sqrt{\max\left(1,\left\|\tilde{T}\right\|^s_{\mathcal{L}\left(L^2_{\rho_X}\right)}\right)\lambda\left\|\tilde{T}^{-s}f_\rho\right\|_\rho D_b(\lambda,s)}\right\}+D_b(\lambda,s), \tag{28}$$

where $D_b(\lambda,s) = \max(1, \|\tilde{T}\|^{s-1}_{\mathcal{L}(L^2_{\rho_X})})\lambda^{\min(1,s)}\|\tilde{T}^{-s}f_\rho\|_\rho$.

Remark 8 *We give a short insight into the assumptions of Theorem 7, followed by consequences of the theorem.*

- Range space assumption on f_ρ: *The range space assumption for the compact, positive, self-adjoint operator,* $\tilde{T} = \tilde{T}(K) : L^2_{\rho_X} \to L^2_{\rho_X}$ *in the 2nd part of Theorem 7 can be interpreted similarly to that on* T; *see Eq.* (18). *One can also prove alternative descriptions for* $Im(\tilde{T}^s)$ *in terms of interpolation spaces (Steinwart and Scovel, 2012, Theorem 4.6, page 387), or the decay of the 2-approximation error function,* $A_2(\lambda) = \inf_{f\in\mathcal{H}(K)}\left(\lambda\|f\|^2_{\mathcal{H}(K)} + \mathcal{R}[f] - \mathcal{R}[f_\rho]\right)$ *(Smale and Zhou, 2003; Steinwart et al., 2009).*

- $\sqrt{\mathcal{E}\left(f^{\lambda}_{\hat{\mathbf{z}}}, f_\rho\right)}$*: Notice that in the bounds [*(27), (28)*], instead of the excess risk, its square root appears; this has technical reasons, as it is easier to have the* $D_a(\lambda, q)$ *quantity (without multiplicative constants) appear on the r.h.s. of Eq.* (27) *with this form.*
- Consistency in the misspecified case: *The consequence of Theorem 7(1) is as follows. Discarding the constants in Eq.* (27)*, we obtain the upper bound (notice that the constant multiplier of* $\|f_\rho - S_K^* q\|_\rho$ *in the last term was* one*):*

$$\sqrt{r(l,N,\lambda,q)} = \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda} + \frac{1}{\sqrt{l\lambda}} + \frac{\sqrt{\|f_\rho - S_K^* q\|_\rho + \sqrt{\lambda}\,\|q\|_{\mathcal{H}}}}{\lambda\sqrt{l}} + \|f_\rho - S_K^* q\|_\rho + \sqrt{\lambda}\,\|q\|_{\mathcal{H}}\,.$$

By choosing $N = l^{1/h}\log l$*,* $\sqrt{r(l,\lambda)}$ *is bounded by*

$$\inf_{q\in\mathcal{H}}\left\{\|f_\rho - S_K^* q\|_\rho + \frac{\sqrt{\|f_\rho - S_K^* q\|_\rho}}{\lambda\sqrt{l}} + \frac{\sqrt{\|q\|_{\mathcal{H}}}}{\lambda^{\frac{3}{4}}\sqrt{l}} + \sqrt{\lambda}\|q\|_{\mathcal{H}}\right\} + \mathcal{O}_p\left(\frac{1}{\sqrt{\lambda l}}\right).$$

Our goal is to investigate the behavior of the bound as $l\to\infty$*,* $\lambda\to 0$ *and* $\lambda\sqrt{l}\to\infty$*. Define* $K(\alpha,\beta,\gamma) := \inf_{q\in\mathcal{H}}\left\{\|f_\rho - S_K^* q\|_\rho + \alpha\sqrt{\|f_\rho - S_K^* q\|_\rho} + \beta\sqrt{\|q\|_{\mathcal{H}}} + \gamma\|q\|_{\mathcal{H}}\right\}$*.* $K(\alpha,\beta,\gamma)$ *is the pointwise infimum of affine functions, therefore it is upper semi-continuous and concave on* $\mathbb{R}^3$ *(Aliprantis and Border, 2006, Lemmas 2.41 and 5.40); it is continuous on* $\times_{i=1}^3\mathbb{R}_{>0}$ *(Rockafellar and Wets, 2008, Theorem 2.35). Moreover, by applying (Rockafellar and Wets, 2008, Corollary 2.37) it extends continuously to* $\times_{i=1}^3\mathbb{R}_{\geq 0}$*; specifically it is continuous at* $(\alpha,\beta,\gamma)=\mathbf{0}$*. In other words, as* $l\to\infty$*,* $\lambda\to 0$ *and* $\lambda\sqrt{l}\to\infty$*,* $K\left(\frac{1}{\lambda\sqrt{l}}, \frac{1}{\lambda^{\frac{3}{4}}\sqrt{l}}, \sqrt{\lambda}\right)\to D_{\mathcal{H}}$ *and we get consistency in the misspecified case,*[10]

$$\sqrt{r(N,l,\lambda)}\to D_{\mathcal{H}}.$$

Discarding the constants in Eq. (28) we get[10]

$$\sqrt{r(l,N,\lambda)} = \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda} + \frac{1}{\sqrt{l\lambda}} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda\sqrt{l}} + \lambda^{\min(1,s)},\ \text{subject to}\ \frac{1}{\lambda^2}\leq l. \tag{29}$$

Our goal is to drive $r(l,N,\lambda)$ to zero with a suitable choice of the (l,N,λ) triplet under the stronger range space assumption. Since in Eq. (29) $\min(1,s)$ appears, one can assume without loss of generality that $s\in(0,1]$; consequently $1-\frac{s}{2}\in\left[\frac{1}{2},1\right)$ and $\frac{1}{l^{\frac{1}{2}}\lambda^{\frac{1}{2}}}\leq\frac{1}{\lambda^{1-\frac{s}{2}}l^{\frac{1}{2}}}$. Let us choose $N = l^{2a/h}\log(l)$; in this case using the previous dominance note, Eq. (29) reduces to the study of

$$\sqrt{r(l,\lambda)} = \frac{1}{l^a\lambda} + \frac{1}{\lambda^{1-\frac{s}{2}}l^{\frac{1}{2}}} + \lambda^s\to 0,\ \text{s.t. } l\lambda^2\geq 1. \tag{30}$$

One can assume that $a>0$, otherwise $r(l,\lambda)\to 0$ fails to hold: in other words, N should grow faster than $\log(l)$. Matching the 'bias' (λ^s) and 'variance' (other) terms in $r(l,\lambda)$ to choose λ, guaranteeing that the matched terms dominate and the constraint in Eq. (30) hold, one can arrive at the following computational-statistical efficiency trade-off:[8]

10. We have discarded the $\log(l)/\lambda^{\frac{2}{h}}\leq N$ constraint implied by the convergence of the first term in $\sqrt{r}$.

Theorem 9 (Computational-statistical efficiency trade-off; misspecified case, $f_\rho \in Im(\tilde{T}^s)$) *Suppose that $f_\rho \in Im(\tilde{T}^s)$ and $N = l^{\frac{2a}{h}} \log(l)$, where $s \in (0,1]$, $a > 0$. If*

- $a \leq \frac{s+1}{s+2}$, *then* $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) = \mathcal{O}_p\left(l^{-\frac{2sa}{s+1}}\right)$ *with* $\lambda = l^{-\frac{a}{s+1}}$,
- $a \geq \frac{s+1}{s+2}$, *then* $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) = \mathcal{O}_p\left(l^{-\frac{2s}{s+2}}\right)$ *with* $\lambda = l^{-\frac{1}{s+2}}$.

Remark 10 *Theorem 9 provides a complete computational-statistical efficiency trade-off description for the choice of the bag size (N) as a number of the distributions (l).*

- a-dependence: *A smaller value of 'a' (smaller bags $N = l^{2a/h}\log(l)$) leads to a computational advantage, but one looses in statistical efficiency. As 'a' reaches $\frac{s+1}{s+2}$, the rate becomes $r(l) = l^{-\frac{2s}{s+2}}$ and one does not gain from further increasing the value of a. The sensible choice of $a = \frac{s+1}{s+2} \leq \frac{2}{3}$ means that N can again be* sub-quadratic *($2a < \frac{4}{3} < 2$) in l.*
- h-dependence: *By using smoother K kernels (larger $h \in (0,1]$) one can reduce the size of the bags: $h \mapsto 2a/h$ is decreasing in h. This is compatible with our smoothness requirement on f_ρ.*
- s-dependence: *"Easier" tasks (larger s) give rise to faster convergence. Indeed, in the $r(l) = l^{-\frac{2s}{s+2}}$ rate the $s \mapsto \frac{2s}{s+2}$ exponent is strictly increasing function of the problem difficulty (s). For example, for extremely non-smooth regression problems ($s \approx 0$) the convergence can be arbitrary slow ($\lim_{s\to 0} \frac{2s}{s+2} = 0$). In the smooth case ($s = 1$) $\lim_{s\to 1} \frac{2s}{s+2} = \frac{2}{3}$ and one can achieve the $r(l) = l^{-\frac{2}{3}}$ rate.*
- *We may compare our $r(l) = l^{-\frac{2s}{s+2}}$ result with the $r_o(l) = l^{-\frac{2s}{2s+1}}$ (one-stage sampled) rate (Steinwart et al., 2009, $\beta/2 := s$, $q := 2$, $p := 1$ in Corollary 6), which was shown to be asymptotically optimal on $Y = \mathbb{R}$ for continuous k on compact metric $\mathcal{X}$. Steinwart* et al.*'s result is more general in terms of q ($\|f\|_{\mathcal{H}}^q$ based regularization) and p ($\|f\|_\infty \leq C\|f\|_{\mathcal{H}}^p \|f\|_\rho^{1-p}$, $\forall f \in \mathcal{H}$; in our case $p = 1$), although it imposes an* additional *eigenvalue constraint [(Steinwart et al., 2009, Eq. (6))] as well as $f_\rho \in Im(\tilde{T}^s)$. Moreover, one can observe that $r_o(l) \leq r(l)$ with a small gap, and that for $s \to 0$ and $s = 1$, $r_o(l) = r(l)$; see Fig. 1. We further remind the reader that our MERR analysis also holds for* separable Hilbert *output spaces Y,* separable topological *domains $\mathcal{X}$ enriched with a bounded, continuous kernel k, and that we handle the* two-stage sampled *setting.*

The main steps of the proof of Theorem 7 are as follows:

Proof of Theorem 7 (the details of the derivation are available in Section 7.2) Steps 1-7 will be identical in both proofs,[11] and we present them jointly.

1. **Decomposition of the excess risk**: By the triangle inequality, we have

$$\sqrt{\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right)} = \left\|S_K^* f_{\hat{\mathbf{z}}}^{\lambda} - f_\rho\right\|_\rho \leq \left\|S_K^*\left(f_{\hat{\mathbf{z}}}^{\lambda} - f_{\mathbf{z}}^{\lambda}\right)\right\|_\rho + \left\|S_K^* f_{\mathbf{z}}^{\lambda} - f_\rho\right\|_\rho. \tag{31}$$

11. Importantly, with a slight modification of the more general, first part of Theorem 7, one can get the specialized second setting of the theorem (see Step 8).

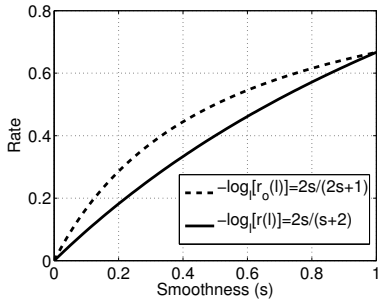

Figure 1: Comparison of the $r_o(l) = l^{-\frac{2s}{2s+1}}$ and $r(l) = l^{-\frac{2s}{s+2}}$ rates as function of the problem difficulty/smoothness (s).

2. **Bound on** $\left\|S_K^*\left(f_{\hat{\mathbf{z}}}^\lambda - f_{\mathbf{z}}^\lambda\right)\right\|_\rho$: Using[12] the fact that

$$\|S_K^* h\|_\rho^2 = \left\|\sqrt{T}h\right\|_{\mathcal{H}}^2 \quad (\forall h \in \mathcal{H}), \tag{32}$$

and the definitions of S_{-1} and S_0 [see Eqs. (15)-(16)], we obtain

$$\left\|S_K^*\left(f_{\hat{\mathbf{z}}}^\lambda - f_{\mathbf{z}}^\lambda\right)\right\|_\rho = \left\|\sqrt{T}\left(f_{\hat{\mathbf{z}}}^\lambda - f_{\mathbf{z}}^\lambda\right)\right\|_{\mathcal{H}} \leq \sqrt{S_{-1}} + \sqrt{S_0}, \tag{33}$$

through an application of triangle inequality. One can derive without a $\mathcal{P}(b,c)$ prior assumption (Section 7.2.1) the upper bound[13]

$$\sqrt{S_{-1}} + \sqrt{S_0} \leq \frac{2LC(1+\sqrt{\alpha})^h (2B_k)^{\frac{h}{2}}}{\sqrt{\lambda} N^{\frac{h}{2}}} \left[1 + \frac{2\sqrt{B_K}}{\sqrt{\lambda}}\right]$$

for the r.h.s. of Eq. (33) under the conditions that $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ (which holds with probability $1-\eta$ if $[12B_K \log(2/\eta)/\lambda]^2 \leq l$), and that Eqs. (24)-(25) hold.

3. **Decomposition of** $\left\|S_K^* f_{\mathbf{z}}^\lambda - f_\rho\right\|_\rho$: By the triangle inequality and Eq. (32), we have

$$\begin{aligned}\left\|S_K^* f_{\mathbf{z}}^\lambda - f_\rho\right\|_\rho &= \left\|S_K^*\left(f_{\mathbf{z}}^\lambda - f^\lambda\right) + S_K^* f^\lambda - f_\rho\right\|_\rho \leq \left\|S_K^*\left(f_{\mathbf{z}}^\lambda - f^\lambda\right)\right\|_\rho + \left\|S_K^* f^\lambda - f_\rho\right\|_\rho \\ &= \left\|\sqrt{T}\left(f_{\mathbf{z}}^\lambda - f^\lambda\right)\right\|_{\mathcal{H}} + \left\|S_K^* f^\lambda - f_\rho\right\|_\rho. \end{aligned} \tag{34}$$

4. **Decomposition of** $\left\|\sqrt{T}\left(f_{\mathbf{z}}^\lambda - f^\lambda\right)\right\|_{\mathcal{H}}$: Making use of the analytical expressions for $f_{\mathbf{z}}^\lambda$ and f^λ [see Eq. (13) and Eq. (17)], and the operator Woodbury formula (Ding and Zhou, 2008, Theorem 2.1, page 724) we arrive at the decomposition (see Section 7.2.2)

$$\begin{aligned}\left\|\sqrt{T}\left(f_{\mathbf{z}}^\lambda - f^\lambda\right)\right\|_{\mathcal{H}} \leq \left\|\sqrt{T}(T_{\mathbf{x}} + \lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \Big(&\|g_{\mathbf{z}} - g_\rho\|_{\mathcal{H}} + \\ &\|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \lambda^{-1} \left\|S_K\left[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right]\right\|_{\mathcal{H}}\Big), \end{aligned}$$

where $g_\rho = S_K f_\rho$. As it is known (Caponnetto and De Vito, 2007, page 348) $\|\sqrt{T}(T_{\mathbf{x}} + \lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq 1/\sqrt{\lambda}$ provided that $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$.

12. See for example de Vito et al. (2006) on page 88 with the $(\mathcal{H}, \mathcal{G}, A, T) := (\mathcal{H}, L^2_{\rho_X}, S_K^*, T)$ choice.
13. See the remark at the end of Section 7.2.1.

5. **Bound on $\|g_{\mathbf{z}} - g_\rho\|_{\mathcal{H}}$, $\|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})}$**: By concentration arguments the bounds

$$\|g_{\mathbf{z}} - g_\rho\|_{\mathcal{H}} \leq \left(\frac{4C\sqrt{B}_K}{l} + \frac{2C\sqrt{B_K}}{\sqrt{l}}\right)\log\left(\frac{2}{\eta}\right),\ \|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq \left(\frac{4B_K}{l} + \frac{4\sigma}{\sqrt{l}}\right)\log\left(\frac{2}{\eta}\right)$$

hold with probability at least $1-\eta$, each (see Section 7.2.3, 7.2.4).

6. **Decomposition of $\left\|S_K\left[f_\rho - (\tilde{T}+\lambda I)^{-1}S_K^* S_K f_\rho\right]\right\|_{\mathcal{H}}^2$**: Exploiting the analytical formula for f^λ, one can construct (Section 7.2.5) the upper bound

$$\left\|S_K\left[f_\rho - (\tilde{T}+\lambda I)^{-1}S_K^* S_K f_\rho\right]\right\|_{\mathcal{H}}^2 \leq \left\|\tilde{T}\left[f_\rho - (\tilde{T}+\lambda I)^{-1}S_K^* S_K f_\rho\right]\right\|_\rho \left\|S_K^* f^\lambda - f_\rho\right\|_\rho.$$

7. **Bound on $\left\|\tilde{T}\left[f_\rho - (\tilde{T}+\lambda I)^{-1}S_K^* S_K f_\rho\right]\right\|_\rho$**: Using our assumptions that $f_\rho \in Im(\tilde{T}^s)$ $(s \geq 0)$[14] and exploiting the separability of $L^2_{\rho_X}$, by Lemma 7.3.2 ($\mathcal{K} = L^2_{\rho_X}$, $f = f_\rho$, $M = \tilde{T}$, $a = 1$) and $\tilde{T} = S_K^* S_K$ we obtain the upper bound

$$\begin{aligned}&\left\|\tilde{T}\left[f_\rho - (\tilde{T}+\lambda I)^{-1}S_K^* S_K f_\rho\right]\right\|_\rho = \left\|\tilde{T}\left[f_\rho - (\tilde{T}+\lambda I)^{-1}\tilde{T} f_\rho\right]\right\|_\rho \\ &\leq \max\left(1, \|\tilde{T}\|^s_{\mathcal{L}(L^2_{\rho_X})}\right)\lambda^{\min(1,s+1)}\left\|\tilde{T}^{-s} f_\rho\right\|_\rho = \max\left(1, \|\tilde{T}\|^s_{\mathcal{L}(L^2_{\rho_X})}\right)\lambda\left\|\tilde{T}^{-s} f_\rho\right\|_\rho,\end{aligned}$$

where we used at the last step that $\min(1, s+1) = 1$; this follows from $s \geq 0$.

8. **Bound on $\left\|S_K^* f^\lambda - f_\rho\right\|_\rho$**:

   (a) **No range space assumption**: One can construct (Section 7.2.6) the bound

$$\left\|S_K^* f^\lambda - f_\rho\right\|_\rho \leq \|f_\rho - S_K^* q\|_\rho + \max\left(1, \|T\|_{\mathcal{L}(\mathcal{H})}\right)\lambda^{\frac{1}{2}}\|q\|_{\mathcal{H}},$$

   which holds for arbitrary $q \in \mathcal{H}$.

   (b) **Range space assumption in $L^2_{\rho_X}$**: Using the $S_K^* f^\lambda = (\tilde{T}+\lambda I)^{-1}\tilde{T} f_\rho$ identity [see Eq. (43)], and Lemma 7.3.2 ($M = \tilde{T}$, $\mathcal{K} = L^2_{\rho_X}$, $a = 0$), we get

$$\left\|S_K^* f^\lambda - f_\rho\right\|_\rho = \left\|(\tilde{T}+\lambda I)^{-1}\tilde{T} f_\rho - f_\rho\right\|_\rho \leq \max\left(1, \left\|\tilde{T}\right\|^{s-1}_{\mathcal{L}(L^2_{\rho_X})}\right)\lambda^{\min(1,s)}\left\|\tilde{T}^{-s} f_\rho\right\|_\rho.$$

9. **Union bound**: Applying an $\alpha = \log(l) + \delta$ reparameterization, changing η to $\frac{\eta}{3}$ and combining the derived results (in case of the first statement with $s = 0$) with a union bound, Theorem 7 follows.

Remark 11 *To contrast the derivation of the well- and the misspecified cases, we note that previous results [Section 4.1, or Caponnetto and De Vito (2007)'s bound] were used at two points:*

(a) In Step 2 by using Eq. (32) *and transforming the $L^2_{\rho_X}$ error $\left\|S_K^*\left(f^\lambda_{\hat{\mathbf{z}}} - f^\lambda_{\mathbf{z}}\right)\right\|_\rho$ to $\mathcal{H}$, we could rely on our previous bounds for S_{-1} and S_0. However, we were required to use a different concentration argument to guarantee $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ since we no longer assume the $\mathcal{P}(b, c)$ prior class.*

14. Note that we choose $s = 0$ and $s > 0$ in the first and second theorem part, respectively.

(b) In Step 4 the first term could be bounded by Caponnetto and De Vito (2007). Its $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ *condition was guaranteed by Step 2; and see Section 7.2.1.*

We note that our misspecified proof method was inspired by Sriperumbudur et al. (2014, Theorem 12), where the authors focused on the consistency of an infinite-dimensional exponential family estimator.

5. Related Work

In this section we discuss existing approaches and heuristic techniques to tackle learning problems on distributions.

Methods based on parametric assumptions: A number of methods have been proposed 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, see Kondor and Jebara, 2003; Jebara et al., 2004; Wang et al., 2009; Nielsen and Nock, 2012). A major limitation of these methods, however, is that they apply quite simple parametric assumptions, which may not be sufficient or verifiable in practise.

Methods based on parametric assumption in a RKHS: 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 Jebara et al., 2004; Zhou and Chellappa, 2006, 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.

Kernels based techniques: A more theoretically grounded approach to learning on distributions has been to define positive definite kernels 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 (Cuturi et al., 2005), non-extensive information theoretical kernel constructions (Martins et al., 2009), and kernels based on Hilbertian metrics (Hein and Bousquet, 2005). For example, the intuition of semigroup kernels (Cuturi et al., 2005) 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 (Hein and Bousquet, 2005), 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 (Póczos et al., 2011, 2012), but these similarity indices are *not* kernels, and their consistency in specific *learning tasks* remains an open question.

Multi-instance learning: An alternative paradigm in learning when the inputs are "bags of objects" is to simply treat each input as a *finite set*: this leads to the multi-instance learning task (MIL, see Dietterich et al., 1997; Ray and Page, 2001; Dooly et al., 2002). In MIL one is given a set of labelled bags, and the task of the learner is to find the mapping from the bags to the labels. Many important examples fit into the MIL framework: for example, different configurations of a given molecule can be handled as a bag of shapes, images can be considered as a set of patches or regions of interest, a video can be seen as a collection of images, a document might be described as a bag of words or paragraphs, a web page can be identified by its links, a group of people on a social network can be captured by their friendship graphs, in a biological experiment a subject can be identified by his/her time series trials, or a customer might be characterized by his/her shopping records. The MIL approach has been applied in several domains; see the reviews from Babenko (2004); Zhou (2004); Foulds and Frank (2010); Amores (2013).

"Bag-of-objects" methods (MIL, classification): Despite the large number of MIL applications and the spate of heuristic solution techniques, there exist few *theoretical results* in the area (Auer, 1998; Long and Tan, 1998; Blum and Kalai, 1998; Babenko et al., 2011; Zhang et al., 2013; Sabato and Tishby, 2012) and they focus on the multi-instance *classification* (MIC) task. In particular, let us first consider the standard MIC assumption (Dietterich et al., 1997), where a bag is declared to be positive (labelled with "1") if at least one of its instances is positive ("1"); otherwise, the bag is negative ("0").[15] In other words, if the instances $(x_{i,n})$ in the i^{th} bag $\{x_{i,1}, \ldots, x_{i,N}\}$ have hidden label $L(x_{i,n}) \in \{0,1\}$, then the observed label of the bag is $y_i = h(x_{i,1}, \ldots, x_{i,N}) = \max(L(x_{i,1}), \ldots, L(x_{i,N})) \in \{0,1\}$. In case of the original APR (axis-aligned rectangles; Dietterich et al., 1997) hypothesis class, function L is equal to the indicator of an unknown rectangle R ($L = \mathbb{I}_R$). In other words, a bag is declared to be positive if there exists at least one instance in the bag, which belongs to R.[16] The goal is to learn R with high probability given the bags ($\{x_{i,1}, \ldots, x_{i,N}\}$-s) and their labels ($y_i$-s). Long and Tan (1998) proved the PAC learnability (probably approximately correct; Valiant, 1984) of the APR hypothesis class, if all instances in each bag are i.i.d. and follow the same product distribution over the instance coordinates. On the other hand, for arbitrary distributions over bags, when the instances within a bag might be statistically dependent, APR learning under MIC is NP-hard (Auer, 1998); the same authors also showed that the product property (Long and Tan, 1998) on the coordinates is not required to obtain PAC results. Blum and Kalai (1998) extended PAC learnability of APR-s to hypothesis classes learnable from one-sided classification noise. In contrast to the previous approaches (Long and Tan, 1998; Auer, 1998; Blum and Kalai, 1998), Babenko et al. (2011) modelled the bags as low-dimensional manifolds, and proved PAC bounds. By relaxing the standard MIC assumption, Sabato and Tishby (2012) showed PAC-learnability for general MIC hypothesis classes with extended "max" functions. Zhang et al. (2013) derived high-probability generalization bounds in the MIC setting, when local and global representations are combined. Our work falls outside this setting since the label and bag generation mechanisms we consider are different: we do not assume an exact form of the

15. The motivation of this assumption comes from drug discovery: if a molecule has at least one well-binding configuration, then it is considered to bind well.

16. In terms of drug binding prediction, this means that a molecule binds to a target iff at least one of its configurations falls within a fixed, but unknown rectangle.

labelling mechanism (function L and max in h). Rather, the labelling is presumed to be stochastically determined by the underlying true distribution, not deterministically by the instance realizations in the bags (these are presumed i.i.d., and may be bag-specific).

"Bag-of-objects" methods (MIL, not classification): Beyond classification, there exist several *heuristics*—without consistency guarantees—for many other multi-instance problems in the literature, including regression (Ray and Page, 2001; Dooly et al., 2002; Zhou et al., 2009; Kwok and Cheung, 2007), clustering (Zhang and Zhou, 2009; Zhang et al., 2009, 2011; Chen and Wu, 2012), ranking (Bergeron et al., 2008; Hu et al., 2008; Bergeron et al., 2012), outlier detection (Wu et al., 2010), transfer learning (Raykar et al., 2008; Zhang and Si, 2009), and feature selection, -weighting and -extraction (also called dimensionality reduction, low-dimensional embedding, manifold learning, see Raykar et al., 2008; Ping et al., 2010; Sun et al., 2010; Carter et al., 2011; Zafra et al., 2013; Chai et al., 2014a,b, and references therein).

Approaches using set metrics: Adapting the bag viewpoint of MIL, one can come up with set metric based learning algorithms.[17] Probably one of the most well-known set metrics is the Hausdorff metric (Edgar, 1995), which is defined for non-empty compact sets of metric spaces, specifically for sets containing finitely many points. There also exist other (semi)metric constructions on points sets (Eiter and Mannila, 1997; Ramon and Bruynooghe, 2001). Unfortunately, the classical Hausdorff metric is highly sensitive to outliers, seriously limiting its practical applicability. In order to mitigate this deficiency, several variants of the Hausdorff metric have been designed in the MIL literature, such as the maximal-, the minimal- and the ranked Hausdorff metrics, with successful applications in MIC (Wang and Zucker, 2000) and multi-instance outlier detection (Wu et al., 2010); and the average Hausdorff metric (Zhang and Zhou, 2009) and contextual Hausdorff dissimilarity (Chen and Wu, 2012), which have been found useful in multi-instance clustering. Unfortunately, these methods lack any theoretical guarantee when applied in specific learning problems.

Functional data analysis techniques: Finally, the distribution regression task might also be interpreted as a functional data analysis problem (Ramsay and Silverman, 2002, 2005; Müller, 2005), by considering the probability measures x_i as functions. This is a highly non-standard setup, however, since these functions (x_i) are defined on σ-algebras and are non-negative, σ-additive.

6. Conclusion

We have established a learning theory of distribution regression, where the inputs are probability measures on separable, topological domains endowed with reproducing kernels, and the outputs are elements of a separable Hilbert space. We studied a ridge regression scheme defined on embeddings of the input distributions to a reproducing kernel Hilbert space, which has a simple analytical solution, as well as theoretically sound, efficient methods for approximation (Zhang et al., 2015; Richtárik and Takáč, 2016; Alaoui and Mahoney, 2015; Yang et al., 2016; Rudi et al., 2015). We derived explicit bounds on the excess risk as a function of the number of samples and problem difficulty. We tackled both the well-

17. Often these "metrics" are only semi-metrics, as they do not satisfy the triangle inequality.

specified case (when the regression function belongs to the assumed RKHS modelling class), and the more general misspecified setup. As a special case of our results, we proved the consistency of regression for set kernels (Haussler, 1999; Gärtner et al., 2002), which was a 17-year-old open problem, and for a recent kernel family (Christmann and Steinwart, 2010), which we have expanded upon (Table 1). We proved an exact computational-statistical efficiency trade-off for the MERR estimator: in the well-specified setting, we showed how to choose the bag size in the two-stage sampled setup to match the one-stage sampled minimax optimal rate (Caponnetto and De Vito, 2007); and in the misspecified setting, our rates approximate closely an asymptotically optimal estimator imposing stricter eigenvalue decay conditions (Steinwart et al., 2009). Several exciting open questions remain, including whether improved/optimal rates can be derived in the misspecified case, whether we can obtain consistency guarantees for non-point estimates, and how to handle non-ridge extensions.

Finally, we note that although the primary focus of the current paper was theoretical, we have applied the MERR method (Szabó et al., 2015, Section A.2) to supervised entropy learning and aerosol prediction based on multispectral satellite images.[18] In future work, we will address applications with vector-valued outputs.

7. Proofs

We provide proofs for our results detailed in Section 4: Section 7.1 (*resp.* Section 7.2) focuses on the well-specified case (*resp.* misspecified setting). The used lemmas are enlisted in Section 7.3.

7.1 Proofs of the Well-specified Case

We give proof details concerning the excess risk in the well-specified case (Theorem 2).

7.1.1 Proof of the bound on $\|g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}\|_{\mathcal{H}}^2$

By (13), (14) we get $g_{\hat{\mathbf{z}}} - g_{\mathbf{z}} = \frac{1}{l}\sum_{i=1}^{l}\left(K_{\mu_{\hat{x}_i}} - K_{\mu_{x_i}}\right)y_i$; hence by applying the Hölder property of $K_{(\cdot)}$, the boundedness of y_i ($\|y_i\|_Y \leq C$) and (24), we obtain

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

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

18. For code, see `https://bitbucket.org/szzoli/ite/`.

7.1.2 Proof of the 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 exploiting (with $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$) that in a normed space[19] $(N, \|\cdot\|)$, $f_i \in N$, $(i = 1, \ldots, n)$

$$\left\| \sum_{i=1}^{n} f_i \right\|^2 \leq n \sum_{i=1}^{n} \|f_i\|^2 , \tag{35}$$

we get

$$\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{l^2} l \sum_{i=1}^{l} \left\| T_{\mu_{x_i}} - T_{\mu_{\hat{x}_i}} \right\|^2_{\mathcal{L}(\mathcal{H})} . \tag{36}$$

To upper bound $\|T_{\mu_{x_i}} - T_{\mu_{\hat{x}_i}}\|^2_{\mathcal{L}(\mathcal{H})}$, let us see how $T_{\mu_u} = K_{\mu_a} K^*_{\mu_a}$ acts. The existence of an $E \geq 0$ constant satisfying $\|(T_{\mu_u} - T_{\mu_v})(f)\|_{\mathcal{H}} \leq E \|f\|_{\mathcal{H}}$ implies $\|T_{\mu_u} - T_{\mu_v}\|_{\mathcal{L}(\mathcal{H})} \leq E$. We continue with the l.h.s. of this equation using Eq. (35):

$$\begin{aligned}
\|(T_{\mu_u} - T_{\mu_v})(f)\|^2_{\mathcal{H}} &= \left\| K_{\mu_u} K^*_{\mu_u}(f) - K_{\mu_v} K^*_{\mu_v}(f) \right\|^2_{\mathcal{H}} \\
&= \left\| K_{\mu_u} \left[K^*_{\mu_u}(f) - K^*_{\mu_v}(f) \right] + \left(K_{\mu_u} - K_{\mu_v} \right) K^*_{\mu_v}(f) \right\|^2_{\mathcal{H}} \\
&\leq 2 \left[\left\| K_{\mu_u} \left[K^*_{\mu_u}(f) - K^*_{\mu_v}(f) \right] \right\|^2_{\mathcal{H}} + \left\| \left(K_{\mu_u} - K_{\mu_v} \right) K^*_{\mu_v}(f) \right\|^2_{\mathcal{H}} \right] .
\end{aligned}$$

By Eq. (45) and the Hölder continuity of $K_{(\cdot)}$, one arrives at

$$\begin{aligned}
&\left\| K_{\mu_u} \left[K^*_{\mu_u}(f) - K^*_{\mu_v}(f) \right] \right\|^2_{\mathcal{H}} \leq \|K_{\mu_u}\|^2_{\mathcal{L}(Y,\mathcal{H})} \left\| K^*_{\mu_u}(f) - K^*_{\mu_v}(f) \right\|^2_{Y} \\
&\quad \leq \|K_{\mu_u}\|^2_{\mathcal{L}(Y,\mathcal{H})} \left\| K^*_{\mu_u} - K^*_{\mu_v} \right\|^2_{\mathcal{L}(\mathcal{H},Y)} \|f\|^2_{\mathcal{H}} = \|K_{\mu_u}\|^2_{\mathcal{L}(Y,\mathcal{H})} \|(K_{\mu_u} - K_{\mu_v})^*\|^2_{\mathcal{L}(\mathcal{H},Y)} \|f\|^2_{\mathcal{H}} \\
&\quad = \|K_{\mu_u}\|^2_{\mathcal{L}(Y,\mathcal{H})} \|K_{\mu_u} - K_{\mu_v}\|^2_{\mathcal{L}(Y,\mathcal{H})} \|f\|^2_{\mathcal{H}} \leq B_K L^2 \|\mu_u - \mu_v\|^{2h}_H \|f\|^2_{\mathcal{H}} , \\
&\left\| \left(K_{\mu_u} - K_{\mu_v} \right) K^*_{\mu_v}(f) \right\|^2_{\mathcal{H}} \leq \|K_{\mu_u} - K_{\mu_v}\|^2_{\mathcal{L}(Y,\mathcal{H})} \left\| K^*_{\mu_v}(f) \right\|^2_{Y} \\
&\quad \leq \|K_{\mu_u} - K_{\mu_v}\|^2_{\mathcal{L}(Y,\mathcal{H})} \left\| K^*_{\mu_v} \right\|^2_{\mathcal{L}(\mathcal{H},Y)} \|f\|^2_{\mathcal{H}} \leq B_K L^2 \|\mu_u - \mu_v\|^{2h}_H \|f\|^2_{\mathcal{H}} .
\end{aligned}$$

Hence $\|(T_{\mu_u} - T_{\mu_v})(f)\|^2_{\mathcal{H}} \leq 4 B_K L^2 \|\mu_u - \mu_v\|^{2h}_H \|f\|^2_{\mathcal{H}} \;\Rightarrow\; E^2 = 4 B_K L^2 \|\mu_u - \mu_v\|^{2h}_H$. Exploiting this property in (36) with Eq. (24) we arrive to the bound

$$\begin{aligned}
\|T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}\|^2_{\mathcal{L}(\mathcal{H})} &\leq \frac{4 B_K L^2}{l} \sum_{i=1}^{l} \|\mu_{x_i} - \mu_{\hat{x}_i}\|^{2h}_H \leq \frac{4 B_K L^2}{l} \sum_{i=1}^{l} \frac{\left(1 + \sqrt{\alpha}\right)^{2h} (2 B_k)^h}{N^h} \\
&= \frac{\left(1 + \sqrt{\alpha}\right)^{2h} 2^{h+2} (B_k)^h B_K L^2}{N^h} .
\end{aligned} \tag{37}$$

7.1.3 Proof: final union bound in Theorem 2

Until now, we obtained that if (i) the sample number N satisfies Eq. (25), (ii) (24) holds (which has probability at least $1 - l e^{-\alpha} = 1 - e^{-[\alpha - \log(l)]} = 1 - e^{-\delta}$ applying a union bound

19. Eq. (35) holds since $\|\cdot\|^2$ is convex function, thus $\left\| \frac{1}{n} \sum_{i=1}^{n} f_i \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} \|f_i\|^2$.

argument; $\alpha = \log(l) + \delta$), and (iii) $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ is fulfilled [see Eq. (22)], then

$$S_{-1} + S_0 \leq \frac{4}{\lambda}\left[L^2 C^2 \frac{\left(1+\sqrt{\alpha}\right)^{2h}(2B_k)^h}{N^h} + \frac{\left(1+\sqrt{\alpha}\right)^{2h} 2^{h+2}(B_k)^h B_K L^2}{N^h} \times \right.$$
$$\left. \times \left(\log^2\left(\frac{6}{\eta}\right)\left\{\frac{64}{\lambda}\left[\frac{M^2 B_K}{l^2\lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l}\right] + \frac{24}{\lambda^2}\left[\frac{4B_K^2 \mathcal{B}(\lambda)}{l^2} + \frac{B_K \mathcal{A}(\lambda)}{l}\right]\right\} + \mathcal{B}(\lambda) + \|f_\rho\|_{\mathcal{H}}^2\right)\right]$$
$$= \frac{4L^2\left(1+\sqrt{\alpha}\right)^{2h}(2B_k)^h}{\lambda N^h}\left[C^2 + 4B_K \times \right.$$
$$\left. \times \left(\log^2\left(\frac{6}{\eta}\right)\left\{\frac{64}{\lambda}\left[\frac{M^2 B_K}{l^2\lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{l}\right] + \frac{24}{\lambda^2}\left[\frac{4B_K^2 \mathcal{B}(\lambda)}{l^2} + \frac{B_K \mathcal{A}(\lambda)}{l}\right]\right\} + \mathcal{B}(\lambda) + \|f_\rho\|_{\mathcal{H}}^2\right)\right].$$

By taking into account Caponnetto and De Vito (2007)'s bounds for S_1 and S_2, $S_1 \leq 32\log^2\left(\frac{6}{\eta}\right)\left[\frac{B_K M^2}{l^2\lambda} + \frac{\Sigma^2\mathcal{N}(\lambda)}{l}\right]$, $S_2 \leq 8\log^2\left(\frac{6}{\eta}\right)\left[\frac{4B_K^2\mathcal{B}(\lambda)}{l^2\lambda} + \frac{B_K\mathcal{A}(\lambda)}{l\lambda}\right]$, plugging all the expressions to (21), we obtain Theorem 2 with a union bound.

7.2 Proofs of the Misspecified Case

We present the proof details concerning the excess risk in the misspecified case (Theorem 7).

7.2.1 Proof of the bound on $\sqrt{S_{-1}} + \sqrt{S_0}$ without $\mathcal{P}(b, c)$

The upper bounds on S_{-1} and S_0 [which are defined in Eqs. (15), (16)] remain valid without modification provided that (i) $\mathbf{\Theta}(\lambda, \mathbf{z}) = \|(T - T_{\mathbf{x}})(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \|(T - T_{\mathbf{x}})(T+\lambda)^{-1}\|_{\mathcal{L}_2(\mathcal{H})} \leq \frac{1}{2}$, where we used Eq. (1), (ii) Eq. (24) is satisfied (which has probability $1 - le^{-\alpha}$) and (iii) Eq. (25) holds. Our goal below is to guarantee the $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ condition with high probability *without* assuming that the prior belongs to $\mathcal{P}(b, c)$.
Requirement $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$: Let us define $\xi_i = T_{\mu_{\mathbf{x}_i}}(T+\lambda)^{-1} \in \mathcal{L}_2(\mathcal{H})$, $(i = 1, \ldots, l)$. With this choice we get $\mathbb{E}[\xi_i] = T(T+\lambda)^{-1}$, $(T - T_{\mathbf{x}})(T+\lambda)^{-1} = \mathbb{E}[\xi_i] - \frac{1}{l}\sum_{i=1}^{l}\xi_i$ and

$$\|\xi_i\|_{\mathcal{L}_2(\mathcal{H})} \leq \left\|T_{\mu_{\mathbf{x}_i}}\right\|_{\mathcal{L}_2(\mathcal{H})}\left\|(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \leq B_K/\lambda \;\Rightarrow\; \mathbb{E}\left[\,\|\xi_i\|_{\mathcal{L}_2(\mathcal{H})}^2\,\right] \leq (B_K)^2/\lambda^2,$$

where we made use of (2), the $\|T_{\mu_{\mathbf{x}_i}}\|_{\mathcal{L}_2(\mathcal{H})} \leq B_K$ identity following from the boundedness of K (Caponnetto and De Vito, 2007, page 341, Eq. (13)), and the spectral theorem. Consequently, by the Bernstein's inequality (Lemma 7.3.1 with $\mathcal{K} = \mathcal{L}_2(\mathcal{H})$, $B = 2B_K/\lambda$, $\sigma = B_K/\lambda$) we obtain that for $\forall\eta \in (0, 1)$

$$\mathbb{P}\left(\left\|(T - T_{\mathbf{x}})(T+\lambda)^{-1}\right\|_{\mathcal{L}_2(\mathcal{H})} \leq 2\left(\frac{2B_K}{\lambda l} + \frac{B_K}{\sqrt{l}\lambda}\right)\log\left(\frac{2}{\eta}\right)\right) \geq 1 - \eta.$$

Thus, for $\mathbf{\Theta}(\lambda, \mathbf{z}) \leq \frac{1}{2}$ with probability $1 - \eta$ it is sufficient to have

$$2\left(\frac{2B_K}{\lambda l} + \frac{B_K}{\sqrt{l}\lambda}\right)\log\left(\frac{2}{\eta}\right) \leq \frac{6B_K}{\sqrt{l}\lambda}\log\left(\frac{2}{\eta}\right) \leq \frac{1}{2} \Leftrightarrow \left[\frac{12B_K}{\lambda}\log\left(\frac{2}{\eta}\right)\right]^2 \leq l. \qquad (38)$$

Under these conditions, we arrived at the upper bound

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

where as opposed to Section 7.1.3 and Eq. (23) we used a slightly cruder $\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2 \leq \frac{C^2}{\lambda}$ bound; it holds without the $\mathcal{P}(b,c)$ assumption by the definition of $f_{\mathbf{z}}^{\lambda}$ and the boundedness of y since $\lambda\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2 \leq \frac{1}{l}\sum_{i=1}^{l}\|y_i\|_Y^2 \leq C^2$.

Remark: Notice that the price we pay for not assuming that the prior belongs to the $\mathcal{P}(b,c)$ class ($b>1$) is a slightly tighter $\frac{1}{\lambda^2} \leq l$ constraint [Eq. (38)] instead of $\frac{1}{\lambda^{1+\frac{1}{b}}} \leq l$ in Eq. (19), and a somewhat looser $\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2$ bound.

7.2.2 Proof of the decomposition of $\left\|\sqrt{T}\left(f_{\mathbf{z}}^{\lambda}-f^{\lambda}\right)\right\|_{\mathcal{H}}$

Using the analytical formula of $f_{\mathbf{z}}^{\lambda}$ [see Eq. (13)] and that of f^{λ} [see Eq.(17)]

$$f^{\lambda} = (S_K S_K^* + \lambda I)^{-1} S_K f_\rho = (T+\lambda I)^{-1} S_K f_\rho \tag{39}$$

one gets $(T+\lambda I)f^{\lambda} = S_K f_\rho \Rightarrow \lambda f^{\lambda} = S_K f_\rho - T f^{\lambda}$ and

$$\begin{aligned} f_{\mathbf{z}}^{\lambda} - f^{\lambda} &= (T_{\mathbf{x}}+\lambda I)^{-1} g_{\mathbf{z}} - f^{\lambda} = (T_{\mathbf{x}}+\lambda I)^{-1} g_{\mathbf{z}} - (T_{\mathbf{x}}+\lambda I)^{-1}(T_{\mathbf{x}}+\lambda I) f^{\lambda} \\ &= (T_{\mathbf{x}}+\lambda I)^{-1}\left[g_{\mathbf{z}} - (T_{\mathbf{x}}+\lambda I) f^{\lambda}\right] = (T_{\mathbf{x}}+\lambda I)^{-1}\left(g_{\mathbf{z}} - T_{\mathbf{x}} f^{\lambda} - \lambda f^{\lambda}\right) \\ &= (T_{\mathbf{x}}+\lambda I)^{-1}\left(g_{\mathbf{z}} - T_{\mathbf{x}} f^{\lambda} - S_K f_\rho + T f^{\lambda}\right) \\ &= (T_{\mathbf{x}}+\lambda I)^{-1}\left(g_{\mathbf{z}} - S_K f_\rho\right) + (T_{\mathbf{x}}+\lambda I)^{-1}(T - T_{\mathbf{x}}) f^{\lambda} \\ &= (T_{\mathbf{x}}+\lambda I)^{-1}\left(g_{\mathbf{z}} - S_K f_\rho\right) + (T_{\mathbf{x}}+\lambda I)^{-1}(T - T_{\mathbf{x}})(T+\lambda I)^{-1} S_K f_\rho. \end{aligned} \tag{40}$$

Let us rewrite $(T+\lambda I)^{-1}$ by the $(A+UV)^{-1} = A^{-1} - A^{-1}U\left(I+VA^{-1}U\right)^{-1}VA^{-1}$ operator Woodbury formula (Ding and Zhou, 2008, Theorem 2.1, page 724)

$$\begin{aligned} (T+\lambda I)^{-1} &= (\lambda I + S_K S_K^*)^{-1} = (\lambda^{-1} I) - (\lambda^{-1} I) S_K\left[I + S_K^*(\lambda^{-1} I) S_K\right]^{-1} S_K^*(\lambda^{-1} I) \\ &= (\lambda^{-1} I) - \lambda^{-1} S_K (\lambda I + \tilde{T})^{-1} S_K^*. \end{aligned}$$

By the derived expression for $(T+\lambda I)^{-1}$, we get $(T+\lambda I)^{-1} S_K f_\rho = \lambda^{-1} S_K f_\rho - \lambda^{-1} S_K (\lambda I + \tilde{T})^{-1} S_K^* S_K f_\rho = \lambda^{-1} S_K\left[f_\rho - (\tilde{T}+\lambda I)^{-1} S_K^* S_K f_\rho\right]$. Plugging this result to Eq. (40), introducing the $g_\rho = S_K f_\rho$ notation, using the triangle inequality we get

$$\begin{aligned} &\left\|\sqrt{T}\left(f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right)\right\|_{\mathcal{H}} = \\ &= \left\|\sqrt{T}(T_{\mathbf{x}}+\lambda I)^{-1}\left\{\left(g_{\mathbf{z}} - S_K f_\rho\right) + (T - T_{\mathbf{x}})\lambda^{-1} S_K\left[f_\rho - (\tilde{T}+\lambda I)^{-1} S_K^* S_K f_\rho\right]\right\}\right\|_{\mathcal{H}} \\ &\leq \left\|\sqrt{T}(T_{\mathbf{x}}+\lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}\Big(\left\|g_{\mathbf{z}} - g_\rho\right\|_{\mathcal{H}} + \left\|T - T_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})}\lambda^{-1}\left\|S_K\left[f_\rho - (\tilde{T}+\lambda I)^{-1} S_K^* S_K f_\rho\right]\right\|_{\mathcal{H}}\Big). \end{aligned}$$

7.2.3 Proof of the bound on $\|g_{\mathbf{z}} - g_\rho\|_{\mathcal{H}}$

As is known $g_{\mathbf{z}} = \frac{1}{l}\sum_{i=1}^{l} K_{\mu_{x_i}} y_i$ [see Eq. (13)] and $g_\rho = \int_X K_{\mu_x} f_\rho(\mu_x) \mathrm{d}\rho_X(\mu_x)$ (Caponnetto and De Vito, 2007, Eq. (23), page 344). Let $\xi_i = K_{\mu_{x_i}} y_i \in \mathcal{H}$ $(i = 1, \ldots, l)$. In this case $\mathbb{E}[\xi_i] = g_\rho$, $g_\rho - g_{\mathbf{z}} = \mathbb{E}[\xi_i] - \frac{1}{l}\sum_{i=1}^{l} \xi_i$, and $\|\xi_i\|_{\mathcal{H}}^2 = \left\|K_{\mu_{x_i}} y_i\right\|_{\mathcal{H}}^2 \leq \left\|K_{\mu_{x_i}}\right\|_{\mathcal{L}(Y,H)}^2 \|y_i\|_Y^2 \leq B_K C^2 \Rightarrow \|\xi_i\|_{\mathcal{H}} \leq C\sqrt{B_K} \Rightarrow \mathbb{E}\left[\,\|\xi_i\|_{\mathcal{H}}^2\,\right] \leq C^2 B_K$ using the boundedness of kernel K ($\left\|K_{\mu_{x_i}}\right\|_{\mathcal{L}(Y,H)}^2 \leq B_K$) and the boundedness of output y ($\|\,|y|\,\|_Y \leq C$). Applying the Bernstein inequality (see Lemma 7.3.1 with $\mathcal{K} = \mathcal{H}$, $B = 2C\sqrt{B_K}$, $\sigma = C\sqrt{B_K}$) one gets that for any $\eta \in (0, 1)$

$$\mathbb{P}\left(\|g_{\mathbf{z}} - g_\rho\|_{\mathcal{H}} \leq 2\left(\frac{2C\sqrt{B}_K}{l} + \frac{C\sqrt{B_K}}{\sqrt{l}}\right)\log\left(\frac{2}{\eta}\right)\right) \geq 1 - \eta.$$

7.2.4 Proof of the bound on $\|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})}$

Let $\xi_i = T_{\mu_{x_i}} \in \mathcal{L}_2(\mathcal{H})$ $(i = 1, \ldots, l)$, then $\mathbb{E}[\xi_i] = T$, $T - T_{\mathbf{x}} = T - \frac{1}{l}\sum_{i=1}^{l} T_{\mu_{x_i}}$, $\|\xi_i\|_{\mathcal{L}_2(\mathcal{H})} = \|T_{\mu_{x_i}}\|_{\mathcal{L}_2(\mathcal{H})} \leq B_K$, $\mathbb{E}\left[\|\xi_i\|_{\mathcal{L}_2(\mathcal{H})}^2\right] \leq B_K^2$. Applying the $\|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq \|T - T_{\mathbf{x}}\|_{\mathcal{L}_2(\mathcal{H})}$ relation [see Eq. (1)] and the Bernstein inequality (see Lemma 7.3.1 with $\mathcal{K} = \mathcal{L}_2(\mathcal{H})$, $B = 2B_K$, $\sigma = B_K$), we obtain that for any $\eta \in (0, 1)$

$$\mathbb{P}\left(\|T - T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq 2\left(\frac{2B_K}{l} + \frac{\sigma}{\sqrt{l}}\right)\log\left(\frac{2}{\eta}\right)\right) \geq 1 - \eta.$$

7.2.5 Proof of the decomposition of $\left\|S_K\big(f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\big)\right\|_{\mathcal{H}}^2$

Since $\|S_K a\|_{\mathcal{H}}^2 = \langle S_K a, S_K a\rangle_{\mathcal{H}} = \langle S_K^* S_K a, a\rangle_\rho = \left\langle \tilde{T} a, a\right\rangle_\rho$ $(\forall a \in L_{\rho_X}^2)$ by the definition of the adjoint operator and $\tilde{T} = S_K^* S_K$ [see Eq. (6)], we can rewrite the target term as

$$\begin{aligned}
\left\|S_K\left[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right]\right\|_{\mathcal{H}}^2 = \\
&= \left\langle \tilde{T}\left[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right], f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right\rangle_\rho \\
&\leq \left\|\tilde{T}\left[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right]\right\|_\rho \left\|f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\right\|_\rho,
\end{aligned}$$

where the CBS (Cauchy-Bunyakovsky-Schwarz) inequality was applied. Since

$$\begin{aligned}
(S_K^* S_K + \lambda I) S_K^* = S_K^*(S_K S_K^* + \lambda I) \qquad S_K^*(S_K S_K^* + \lambda I)^{-1} &= (S_K^* S_K + \lambda I)^{-1} S_K^* \\
S_K^*(S_K S_K^* + \lambda I)^{-1} S_K = (S_K^* S_K + \lambda I)^{-1} S_K^* S_K &\qquad (41) \\
S_K^*(T + \lambda I)^{-1} S_K = (\tilde{T} + \lambda I)^{-1} \tilde{T} &\qquad (42)
\end{aligned}$$

using Eq. (41) and the analytical expression for f^λ [see Eq. (39)] we have

$$\begin{aligned}
(\tilde{T} + \lambda I)^{-1} \tilde{T} f_\rho = (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho = \left(S_K^* S_K + \lambda I\right)^{-1} S_K^* S_K f_\rho \\
= S_K^*(S_K S_K^* + \lambda I)^{-1} S_K f_\rho = S_K^* f^\lambda \qquad (43)
\end{aligned}$$

and $\left\|S_K\big[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\big]\right\|_{\mathcal{H}}^2 \leq \left\|\tilde{T}\big[f_\rho - (\tilde{T} + \lambda I)^{-1} S_K^* S_K f_\rho\big]\right\|_\rho \left\|S_K^* f^\lambda - f_\rho\right\|_\rho$.

7.2.6 Proof of the bound on $\left\|S_K^* f^\lambda - f_\rho\right\|_\rho$

Let us apply (i) the $Af - f = Af - f - q' + q' = (A - I)(f - q') + Aq' - q'$ relation with $A = (\tilde{T} + \lambda I)^{-1}\tilde{T}$, $f = f_\rho$ and $q' = S_K^* q$, where $q \in \mathcal{H}$ is an arbitrary element from $\mathcal{H}$, (ii) Eq. (43) and (iii) the triangle inequality to arrive at

$$\begin{aligned}\left\|S_K^* f^\lambda - f_\rho\right\|_\rho &= \left\|(\tilde{T} + \lambda I)^{-1}\tilde{T} f_\rho - f_\rho\right\|_\rho \\ &= \left\|\left[(\tilde{T} + \lambda I)^{-1}\tilde{T} - I\right](f_\rho - S_K^* q) + (\tilde{T} + \lambda I)^{-1}\tilde{T} S_K^* q - S_K^* q\right\|_\rho \\ &\leq \left\|\left[(\tilde{T} + \lambda I)^{-1}\tilde{T} - I\right](f_\rho - S_K^* q)\right\|_\rho + \left\|(\tilde{T} + \lambda I)^{-1}\tilde{T} S_K^* q - S_K^* q\right\|_\rho.\end{aligned}$$

Below we give upper bounds on these two terms.

First, notice that $\mu_x \in X \mapsto \|K(\mu_x, \mu_x)\|_{\mathcal{L}(Y)} \leq B_K$. This boundedness with the strong continuity of $K_{(\cdot)}$ imply (Carmeli et al., 2006, Proposition 12) that $\mathcal{H} \subseteq C(X, Y)$, i.e., K is a Mercer kernel. Since K_{μ_x} is a Hilbert-Schmidt operator for all $\mu_x \in X$ [see Eq. (11)], it is also a compact operator ($\forall \mu_x \in X$). The compactness of K_{μ_x}-s with the bounded and Mercer property of K guarantees the boundedness of S_K^* and that $\tilde{T}$ is a *compact*, positive, self-adjoint operator (Carmeli et al., 2010, Proposition 3).

Bound on $\|[(\tilde{T} + \lambda I)^{-1}\tilde{T} - I](f_\rho - S_K^* q)\|_\rho$: Since $\tilde{T}$ is a compact positive self-adjoint operator, by the spectral theorem (Steinwart and Christmann, 2008, Theorem 4.27, page 127) there exist an $(u_i)_{i\in I}$ countable ONB in $cl\big[Im(\tilde{T})\big]$, and $a_1 \geq a_2 \geq \ldots > 0$ such that $\tilde{T} f = \sum_{i\in I} a_i \langle f, u_i\rangle_\rho u_i$ ($\forall f \in L^2_{\rho_X}$) and let $(v_j)_{j\in J}$ (J is also countable by the separability[20] of $L^2_{\rho_X}$) an ONB in $Ker(\tilde{T}^*) = Ker(\tilde{T})$; $L^2_{\rho_X} = cl\big[Im(\tilde{T})\big] \oplus Ker(\tilde{T})$. Thus,

$$\begin{aligned}\left\|\left[(\tilde{T} + \lambda I)^{-1}\tilde{T} - I\right](f_\rho - S_K^* q)\right\|_\rho^2 &= \sum_{i\in I}\left(\frac{a_i}{a_i + \lambda} - 1\right)^2 \langle f_\rho - S_K^* q, u_i\rangle_\rho^2 + \sum_{j\in J}\langle f_\rho - S_K^* q, v_j\rangle_\rho^2 \\ &\leq \sum_{i\in I}\langle f_\rho - S_K^* q, u_i\rangle_\rho^2 + \sum_{j\in J}\langle f_\rho - S_K^* q, v_j\rangle_\rho^2 = \|f_\rho - S_K^* q\|_\rho^2\end{aligned}$$

exploiting the Parseval's identity and that $\left(\frac{\lambda_i}{\lambda_i + \lambda} - 1\right)^2 \leq 1$.

Bound on $\|(\tilde{T} + \lambda I)^{-1}\tilde{T} S_K^* q - S_K^* q\|_\rho$: By using Eq. (42), Eq. (32), and Lemma 7.3.2 ($M = T = S_K S_K^*$, $\mathcal{K} = \mathcal{H}$, $f = q$, $a = \frac{1}{2}$), the target quantity can be bounded as

$$\begin{aligned}\left\|(\tilde{T} + \lambda I)^{-1}\tilde{T} S_K^* q - S_K^* q\right\|_\rho &= \left\|S_K^*(T + \lambda I)^{-1} S_K S_K^* q - S_K^* q\right\|_\rho \\ &= \left\|\sqrt{T}\left[(T + \lambda I)^{-1} S_K S_K^* q - q\right]\right\|_{\mathcal{H}} \\ &= \left\|\sqrt{T}\left[(T + \lambda I)^{-1} T q - q\right]\right\|_{\mathcal{H}} \leq \max\left(1, \|T\|_{\mathcal{L}(\mathcal{H})}\right)\lambda^{\frac{1}{2}}\|q\|_{\mathcal{H}}.\end{aligned}$$

Making use of the two derived bounds, we get $\left\|S_K^* f^\lambda - f_\rho\right\|_\rho \leq \|f_\rho - S_K^* q\|_\rho + \max\left(1, \|T\|_{\mathcal{L}(\mathcal{H})}\right)\lambda^{\frac{1}{2}}\|q\|_{\mathcal{H}}$.

20. $L^2_{\rho_X} = L^2(X, \mathcal{B}(H)|_X, \rho_X; Y)$ is isomorphic to $L^2(X, \mathcal{B}(H)|_X, \rho_X; \mathbb{R}) \otimes Y$, where $\otimes$ is the tensor product of Hilbert spaces. The separability follows from that of Y and $L^2(X, \mathcal{B}(H)|_X, \rho_X; \mathbb{R})$; the latter holds (Cohn, 2013, Proposition 3.4.5) since $\mathcal{B}(H)|_X$ is countably generated since $X \subseteq H$ is separable.

7.3 Supplementary Lemmas

In this section, we list two lemmas used in the proofs.

7.3.1 Bernstein's inequality (Caponnetto and De Vito, 2007, Prop. 2, p. 345)

Let ξ_i $(i = 1, \dots, l)$ be i.i.d. realizations of a random variable on a $(\Omega, \mathcal{A}, P)$ probability space with values in a separable Hilbert space $\mathcal{K}$. If there exist $B > 0$, $\sigma > 0$ constants such that $\|\xi(\omega)\|_{\mathcal{K}} \leq \frac{B}{2}$ a.s., $\mathbb{E}\left[\|\xi\|_{\mathcal{K}}^2\right] \leq \sigma^2$, then for all $l \geq 1$ and $\eta \in (0,1)$ we have

$$\mathbb{P}\left(\left\|\frac{1}{l}\sum_{i=1}^{l}\xi_i - \mathbb{E}[\xi_1]\right\|_{\mathcal{K}} \leq 2\left(\frac{B}{l} + \frac{\sigma}{\sqrt{l}}\right)\log\left(\frac{2}{\eta}\right)\right) \geq 1 - \eta.$$

7.3.2 Lemma on bounded, self-adjoint compact operators; Sriperumbudur et al. (2014, Proposition A.2, page 39)

Let M be a bounded, self-adjoint compact operator on a separable Hilbert space $\mathcal{K}$. Let $a \geq 0$, $\lambda > 0$, and $s \geq 0$. Let $f \in \mathcal{K}$ such that $f \in Im\left(M^s\right)$. If $s + a > 0$, then

$$\left\|M^a\left[(M+\lambda I)^{-1}Mf - f\right]\right\|_{\mathcal{K}} \leq \max\left(1, \|M\|_{\mathcal{L}(\mathcal{K})}^{s+a-1}\right)\lambda^{\min(1,s+a)}\left\|M^{-s}f\right\|_{\mathcal{K}}.$$

Note: specifically for $s = 0$ we have $Im\left(M^s\right) = Im\left(I\right) = \mathcal{K}$, in other words, there is no additional range space constraint.

8. Discussion of Our Assumptions

We give a short insight into the consequences of our assumptions (detailed in Section 3) and present some concrete examples.

- **Well-definedness of ρ:** The boundedness and continuity of k imply the measurability of $\mu : (\mathcal{M}_1^+(\mathcal{X}), \mathcal{B}(\tau_w)) \to (H, \mathcal{B}(H))$. Let τ denote the open sets on $H = H(k)$, $\tau|_X = \{A \cap X : A \in \tau\}$ the subspace topology on X, and $\mathcal{B}(H)|_X = \{A \cap X : A \in \mathcal{B}(H)\}$ the subspace σ-algebra on X. By noting (Schwartz, 1998, Corollary 5.2.13) that $\mathcal{B}\left(\tau|_X\right) = \mathcal{B}(H)|_X = \{A \in \mathcal{B}(H) : A \subseteq X\} \subseteq \mathcal{B}(H)$, the H-measurability of μ guarantees the measurability of $\mu : (\mathcal{M}_1^+(\mathcal{X}), \mathcal{B}(\tau_w)) \to (X, \mathcal{B}(H)|_X)$, and hence the well-definedness of ρ, the measure induced by $\mathcal{M}$ on $X \times Y$; for further details see (Szabó et al., 2015, Section A.1.1).[21]

- **Separability of X:** separability of $\mathcal{X}$ and the continuity of k implies the separability of $H = H(k)$ (Steinwart and Christmann, 2008, Lemma 4.33, page 130). Also, since $X \subseteq H$, X is separable.

- **Finiteness of B_k:** If $\mathcal{X}$ is compact, then the continuity of k implies $B_k < \infty$.

- **Finiteness of B_K, compact metricness of X:** Let $\mathcal{X}$ be a compact metric space. In this case $\mathcal{M}_1^+(\mathcal{X})$ is also compact metric (Parthasarathy, 1967, Theorem 6.4, page 55).

21. Note that the referred proof also holds for separable Hilbert Y, and by the simplified reasoning above the original $X \in \mathcal{B}(H)$ condition could be avoided.

Hence if $\mu : (\mathcal{M}_1^+(\mathcal{X}), \tau_w) \to H(k)$ is continuous[22] (not just measurable), then X is compact metric and thus by the Hölder property of $K_{(\cdot)}$, it is continuous implying that $B_K < \infty$.

- K **properties:** It is known (Caponnetto and De Vito, 2007, page 339-340) that

$$K(\mu_a, \mu_b) = K_{\mu_a}^* K_{\mu_b}, \quad (\forall \mu_a, \mu_b \in X) \tag{44}$$

$$\|K_{\mu_a}\|_{\mathcal{L}(Y,\mathcal{H})} = \left\|K_{\mu_a}^*\right\|_{\mathcal{L}(\mathcal{H},Y)} \leq \sqrt{B_K}, \quad (\forall \mu_a \in X). \tag{45}$$

  Remark: In terms of Eq. (44), the Eq. (11) assumption means that the $\{K(\mu_a, \mu_a)\}_{\mu_a \in X}$ operators are trace class, specifically they are compact operators.

- **Separability of $\mathcal{H}$:** The separability of X and the continuity of K imply the separability of $\mathcal{H}$. Indeed, since $\mu_a \mapsto K_{\mu_a}$ is Hölder continuous w.r.t. the Hilbert-Schmidt norm it is also continuous. As a result it is continuous w.r.t. the operator norm, and thus also w.r.t. the strong topology. Using this property with the finiteness of B_K the separability of $\mathcal{H}$ follows (Carmeli et al., 2006, Proposition 5.1, Corollary 5.2).

- Our assumptions imply Caponnetto and De Vito (2007)'s conditions (not considering the $\mathcal{P}(b, c)$ prior requirement). Indeed
  1. Y is a separable Hilbert space by assumption; the same property also holds for $\mathcal{H}$ as we have seen.
  2. The measurability of $(\mu_x, \mu_t) \mapsto \langle K_{\mu_x} w, K_{\mu_t} v \rangle_{\mathcal{H}}$ for $\forall w, v \in Y$ is guaranteed by the continuity of $K_{(\cdot)}$ w.r.t. the strong topology.
  3. We have $\int_{X \times Y} \|y\|_Y^2 \, \mathrm{d}\rho_X(\mu_x, y) \leq \int_{X \times Y} C^2 \mathrm{d}\rho_X(\mu_x, y) = C^2 < \infty$ due to the boundedness of y, and hence $\exists \Sigma > 0, \exists M > 0$ such that for ρ_X-almost $\mu_x \in X$

$$\int_Y \|y - f_\rho(\mu_x)\|_Y^m \, \mathrm{d}\rho(y|\mu_x) \leq \frac{m! \Sigma^2 M^{m-2}}{2} \quad (\forall m \geq 2). \tag{46}$$

     Indeed, by (Caponnetto and De Vito, 2007, Eq. (33)) the Bernstein condition (46) holds if $\|y - f_\rho(\mu_x)\|_Y \leq \frac{M}{2}$, $\int_Y \|y - f_\rho(\mu_x)\|_Y^2 \, \mathrm{d}\rho(y|\mu_x) \leq \Sigma^2$. In our case using the boundedness of y, the regression function is also bounded and the same holds for $\|y - f_\rho(\mu_x)\|_Y$ by the triangle inequality: $\|y - f_\rho(\mu_x)\|_Y \leq C + \|f_\rho\|_{\mathcal{H}} \sqrt{B_K}$; thus, $M = 2(C + \|f_\rho\|_{\mathcal{H}} \sqrt{B_K})$, $\Sigma = \frac{M}{2}$ is a suitable choice.
  4. The Polishness of $X \times Y$ was used by Caponnetto and De Vito (2007) to assure the existence of $\rho(y|\mu_a)$; we guaranteed this existence under somewhat milder conditions (see footnote 7).

Real-valued outputs: We now consider the specific case of $Y = \mathbb{R}$, when the following simplifications and results hold. By noting that in this case $Tr(K_{\mu_a}^* K_{\mu_a}) = K(\mu_a, \mu_a)$, Eq. (11) simplifies to the boundedness of kernel K in the traditional sense

$$K(\mu_a, \mu_a) \leq B_K \quad (\forall \mu_a \in X). \tag{47}$$

22. For example, if k is universal, then μ metrizes the weak topology τ_w (Sriperumbudur et al., 2010, Theorem 23, page 1552), hence μ is continuous.

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

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

Eq. (12) reduces to the Hölder continuity of the canonical feature map $\Psi_K(\mu_c) := K(\cdot,\mu_c) : X \to \mathcal{H}$, in other words $\exists L>0$, $h\in(0,1]$ such that $\|\Psi_K(\mu_a)-\Psi_K(\mu_b)\|_{\mathcal{H}} \leq L\|\mu_a-\mu_b\|_H^h$, $\forall(\mu_a,\mu_b)\in X\times X$. In case of a linear kernel, $K(\mu_a,\mu_b)=\langle\mu_a,\mu_b\rangle_H$, $(\mu_a,\mu_b\in X)$, the Hölder continuity of Ψ_K holds with $L=1$, $h=1$, and $B_K=B_k$ is a suitable choice. Evaluating the kernel K at the empirical embeddings $\mu_{\hat{x}_i}=\int_{\mathcal{X}}k(\cdot,u)\mathrm{d}\hat{x}_i(u)=\frac{1}{N}\sum_{n=1}^N k(\cdot,x_{i,n})\in H$ yields the standard set kernel

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

by the bilinearity of $\langle\cdot,\cdot\rangle_H$ and the reproducing property of k.

Remark: One can define many nonlinear kernels (see Table 1) on mean embedded distributions. These kernels are the natural extensions to distributions of the Gaussian (Christmann and Steinwart, 2010), exponential, Cauchy, generalized t-student and inverse multiquadric kernels. If $\mathcal{X}$ is a compact metric space and μ is continuous, then the Ψ_K canonical feature maps, associated to K-s in Table 1, can be shown to satisfy our Hölder continuity requirement [Eq. (12)]; for details, see (Szabó et al., 2015, Section A.1.5-A.1.6).

Acknowledgments

We would like to thank the anonymous reviewers for their highly valuable, constructive suggestions to improve the manuscript. This work was supported by the Gatsby Charitable Foundation, NSF grant 1247658, and DOE grant DE-SC001114. A part of 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.

References

Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, pages 775–783, 2015.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

Yasemin Altun and Alexander Smola. Unifying divergence minimization and statistical inference via convex duality. In *Conference on Learning Theory (COLT)*, pages 139–153, 2006.

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4:195–266, 2011.

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

Peter Auer. Approximating hyper-rectangles: Learning and pseudorandom sets. *Journal of Computer and System Sciences*, 57:376–388, 1998.

Boris Babenko. Multiple instance learning: Algorithms and applications. Technical report, Department of Computer Science and Engineering, University of California, San Diego, 2004. (`http://cms.brookes.ac.uk/research/visiongroup/talks/rg_dec_11_09/bbabenko_re.pdf`).

Boris Babenko, Nakul Verma, Piotr Dollár, and Serge Belongie. Multiple instance learning with manifold bags. In *International Conference on Machine Learning (ICML)*, pages 81–88, 2011.

Charles Bergeron, Jed Zaretzki, Curt Breneman, and Kristin P. Bennett. Multiple instance ranking. In *International Conference on Machine Learning (ICML)*, pages 48–55, 2008.

Charles Bergeron, Gregory Moore, Jed Zaretzki, Curt M. Breneman, and Kristin P. Bennett. Fast bundle algorithm for multiple-instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1068–1079, 2012.

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

Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30:23–29, 1998.

Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5:216–233, 2015.

Céline Brouard, Florence d'Alché Buc, and Marie Szafranski. Semi-supervised penalized output kernel regression for link prediction. In *International Conference on Machine Learning (ICML)*, pages 593–600, 2011.

Andrea Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, Massachusetts Institute of Technology, 2006. (`http://www6.cityu.edu.hk/ma/doc/people/caponnettoa/regop_TR%28TR11%29.pdf`).

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

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4: 377–408, 2006.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

Kevin M. Carter, Raviv Raich, William G. Finn, and Alfred O. Hero. Information-geometric dimensionality reduction. *IEEE Signal Processing Magazine*, 28:89–99, 2011.

Jing Chai, Hongtao Chen, Lixia Huang, and Fanhua Shang. Maximum margin multiple-instance feature weighting. *Pattern Recognition*, 47:2091–2103, 2014a.

Jing Chai, Xinghao Ding, Hongtao Chen, and Tingyu Li. Multiple-instance discriminant analysis. *Pattern Recognition*, 47:2517–2531, 2014b.

Ying Chen and Ou Wu. Contextual Hausdorff dissimilarity for multi-instance clustering. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 870–873, 2012.

Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414, 2010.

Donald L. Cohn. *Measure Theory: Second Edition*. Birkhäuser Basel, 2013.

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

Ernesto de Vito, Lorenzo Rosasco, and Andrea Caponnetto. Discretization error analysis for Tikhonov regularization. *Analysis and Applications*, 4:81–99, 2006.

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.

Jiu Ding and Aihui Zhou. A spectrum theorem for perturbed bounded linear operators. *Applied Mathematics and Computation*, 201:723–728, 2008.

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.

Gerald Edgar. *Measure, Topology and Fractal Geometry*. Springer-Verlag, 1995.

Thomas Eiter and Heikki Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34:109–133, 1997.

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Seth Flaxman, Yu-Xiang Wang, and Alex Smola. Who supported Obama in 2012? ecological inference through distribution regression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 289–298, 2015.

James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25:1–25, 2010.

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.

Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002.

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.

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

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`).

Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143, 2005.

Yang Hu, Mingjing Li, and Nenghai Yu. Multiple-instance ranking: Learning to rank images for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

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

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, and Manuel Davy. Nonlinear functional regression: a functional RKHS approach. *International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 9:374–380, 2010.

Hachem Kadri, Alain Rakotomamonjy, Francis Bach, and Philippe Preux. Multiple operator-valued kernel learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2429–2437, 2012.

Hachem Kadri, Mohammad Ghavamzadeh, and Philippe Preux. A generalized kernel approach to structured output learning. *International Conference on Machine Learning (ICML; JMLR W&CP)*, 28:471–479, 2013.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17:1–54, 2016.

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *International Conference on Machine Learning (ICML)*, pages 361–368, 2003.

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`).

James T. Kwok and Pak-Ming Cheung. Marginalized multi-instance kernels. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 901–906, 2007.

Philip M. Long and Lei Tan. PAC learning of axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30:7–21, 1998.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. *International Conference on Machine Learning (ICML; JMLR W&CP)*, 37:1452–1461, 2015.

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.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38:526–565, 2010.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18, 2012.

Hans-Georg Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32:223–240, 2005.

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.

Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. *International Conference on Machine Learning (ICML; JMLR W&CP)*, 28:1049–1057, 2013.

Junier Oliva, William Neiswanger, Barnabás Póczos, Eric Xing, Hy Trac, Shirley Ho, and Jeff Schneider. Fast function to function regression. *International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 38:717–725, 2015.

Junier B. Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. Fast distribution to real regression. *International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 33:706–714, 2014.

Kalyanapuram R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

George Pedrick. Theory of reproducing kernels for Hilbert spaces of vector valued functions. Technical report, 1957.

Wei Ping, Ye Xu, Kexin Ren, Chi-Hung Chi, and Shen Furao. Non-I.I.D. multi-instance dimensionality reduction by learning a maximum bag margin subspace. In *AAAI Conference on Artificial Intelligence*, pages 551–556, 2010.

Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Uncertainty in Artificial Intelligence (UAI)*, pages 599–608, 2011.

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

Barnabás Póczos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. *International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 31:507–515, 2013.

Jan Ramon and Maurice Bruynooghe. A polynomial time computable metric between point sets. *Acta Informatica*, 37:765–780, 2001.

James O. Ramsay and Bernard W. Silverman. *Applied Functional Data Analysis*. Springer Verlag, New York, 2002.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer Verlag, New York, 2005.

Soumya Ray and David Page. Multiple instance regression. In *International Conference on Machine Learning (ICML)*, pages 425–432, 2001.

Vikas C. Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R. Bharat Rao. Bayesian multiple instance learning: Automatic feature selection and inductive transfer. In *International Conference on Machine Learning (ICML)*, pages 808–815, 2008.

Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17:1–25, 2016.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2008.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1648–1656, 2015.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13:2999–3039, 2012.

Laurent Schwartz. *Analyse III, Calcul Integrél*. Hermann, 1998.

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

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.

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.

Bharath K. Sriperumbudur, Kenji Fukumizu, Revant Kumar, Arthur Gretton, and Aapo Hyvärinen. Density estimation in infinite dimensional exponential families. Technical report, 2014. (`http://arxiv.org/pdf/1312.3516`).

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

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417, 2012.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory (COLT)*, 2009.

Hongwei Sun and Qiang Wu. Application of integral operator for regularized least-square regression. *Mathematical and Computer Modelling*, 49:276–285, 2009a.

Hongwei Sun and Qiang Wu. A note on application of integral operator in learning theory. *Applied and Computational Harmonic Analysis*, 26:416–421, 2009b.

Xu Sun, Hisashi Kashima, and Naonori Ueda. Large-scale personalized human activity recognition using online multitask learning. *IEEE Transactions on Knowledge and Data Engine*, 25:2551–2563, 2013.

Yu-Yin Sun, Michael K. Ng, and Zhi-Hua Zhou. Multi-instance dimensionality reduction. In *AAAI Conference on Artificial Intelligence*, pages 587–592, 2010.

Dougal J. Sutherland, Junier B. Oliva, Barnabás Póczos, and Jeff Schneider. Linear-time learning on distributions with approximate kernel embeddings. In *AAAI Conference on Artifical Intelligence (AAAI)*, pages 2073–2079, 2016.

Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 948–957, 2015.

Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

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.

Jun Wang and Jean-Daniel Zucker. Solving the multiple-instance problem: A lazy learning approach. In *International Conference on Machine Learning (ICML)*, pages 1119–1126, 2000.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

Ou Wu, Jun Gao, Weiming Hu, Bing Li, and Mingliang Zhu. Identifying multi-instance outliers. In *SIAM International Conference on Data Mining (SDM)*, pages 430–441, 2010.

Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Annals of Statistics*, 2016. (to appear; arXiv: `http://arxiv.org/abs/1501.06195`).

Amelia Zafra, Mykola Pechenizkiy, and Sebastián Ventura. HyDR-MI: A hybrid algorithm to reduce dimensionality in multiple instance learning. *Information Sciences*, 222:282–301, 2013.

Dan Zhang and Luo Si. Multiple instance transfer learing. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 406–411, 2009.

Dan Zhang, Fei Wang, Luo Si, and Tao Li. M3IC: Maximum margin multiple instance clustering. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1339–1344, 2009.

Dan Zhang, Fei Wang, Luo Si, and Tao Li. Maximum margin multiple instance clustering with applications to image and text clustering. *IEEE Transactions on Neural Networks*, 22:739–751, 2011.

Dan Zhang, Jingrui He, Luo Si, and Richard D. Lawrence. MILEAGE: Multiple Instance LEArning with Global Embedding. *International Conference on Machine Learning (ICML; JMLR W&CP)*, 28:82–90, 2013.

Min-Ling Zhang and Zhi-Hua Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31:47–68, 2009.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. 16:3299–3340, 2015.

Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013.

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.

Zhi-Hua Zhou. Multi-instance learning: A survey. Technical report, AI Lab, Department of Computer Science & Technology, Nanjing University, Nanjing, China, 2004. (`http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/techrep04.pdf`).

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-I.I.D. samples. In *International Conference on Machine Learning (ICML)*, pages 1249–1256, 2009.