Statistical Learning Theory for Distributional Classification

Christian Fiedler

Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich (TUM)
Munich Center for Machine Learning (MCML)
Institute for Data Science in Mechanical Engineering (DSME)*, RWTH Aachen University
christian.fiedler@tum.de

Abstract

In supervised learning with distributional inputs in the two-stage sampling setup, relevant to applications like learning-based medical screening or causal learning, the inputs (which are probability distributions) are not accessible in the learning phase, but only samples thereof. This problem is particularly amenable to kernel-based learning methods, where the distributions or samples are first embedded into a Hilbert space, often using kernel mean embeddings (KMEs), and then a standard kernel method like Support Vector Machines (SVMs) is applied, using a kernel defined on the embedding Hilbert space. In this work, we contribute to the theoretical analysis of this latter approach, with a particular focus on classification with distributional inputs using SVMs. We establish a new oracle inequality and derive consistency and learning rate results. Furthermore, for SVMs using the hinge loss and Gaussian kernels, we formulate a novel variant of an established noise assumption from the binary classification literature, under which we can establish learning rates. Finally, some of our technical tools like a new feature space for Gaussian kernels on Hilbert spaces are of independent interest.

1 Introduction

In supervised learning, distributions can appear as inputs, and this scenario has been considered in many works, cf. (Muandet et al. 2012) and the references therein. However, in some applications, the distributions acting as inputs are not directly accessible, but only samples thereof. For example, in the context of artificial intelligence (AI) assisted medical diagnosis, one might want to train on past patient data a binary classifier acting on biomarkers in order to detect an illness or anomaly. In practice, these biomarkers might not be fully observable, but only samples from them are available, e.g., through repeated measurements. In turn, we can model the biomarkers as distributions, which are only accessible through samples (Szabó et al. 2015). A concrete instance of this situation is training a classifier to detect atrial fibrillation from electrocardiogram measurements (Massiani et al. 2025). Another example can be found in statistical learning approaches to causal learning, where a classifier for the direction of causality between two random variables is desired, and such a causality classifier can be trained on samples from distributions with known causality structure (Lopez-Paz et al. 2015). The common feature of these examples is the *two-stage sampling* setup. First, a distribution (the input) and some output is sampled from a population, and then samples from the distribution acting as an input are drawn, and only these samples (and the output) are available during the learning phase. This setup has received particular attention in connection with kernel methods. Commonly, the distributions and the samples thereof are first embedded into a Hilbert space, and then a standard kernel method (now with inputs from a Hilbert space) is used on the transformed data set. The learned hypothesis can then be used on distributional inputs by composing it with the embedding (Szabó et al. 2015). This strategy has received a lot of attention, especially in the context of regression problems, where a Hilbertian embedding is combined with kernel ridge regression (KRR), and a substantial body of theory is available, including learning rates (Szabó et al. 2015, 2016). However, for some applications, other types of learning problems might be more appropriate, and a theoretical analysis should reflect this. For instance, the two examples described above are most naturally framed as classification problems, and (Massiani et al. 2025) actually used a classification SVM instead of KRR, with excellent empirical results.

Providing an analysis tailored to the classification setting is particularly important with regards to the assumptions used therein. As is well-known, in order to establish learning rates, one has to make distributional assumptions due to the *No Free Lunch-Theorem* (Steinwart and Christmann 2008, Chapter 6). For regression problems, typically a certain smoothness of the regression function is assumed, which is natural and reasonable in this setting. However, a smoothness assumption might not be appropriate for classification problems, since it does not necessarily capture the intrinsic difficulty (or simplicity) of a classification problem. Instead, margin and noise exponent assumptions are more appropriate, cf. (Steinwart and Scovel 2007) and (Steinwart and Christmann 2008, Chapter 8) for an overview and discussion. In turn, this calls for an analysis that can take these considerations into account. To the best of our knowledge, consistency and learning rate results for kernel-based distributional classification in the two-stage sampling setup under assumptions natural for the classification setting are still

*Majority of the work conducted while at DSME.

missing. In this work, we close this gap by providing a thorough theoretical investigation of SVMs with distributional inputs in the two-stage sampling setup.

Related Work After its introduction in (Póczos et al. 2013), learning in the two-stage sampling setup has been primarily investigated in the context of kernel methods, starting with (Szabó et al. 2015), which uses kernel mean embeddings (KMEs) for the Hilbertian embeddings of probability distributions. Note that (Muandet et al. 2012) is an earlier work that uses this embedding approach, but not in the two-stage sampling setup. While primarily KMEs have been used for the embeddings, other Hilbertian embeddings have been considered, like using sliced Wasserstein kernels (Meunier, Pontil, and Ciliberto 2022), and a variety of such embeddings are available, with (Bonnier, Oberhauser, and Szabó 2023) as a recent example, and even learned embeddings (Kachaiev and Recanatesi 2024). The main focus has been on regression, with KRR as the kernel method used after the embedding, and this setting is by now well-understood (Szabó et al. 2016; Fang, Guo, and Zhou 2020), different algorithmic approaches (Mücke 2021), and robust variants (Yu et al. 2021). Despite its practical importance, distributional classification in the two-stage setup has received much less attention. This setting is considered in (Lopez-Paz et al. 2015), though this work focuses on empirical risk minimization. In (Fiedler et al. 2024), first steps towards a more general learning-theoretic analysis are taken, including oracle inequalities for SVMs in the two-stage setups. In particular, it was recognized that most existing works rely on the integral operator technique (Caponnetto and De Vito 2007), which is limited to KRR (and related spectral regularization techniques for regression) and hence cannot be used to treat the classification setting, using for example SVMs with the hinge loss. However, no consistency results or learning rates are provided, and the oracle inequalities have technical limitations, requiring significant regularity of the loss function, which excludes the hinge loss, or requiring a suitable discretization of the hypothesis space, which can be problematic due to the Hilbertian embedding. Finally, we would like to stress that we focus on kernel-based approaches for learning with distributional inputs in the two-stage sampling setup, and other approaches have been considered in this context like deep learning, cf. (Liu and Zhou 2025) for an example and further pointers to the literature, and we refer to (Kachaiev and Recanatesi 2024, Section C) for an overview.

Contributions As a starting point, we prove a very general oracle inequality (Theorem 9) for SVMs in the two-stage sampling setup. In particular, in contrast to the results in (Fiedler et al. 2024) it can now handle the hinge loss without any discretization of the hypothesis space. We then prove (universal) consistency for rather general loss functions, which includes classification with the hinge loss, both for generic Hilbertian embeddings (Proposition 11) and kernel mean embeddings (Proposition 12). Furthermore, under a standard assumption we then establish learning rates for rather general loss functions (Theorem 14), again with classification as a special case. To the best of our knowledge, these are the first such results for distributional classification with SVMs in the two-stage setup. Finally, we investigate learning rates for distributional classification in the two-stage sampling setting with SVMs using the hinge loss and Gaussian kernels, under assumptions that are natural for classification problems. For this, we introduce a variant of a well-known geometric noise exponent assumption from the theory of binary classification (Assumption 16), which in turn allows us to establish learning rates (Theorem 18), without any explicit smoothness assumption as used in regression. To the best of our knowledge, this is again the first such result. Furthermore, for the proof Theorem 18 we develop a new feature space for Gaussian kernels on Hilbert spaces (Theorem 15), which is of independent interest. Due to space constraints, most of the proofs of our results have been placed in the supplementary material.

2 Preliminaries

We start by recalling some preliminaries, including the classic setup of statistical learning theory, and the type of learning methods we consider.

General background We follow (Fiedler et al. 2024) and use *comparison functions* as common in control theory. Recall that class $\mathcal{K}$ functions are defined as $\mathcal{K} = \{f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \mid f \text{ continuous, strictly increasing}, f(0) = 0\}$, and relations and operations on $\mathcal{K}$ are defined pointwise. For the reader's convenience, we have collected additional background on comparison functions in the supplementary material.

We denote the real part of a complex number $z \in \mathbb{C}$ by $\Re z$. For a measure space $(\Omega, \mathcal{A}, \mu)$ and $\mathbb{K} = \{\mathbb{R}, \mathbb{C}\}$, let $L^2(\Omega, \mu; \mathbb{K})$ be the usual Lebesgue space of (μ-equivalence classes of) $\mathbb{K}$-valued square-integrable functions, and let $L^2_{\mathbb{R}}(\Omega, \mu; \mathbb{C})$ be the real Hilbert space arising by restricting scalar multiplication in $L^2(\Omega, \mu; \mathbb{C})$ to the reals. We denote by $L_1^+(\mathcal{H})$ the set of continuous, linear, self-adjoint, trace-class, positive operators on a Hilbert space $\mathcal{H}$.

Finally, for $Q \in L_1^+(\mathcal{H})$ we denote by $\mathcal{N}(0, Q)$ the Gaussian measure on $\mathcal{H}$ with covariance operator Q, and by $\mathcal{H} \ni h \mapsto W_h \in L^2(\mathcal{H}, \mathcal{N}(0, Q); \mathbb{R})$ the associated white noise mapping, cf. (Da Prato and Zabczyk 2002) and the supplementary material for more details.

Statistical Learning Theory We follow the standard setup as formalized in (Steinwart and Christmann 2008, Chapters 2, 6). The *input space* is a measurable space $\mathcal{X}$, the *output space* $\mathcal{Y} \subseteq \mathbb{R}$ closed with the corresponding Borel σ-algebra, and we consider only *supervised loss functions*, i.e., measurable functions $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, which we call continuous, differentiable etc. if $\ell(y, \cdot)$ has this property for all $y \in \mathcal{Y}$. Furthermore, we define $|\ell|_{1,T} = \sup\{|\ell(y, t_1) - \ell(y, t_2)|/|t_1 - t_2| \mid t_1, t_2 \in [-T, T], t_1 \neq t_2, y \in \mathcal{Y}\}$ and say that ℓ is *locally Lipschitz-continuous*[1] if $|\ell|_{1,T} < \infty$ for all $T \in \mathbb{R}_{>0}$. We call ℓ *globally L_ℓ-Lipschitz continuous* for $L_\ell \in \mathbb{R}_{\geq 0}$, if $|\ell(y, t) - \ell(y, t')| \leq L_\ell |t - t'|$ for all $y \in \mathcal{Y}$,

[1] Note that this definition entails uniformity in the first argument of ℓ.

$t, t' \in \mathbb{R}$. We say that ℓ can be *clipped at* $M \in \mathbb{R}_{>0}$ if $\ell(y, \bar{t}) \leq \ell(x, y, t)$ for all $y \in \mathcal{Y}, t \in \mathbb{R}$, where

$$\bar{t} = \begin{cases} M & \text{if } t > M \\ t & \text{if } -M \leq t \leq M \\ -M & \text{if } t < -M \end{cases}$$

is the *clipped value* of t. Furthermore, for a map $f : \mathcal{X} \to \mathbb{R}$, we define $\ell \triangleleft f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ by $(\ell \triangleleft f)(x, y) = \ell(y, f(x))$. For a *data-generating distribution* P on $\mathcal{X} \times \mathcal{Y}$, we define the *risk* of a *hypothesis* $f : \mathcal{X} \to \mathbb{R}$ (measurable function) as $\mathcal{R}_{\ell,P}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(x)) \mathrm{d}P(x, y)$, the *Bayes risk* as $\mathcal{R}^*_{\ell,P} = \inf\{\mathcal{R}_{\ell,P}(f) \mid f : \mathcal{X} \to \mathbb{R}, \text{ measurable}\}$, and for a hypothesis class H also $\mathcal{R}^{H*}_{\ell,P} = \inf\{\mathcal{R}_{\ell,P}(f) \mid f \in H\}$. For a *data set* $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$, we define the *empirical risk* as $\mathcal{R}_{\ell,\mathcal{D}} = \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, f(x_n))$. If H is a normed vector space, we define the *regularized risk* with *regularization parameter* $\lambda \in \mathbb{R}_{>0}$ as $\mathcal{R}_{\ell,P,\lambda}(f) = \mathcal{R}_{\ell,P}(f) + \lambda \|f\|^2_H$, and the *regularized empirical risk* as $\mathcal{R}_{\ell,\mathcal{D},\lambda}(f) = \mathcal{R}_{\ell,\mathcal{D}}(f) + \lambda \|f\|^2_H$. Finally, we turn to notions of learnability. A *learning method* is a measurable[2] map between data sets and a hypothesis space H, formally $\bigcup_{N\in\mathbb{N}_+} (\mathcal{X} \times \mathcal{Y})^N \ni \mathcal{D} \mapsto f_{\mathcal{D}} \in H$. We call such a learning method *ℓ-risk consistent* or just *consistent* if $\mathcal{R}_{\ell,P}(f_{\mathcal{D}_N}) \to \mathcal{R}^*_{\ell,P}$ in probability for $N \to \infty$ and $\mathcal{D}_N \sim P^{\otimes N}$, and *universally ℓ-risk consistent* or just *universally consistent* if this holds for all data-generating distributions P on $\mathcal{X} \times \mathcal{Y}$. Given a set $\mathcal{P}$ of distributions on $\mathcal{X} \times \mathcal{Y}$, a *learning rate* for the learning method is a sequence $(\epsilon_N)_N \subseteq \mathbb{R}_{\geq 0}$ together with a constant $C_{\mathcal{P}}$ and $(c_\delta)_{\delta\in(0,1]}$ such that for all $P \in \mathcal{P}$, $N \in \mathbb{N}_+$, and $\delta \in (0, 1]$ we have

$$\mathbb{P}_{\mathcal{D}\sim P^{\otimes N}}[\mathcal{R}_{\ell,P}(f_{\mathcal{D}}) \leq \mathcal{R}^*_{\ell,P} + C_{\mathcal{P}} c_\delta \epsilon_N] \geq 1 - \delta. \quad (1)$$

Example 1. *We are primarily interested in binary classification, which can be formlized with $\mathcal{Y} = \{-1, 1\}$ (encoding the two classes) and the 0-1-loss $\ell_{\mathrm{c}} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, defined by*

$$\ell_{\mathrm{c}}(y, t) = \begin{cases} 0 & \textit{if } \operatorname{sgn}(t) = y \\ 1 & \textit{otherwise} \end{cases}$$

where $\operatorname{sgn}(t) = 1$ if $t \geq 0$, and -1 otherwise. Note that ℓ_{c} is discontinuous and nonconvex.

Kernels and SVMs We now collect some well-known definitions and facts related to kernels, reproducing kernel Hilbert spaces, and support vector machines, based on the exposition in (Steinwart and Christmann 2008), and we refer to this reference for more details. Recall that a function $k : X \times X \to \mathbb{R}$ is called a *kernel* on an arbitrary nonempty set X, if there exists a Hilbert space $\mathcal{H}$ (called *feature space*) and a map $\Phi : X \to \mathcal{H}$ (called *feature map*) such that $k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ for all $x, x' \in X$. Furthermore, k is the *reproducing kernel* of a Hilbert space H of functions on X if $k(\cdot, x) \in H$ for all $x \in X$, and $f(x) = \langle f, k(\cdot, x) \rangle_H$ for all $f \in H$, $x \in X$, and H is called a *reproducing kernel Hilbert space* (RKHS) if it has a reproducing kernel (which is then unique). Recall also that k is a kernel if and only if it is the reproducing kernel of an RKHS, and the latter is then unique and denoted by $(H_k, \|\cdot\|_k)$. Observe that then H_k is a feature space for k with feature map $\Phi_k(x) = k(\cdot, x)$ (called *canonical feature map*). If $(\mathcal{H}, \Phi)$ is an arbitrary feature space-feature map pair for a kernel k, then $\mathcal{H} \ni h \mapsto \langle h, \Phi(\cdot) \rangle_{\mathcal{H}} \in H_k$ is a canonical surjection on the RKHS of k. We also use the notation $\|k\|_\infty = \sup_{x\in X} \sqrt{k(x, x)}$ and remark that k is bounded if and only if $\|k\|_\infty < \infty$.

In this work, we consider *regularized empirical risk minimization* (RERM) over an RKHS H_k for a kernel k on $\mathcal{X}$, which is called a *support vector machine* (SVM) in this context. For a data set $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^N$, this corresponds to

$$\inf_{f\in H_k} \mathcal{R}_{\ell,\mathcal{D},\lambda}(f),$$

and for ℓ convex there exists a unique solution denoted by $f^{H_k}_{\mathcal{D},\lambda}$, or $f_{\mathcal{D},\lambda}$ if H_k is clear from the context. For analysis purposes, we also define the *approximation error function* $A^{H_k}_{\ell,P}(\lambda) = \mathcal{R}^{H_k*}_{\ell,P,\lambda} - \mathcal{R}^{H_k*}_{\ell,P}$.

Example 2. *We are primarily interested in classification, which can be described by the 0-1-loss ℓ_{c}, cf. Example 1. However, since ℓ_{c} is discontinuous and nonconvex, it is in practice not suitable for RERM, and instead* surrogate losses *are used. The idea is that these losses are well-behaved (in particular, convex), yet still describe a classification task, which can be formalized by* classification calibration*, cf. (Steinwart and Christmann 2008, Chapters 2,3) and (Bach 2024, Chapter 4), as well as the supplementary material for more background on this. The most important example in our context is the* hinge loss $\ell_{\mathrm{h}} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$*, defined by $\ell_{\mathrm{h}}(y, t) = \max\{0, 1 - yt\}$, which is convex and globally Lipschitz continuous, and can be clipped at 1.*

3 Setup

In this section, we formalize the precise setting we work with for the remainder of this manuscript. We first describe the general two-stage learning setup, then we introduce abstract Hilbertian embeddings, and finally we outline kernel mean embeddings as a concrete examle of suitable Hilbertian embeddings.

Two-stage sampling We now formalize the two-stage sampling setup for distributional inputs as introduced in (Póczos et al. 2013; Szabó et al. 2015), following the formalization from (Fiedler et al. 2024). The underlying sampling space is a measurable space $(\mathcal{S}, \mathcal{B}(\tau_{\mathcal{S}}))$, where $(\mathcal{S}, \tau_{\mathcal{S}})$ is a topological space and $\mathcal{B}(\tau)$ is the Borel σ-algebra generated by a topology τ. For the first sampling stage, the input space is then $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$, where $\mathcal{M}_1(\mathcal{S})$ is the set of Borel probability measures on $\mathcal{S}$, and τ_w the topology of weak convergence in $\mathcal{M}_1(\mathcal{S})$. The data-generating distribution, also called *meta-distribution* in this context, is now a probability measure P on $\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y}$. A data set $\mathcal{D}$ with $N \in \mathbb{N}_+$ data points is generated as follows. In the first

[2] For precise definitions, we refer to (Steinwart and Christmann 2008, Chapter 6).

stage, a data set

$$\bar{\mathcal{D}} = ((Q_1, y_1), \dots, (Q_N, y_N)) \in (\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})^N \quad (2)$$

is generated by $(Q_1, y_1), \dots, (Q_N, y_N) \overset{\text{i.i.d.}}{\sim} P$. In the second stage, given $M^{(1)}, \dots, M^{(N)} \in \mathbb{N}_+$, we sample independently

$$S_1^{(n)}, \dots, S_{M^{(n)}}^{(n)} \sim Q_n, \quad n = 1, \dots, N$$

and then set

$$\mathcal{D} = \left((S^{(1)}, y_1), \dots, (S^{(N)}, y_N)\right) \in (\mathcal{S}^* \times \mathcal{Y})^N, \quad (3)$$

where we defined $\mathcal{S}^* = \bigcup_{M \in \mathbb{N}_+} \mathcal{S}^M$ and $S^{(n)} = \left(S_1^{(n)}, \dots, S_{M^{(n)}}^{(n)}\right)$, for $n = 1, \dots, N$.

Hilbertian embeddings We now outline the use of Hilbertian embeddings in this context, following the axiomatic approach from (Fiedler et al. 2024). A *Hilbertian embedding* is a map $\Pi : \mathcal{M}_1(\mathcal{S}) \to \mathcal{H}$, where the *embedding space* $\mathcal{H}$ is a (real) Hilbert space, inducing a new input space $\mathcal{X} = \Pi(\mathcal{M}_1(\mathcal{S})) \subseteq \mathcal{H}$. We also assume access to *embedding estimators* $(\hat{\Pi}_M)_{M \in \mathbb{N}_+}$, $\hat{\Pi}_M : \mathcal{S}^M \to \mathcal{H}$, and define $\hat{\Pi} : \mathcal{S}^* \to \mathcal{X}$ by $\hat{\Pi}(S) = \hat{\Pi}_M(S)$ for all $S \in \mathcal{S}^M$ and $M \in \mathbb{N}_+$. Furthermore, for a first-stage data set $\bar{\mathcal{D}}$ from (2), we define $\bar{\mathcal{D}}_\Pi = \left((\Pi(Q_n), y_n)\right)_{n=1,\dots,N} \in (\mathcal{X} \times \mathcal{Y})^N$, and for a second-stage data set $\mathcal{D}$ from (3), we define $\mathcal{D}_{\hat{\Pi}} = \left((\hat{\Pi}(S^{(n)}), y_n)\right)_{n=1,\dots,N}$. To avoid measurability issues, one can use the following assumption.

Assumption 3. *$\mathcal{H}$ is separable, Π is $\mathcal{B}(\tau_w)$-$\mathcal{B}(\mathcal{H})$-measurable, and $\mathcal{X} \in \mathcal{B}(\mathcal{H})$. Furthermore, for all $M \in \mathbb{N}_+$, $\hat{\Pi}_M$ is $\mathcal{B}(\tau_\mathcal{S})^{\otimes M}$-$\mathcal{B}(\mathcal{X})$-measurable.*

The following result then takes care of measurability issues.

Lemma 4. *Under Assumption 3, the map Π is $\mathcal{B}(\tau_w)$-$\mathcal{B}(\tau_\mathcal{H}|_\mathcal{X})$-measurable, where $\tau_\mathcal{H}|_\mathcal{X}$ is the subspace topology on $\mathcal{X}$ induced by the topology on $\mathcal{H}$. Furthermore, every $P \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})$ induces a distribution P_Π on $\mathcal{X} \times \mathcal{Y}$ as the pushforward of P along $(Q, y) \mapsto (\Pi(Q), y)$.*

A proof of this result is provided in Section A.1.1 in (Szabó et al. 2015) and the supplementary to (Lopez-Paz et al. 2015). For the analysis later on, we need probabilistic estimation bounds for the Hilbertian embeddings, which we abstract in the next assumption.

Assumption 5. *We have access to $B_\Pi : \mathbb{N}_+ \times (0,1) \to \mathbb{R}_{\geq 0}$ such that for all $Q \in \mathcal{M}_1(\mathcal{S})$, $M \in \mathbb{N}_+$, $\delta \in (0,1)$*

$$\mathbb{P}_{S \sim Q^{\otimes M}}[\|\Pi(Q) - \hat{\Pi}(S)\|_\mathcal{H} > B_\Pi(M, \delta)] < \delta$$

We would like to stress the embedding strategy outlined here requires a kernel k on $\mathcal{X}$, which is a subset of an in general infinite-dimensional Hilbert space. The availability of such kernels is one major advantage of this approach, and we refer to (Meunier, Pontil, and Ciliberto 2022) for more details. In the next example, we recall an important instance of such a kernel.

Example 6. *Let $\emptyset \neq X \subseteq \mathcal{H}$ be a subset of an arbitrary real Hilbert space. For $\gamma \in \mathbb{R}_{>0}$, $k_\gamma(x, x') = \exp\left(-\|x - x'\|_\mathcal{H}^2 / \gamma^2\right)$ is a kernel on X, called* Gaussian kernel*, and we denote its unique RKHS by $(H_\gamma, \|\cdot\|_{k_\gamma})$, cf. (Christmann and Steinwart 2010) and (Meunier, Pontil, and Ciliberto 2022) for more details.*

In general, in the following we will work with kernels k that fulfill the next (rather mild) assumption. For instance, since $\mathcal{X} \subseteq \mathcal{H}$ is separable, cf. Assumption 3, the Gaussian kernel from Example 6 fulfills it.

Assumption 7. *The kernel k on $\mathcal{X}$ is measurable, bounded, and has a separable RKHS H_k. Furthermore, there exists $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x) - \Phi_k(x')\|_k \leq \alpha_k(\|x - x'\|_\mathcal{H})$ holds for all $x, x' \in \mathcal{X}$.*

The last condition in the preceding assumption is easily fulfilled for Hölder-continuous kernels, cf. (Fiedler 2023) for a thorough discussion of this aspect, and (Szabó et al. 2015) for an extensive list of concrete examples of such kernels.

As a concrete example of Hilbertian embeddings, we use *kernel mean embeddings* (KMEs). For the reader's convenience, we collect now some well-known definitions and facts, cf. (Lopez-Paz et al. 2015; Szabó et al. 2016) for more details and proofs.

Proposition 8. *Let $(\mathcal{S}, \mathcal{A}_\mathcal{S})$ be a measurable space, and κ a measurable and bounded kernel on $\mathcal{S}$ with separable RKHS H_κ. (i) The map*

$$\Pi_\kappa : \mathcal{M}_1(\mathcal{S}) \to H_\kappa,\ \Pi_\kappa Q = \int \kappa(\cdot, s) \mathrm{d}Q(s) \quad (4)$$

is well-defined, and we call $\Pi_\kappa Q$ the kernel mean embedding (KME) *of $Q \in \mathcal{M}_1(\mathcal{S})$ w.r.t. κ.*

(ii) Define $\hat{\Pi}_\kappa : \mathcal{S}^* \to H_\kappa$ by

$$\hat{\Pi}_\kappa((s_1, \dots, s_M)) = \frac{1}{M} \sum_{m=1}^{M} \kappa(\cdot, s_m). \quad (5)$$

For all $Q \in \mathcal{M}_1(\mathcal{S})$ and $S \sim Q^{\otimes M}$, $M \in \mathbb{N}_+$, and $\delta \in (0,1)$, we have that

$$\|\hat{\Pi}_\kappa S - \Pi_\kappa Q\|_\kappa \leq 2\sqrt{\frac{\|\kappa\|_\infty^2}{M}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(1/\delta)}{M}} \quad (6)$$

holds with probability at least $1 - \delta$.

(iii) Let $(\mathcal{S}, \tau_\mathcal{S})$ be a separable topological space, consider $\mathcal{A}_\mathcal{S} = \mathcal{B}(\tau_\mathcal{S})$, and assume that κ is continuous, then Π_κ is $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$-$(H_\kappa, \mathcal{B}(H_\kappa))$-measurable.

4 Consistency and Learning Rates

We will now present our first main results. Building on a rather general oracle inequality stated in Section 4.1, we establish (universal) consistency for SVMs in the two-stage sampling setup in Section 4.2, and then learning rates in Section 4.3.

4.1 Oracle Inequality

As common in statistical learning theory, consistency and learning rates can be derived from an oracle inequality. The following result will be our central tool for this task, and it is a two-stage sampling variant of (Steinwart and Christmann 2008, Theorem 7.22).

Theorem 9. *Consider the two-stage sampling setup outlined in Section 3. Let k be a kernel on $\mathcal{X}$ that fulfills Assumption 7, let the Hilbertian embedding fulfill Assumptions 3 and 5, and consider a convex, locally Lipschitz-continuous loss ℓ that can be clipped at $M \in \mathbb{R}_{>0}$. Finally, assume that there exists $B \in \mathbb{R}_{\geq 0}$ such that*

$$\ell(y,t) \leq B \quad \forall y \in \mathcal{Y}, t \in [-M,M] \tag{7}$$

holds. Then there exists a universal constant $C \in \mathbb{R}_{>0}$ such that for all $N \geq 2$, $\lambda \in \mathbb{R}_{>0}$ and $\tau \geq 1$ it holds with probability at least $1 - 4e^{-\tau}$ that

$$\begin{aligned}\mathcal{R}_{\ell,P_\Pi}(\bar{f}_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \mathcal{R}^*_{\ell,P_\Pi} &\leq 9A^{H_k}_{\ell,P_\Pi}(\lambda) \\ &+ 9(\mathcal{R}^{H_k*}_{\ell,P_\Pi} - \mathcal{R}^*_{\ell,P_\Pi}) + C|\ell|^2_{1,M}\|k\|_\infty \frac{\ln N}{N\lambda} \\ &+ 300\frac{B\tau}{\sqrt{N}} + 15\frac{\tau}{N}|\ell|_{1,C_\lambda}\|k\|_\infty\sqrt{\frac{A^{H_k}_{\ell,P_\Pi}(\lambda)}{\lambda}} \\ &+ \frac{3}{N}\sum_{n=1}^{N}\alpha_\lambda\left(B_\Pi(M^{(n)}, e^{-\tau}/N)\right),\end{aligned} \tag{8}$$

where

$$\alpha_\lambda = \left(|\ell|_{1,C_\lambda}\sqrt{A^{H_k}_{\ell,P_\Pi}(\lambda)/\lambda} + |\ell|_{1,M}\sqrt{B/\lambda}\right)\alpha_k.$$

and

$$C_\lambda = \|k\|_\infty\sqrt{A^{H_k}_{\ell,P_\Pi}(\lambda)/\lambda}.$$

On a high level, for the proof we use continuity properties to go from the accessible data set $\mathcal{D}_{\hat{\Pi}}$ to the inaccessible first-stage sampling data set $\bar{\mathcal{D}}_\Pi$, on which existing results can be applied. While this is the standard strategy for the two-stage sampling setup, going back to (Szabó et al. 2015; Lopez-Paz et al. 2015) and also used by (Fiedler et al. 2024), we use a rather advanced oracle inequality for the first stage of sampling, which requires some work. Similarly as in the proof of (Steinwart and Christmann 2008, Theorem 7.22), we first establish an oracle inequality for general approximate RERM schemes, and then check that SVMs indeed fulfill the necessary assumptions for this class of learning methods. A detailed proof is provided in the supplementary material.

4.2 Consistency

We now turn to consistency of SVMs in the two-stage sampling setup. It is clear that for consistency to hold, the hypothesis space H_k has to be expressive enough. This is formalized in the next assumption.

Assumption 10. *It holds that $\mathcal{R}^{H_k*}_{\ell,P_\Pi} = \mathcal{R}^*_{\ell,P}$.*

For simplicity, we consider from now on only globally Lipschitz-continuous loss functions, including in particular the hinge loss ℓ_{h}, which is of prime importance for classification. We are now ready to state the following general consistency result for SVMs with distributional inputs in the two-stage sampling setup.

Proposition 11. *Consider the situation of Theorem 9. Assume that the loss function ℓ is globally L_ℓ-Lipschitz continuous, and assume that for a data set of size $N \in \mathbb{N}_+$, for every data point, $M_N \in \mathbb{N}_+$ samples are drawn in the second stage of sampling, so $M^{(1)} = \ldots = M^{(N)} = M_N$ for N samples. If Assumption 10 holds, and if $(\lambda_N)_N \subseteq \mathbb{R}_{>0}$ and $(M_N)_N \subseteq \mathbb{N}_+$ are sequences such that $\lim_{N\to\infty}\lambda_N = 0$ and*

$$\lim_{N\to\infty}\frac{\ln(N)}{N\lambda_N} = \lim_{N\to\infty}\frac{1}{\sqrt{\lambda_N}}\alpha_k(B_\Pi(M_N, 1/N)) = 0, \tag{9}$$

then

$$(\mathcal{S}^{M_N}\times\mathcal{Y})^N \ni \mathcal{D}^{(N)} \mapsto \bar{f}_{\mathcal{D}^{(N)}_{\hat{\Pi}},\lambda_N}\circ\Pi$$

is an ℓ-risk consistent learning method, so

$$\mathcal{R}_{\ell,P}(\bar{f}_{\mathcal{D}^{(N)}_{\hat{\Pi}},\lambda_N}\circ\Pi) \to \mathcal{R}^*_{\ell,P}$$

in probability for $N\to\infty$, for all data-generating distributions P under Assumption 10.

This result poses two conditions on $(\lambda_N)_N$ and $(M_N)_N$. The first one appears also in the usual statistical learning setup with only a single stage of sampling, cf. the discussion in (Steinwart and Christmann 2008, Section 7.4). The second condition arises through the two-stage sampling setup, which is a well-known effect in the case of distributional regression, cf. (Szabó et al. 2015, 2016).

Proof. Observe that

$$\begin{aligned}\mathcal{R}^*_{\ell,P} &= \inf_{f \text{ measurable}}\mathcal{R}_{\ell,P}(f) \\ &\leq \inf_{f \text{ measurable}}\mathcal{R}_{\ell,P}(f\circ\Pi) = \mathcal{R}^*_{\ell,P_\Pi},\end{aligned}$$

and $\mathcal{R}_{\ell,P}(\bar{f}_{\mathcal{D}_{\hat{\Pi}},\lambda}\circ\Pi) - \mathcal{R}^*_{\ell,P} \leq \mathcal{R}_{\ell,P_\Pi}(\bar{f}_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|\bar{f}_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \mathcal{R}^*_{\ell,P_\Pi}$. By definition of ℓ-risk consistency and using Assumption 10, it is hence enough to ensure that for fixed $\tau \geq 1$, the righthand side in (8) converges to zero for $N\to\infty$. Since $\lambda_N \to 0$, we have $A^{H_k}_{\ell,P_\Pi}(\lambda_N) \to 0$, cf. (Steinwart and Christmann 2008, Lemma 5.15). Furthermore, $\frac{\ln(N)}{N\lambda_N}\to 0$ implies $N\lambda_N\to\infty$, so $15\frac{\tau}{N}L_\ell\|k\|_\infty\sqrt{A^{H_k}_{\ell,P_\Pi}(\lambda_N)/\lambda_N} = 15\tau L_\ell\|k\|_\infty/\sqrt{N}\times\sqrt{A^{H_k}_{\ell,P_\Pi}(\lambda_N)/(N\lambda_N)} \to 0$. Finally, the last condition in (9) ensures that also the remaining terms converge to zero. Altogether, this shows that the righthand side in (8) indeed converges to zero, establishing consistency. □

For concreteness we now consider KMEs as Hilbertian embeddings and we assume Hölder-continuity of Φ_k. In this situation, we can achieve the following consistency result.

Proposition 12. *Consider the situation of Theorem 9 and assume that the loss function ℓ is globally L_ℓ-Lipschitz continuous. Assume that for a data set of size $N \in \mathbb{N}_+$, for every data point, $M_N \in \mathbb{N}_+$ samples are drawn in the second stage of sampling, and that KMEs (from Proposition 8) are used for the Hilbertian embedding. Furthermore, assume that there exist $C_k, \alpha \in \mathbb{R}_{>0}$ such that $\alpha_k(s) = C_k s^\alpha$. If Assumption 10 holds, and if $(\lambda_N)_N \subseteq \mathbb{R}_{>0}$ and $(M_N)_N \subseteq \mathbb{N}_+$ are sequences such that $\lim_{N\to\infty} \lambda_N = 0$ and*

$$\lim_{N\to\infty} \frac{\ln(N)}{N\lambda_N} = 0 \quad \textit{and} \quad \lim_{N\to\infty} \frac{\ln(N)^\alpha}{\lambda_N M_N^\alpha} = 0, \tag{10}$$

then $(\mathcal{S}^{M_N} \times \mathcal{Y})^N \ni \mathcal{D}^{(N)} \mapsto \bar{f}_{\mathcal{D}^{(N)}_{\hat{\Pi}_\kappa}, \lambda_N} \circ \Pi_\kappa$ is a ℓ-risk consistent learning method for all data-generating distributions P under Assumption 10.

This result follows as an immediate corollary from Proposition 11, and we provide a detailed proof in the supplementary material.

4.3 Learning Rates

As is well-known, learning rates can only be derived under distributional assumptions due to the *No Free Lunch Theorem*. The form of the oracle inequality in Theorem 9 shows that any distributional assumption must enter through the approximation error function. The following is a standard assumption for this task, cf. (Steinwart and Christmann 2008, Chapter 6) and (Steinwart et al. 2009).

Assumption 13. *There exist constants $C_\mathcal{A} \in \mathbb{R}_{>0}$, $\beta \in (0, 1]$ such that $A^{H_k}_{\ell, P_\Pi}(\lambda) \leq C_\mathcal{A} \lambda^\beta$ for all $\lambda \in \mathbb{R}_{>0}$.*

For concreteness, we use this to establish a learning rate in the case of KMEs for the Hilbertian embeddings. Learning rates for other embeddings can be derived similarly, and in the supplementary material we provide a corresponding result for generic Hilbertian embeddings.

Theorem 14. *Consider the situation of Proposition 12, assume that $\alpha \in (0, 2]$, and let in addition Assumption 13 hold for $\Pi = \Pi_\kappa$ from Proposition 8. If $(M_N)_N$ grows at least as $N^{\frac{2}{\alpha}}$, and $(\lambda_N)_N$ decays as $N^{-\frac{1}{\beta+1}}$, then a learning rate of $\ln(N) N^{-\frac{\beta}{\beta+1}}$ is achieved.*

This result can be derived from Theorem 9 using well-known elementary arguments.

5 Classfication With Gaussian Kernels and the Hinge Loss

We now turn to classification ($\mathcal{Y} = \{-1, 1\}$) using the hinge loss ℓ_h and the Gaussian kernel k_γ on $\mathcal{X}$, cf. Example 6. The theory in Section 4 covers this case already, however, the important Assumption 13 is in general difficult to interpret. For binary classification, margin and noise exponent assumptions are more intuitive, cf. (Steinwart and Christmann 2008, Chapter 8) for an overview, and our goal in this section is to realize this also for distributional classification in the two-stage sampling setup. For this, we will first introduce a new feature space for Gaussian kernels on (subsets of) Hilbert spaces, which is of independent interest, and then establish a bound on the approximation error function using a variant of an establish geometric margin condition.

5.1 A New Feature Space

In order to bound the approximation error function for the hinge loss and the Gaussian kernel, we need a convenient feature space. For the Gaussian kernel on $X \subseteq \mathbb{R}^d$, one can use $L^2(X, \lambda_X^{(d)}, \mathbb{R})$, where $\lambda_X^{(d)}$ is the Lebesgue measure on X, cf. (Steinwart and Christmann 2008, Section 4.4). However, since we consider the Gaussian kernel on $\mathcal{X} \subseteq \mathcal{H}$, where $\mathcal{H}$ is in general infinite-dimensional, we do not have the Lebesgue measure available anymore. Instead, as common in infinite-dimensional analysis (Da Prato 2006), we use the Gaussian measure on a Hilbert space instead. The next result describes how exactly this can be used to build a feature space-feature map pair for the Gaussian kernel on a separable Hilbert space.

Theorem 15. *Let $\mathcal{H}$ be a separable real Hilbert space, $\emptyset \neq X \subseteq \mathcal{H}$, and k_γ the Gaussian kernel on X with length scale $\gamma \in \mathbb{R}_{>0}$. For all $Q \in L_1^+(\mathcal{H})$ with $\ker(Q) = \{0\}$, $L^2_\mathbb{R}(\mathcal{H}, \mathcal{N}(0, Q); \mathbb{C})$ is a (real) feature space of k_γ, $\Phi_Q : X \to L^2_\mathbb{R}(\mathcal{H}, \mathcal{N}(0, Q); \mathbb{C})$ defined by*

$$\Phi_Q(x) = \exp(i\sqrt{2}/\gamma \cdot W_x(\cdot)) \tag{11}$$

is a corresponding feature map, and the canonical surjection $V_Q : L^2_\mathbb{R}(\mathcal{H}, \mathcal{N}(0, Q); \mathbb{C}) \to H_\gamma$ is given by

$$(V_Q g)(x) = \Re \int_\mathcal{H} \exp\left(-i\frac{\sqrt{2}}{\gamma} W_x(z)\right) g(z) \mathrm{d}\mathcal{N}(z \mid 0, Q). \tag{12}$$

We provide a detailed proof of this result in the supplementary material.

5.2 Learning Rates

In order to bound the approximation error function for the hinge loss, we follow the high level strategy from (Steinwart and Scovel 2007). First, we need some preliminaries. We consider a distribution P on $\mathcal{X} \times \mathcal{Y}$, and recall that $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{X} \subseteq \mathcal{H}$ (later on, P_Π will play the role of this P). Let $\eta : \mathcal{X} \to [0, 1]$ be a version of the conditional probability $\mathbb{P}_{(X,Y)\sim P}[Y = 1 \mid X = x]$, and define

$$X_1 = \{x \in \mathcal{X} \mid \eta(x) > \frac{1}{2}\} \quad X_{-1} = \{x \in \mathcal{X} \mid \eta(x) < \frac{1}{2}\}$$

and

$$\Delta(x) = \begin{cases} d_\mathcal{H}(x, X_1) & \text{if } x \in X_{-1} \\ d_\mathcal{H}(x, X_{-1}) & \text{if } x \in X_1 \\ 0 & \text{otherwise} \end{cases}$$

where $d_\mathcal{H}(x, A) = \inf_{x' \in A} \|x - x'\|_\mathcal{H}$ for $x \in \mathcal{H}$ and $A \subseteq \mathcal{H}$. Furthermore, define $f_P^* : \mathcal{X} \to [-1, 1]$ as

$$f_P^*(x) = \begin{cases} 1 & \text{if } x \in X_1 \\ -1 & \text{if } x \in X_{-1} \\ 0 & \text{otherwise} \end{cases}$$

Then f_P^* is measurable and achieves the Bayes risk, i.e., $\mathcal{R}_{\ell_c, P}(f_P^*) = \mathcal{R}^*_{\ell_c, P}$.

We will now state the central assumption of this section. It can be interpreted as a variant of the geometric noise exponent assumption from (Steinwart and Scovel 2007), adapted to a Hilbert space setting.

Assumption 16. *P and η are such that there exist $Q \in L_1^+(\mathcal{H})$ with $\ker(Q) = \{0\}$ and constants $C_Q, \alpha_Q, \bar{t}_Q \in \mathbb{R}_{>0}$ such that for all $0 < t \leq \bar{t}_Q$ it holds that*

$$\int_{X_1 \cup X_{-1}} \left(1 - 2 \int_{B_{\Delta(x)}(x)} \exp\left(-\frac{\|x-y\|_{\mathcal{H}}^2}{t}\right) \mathrm{d}\mathcal{N}(y \mid 0, Q)\right) \times |2\eta(x) - 1| \mathrm{d}P_X(x) \leq C_Q t^{\alpha_Q}. \tag{13}$$

and

$$\int_{\mathcal{X}} \left(\int_{\mathcal{H}} \exp\left(-\frac{1}{t}\|y\|_{\mathcal{H}}^2\right) \mathrm{d}\mathcal{N}(y \mid x, Q)\right) \times |2\eta(x) - 1| \mathrm{d}P_X(x) \leq C_Q t^{\alpha_Q} \tag{14}$$

While rather technical, the preceding assumption has a clear intuitive interpretation, which we discuss in the supplementary material. We are now ready to use this assumption to derive a bound on the approximation error function for Gaussian kernels on separable Hilbert spaces.

Theorem 17. *Let Assumption 16 hold, then for all $\gamma \in \mathbb{R}_{>0}$ with $\gamma^2 < \bar{t}_Q$, we have*

$$A_{\ell_{\mathrm{h}},P}^{H_\gamma}(\lambda) \leq 2C_Q \gamma^{2\alpha_Q} + \lambda \tag{15}$$

for all $\lambda \in \mathbb{R}_{>0}$.

The proof uses the strategy from (Steinwart and Scovel 2007, Section 4), cf. also (Steinwart and Christmann 2008, Section 8.2): An explicit Bayes optimal classifier is embedded into the Gaussian RKHS (here via Theorem 15), which is interpreted as a smoothing, and the resulting performance degradation, i.e., increase in risk, is bounded using a margin or noise assumption, here Assumption 16. A detailed proof is provided in the supplementary material.

Finally, all of this can be combined to arrive at the following result on learning rates for distributional classification with hinge-loss SVMs with Gaussian kernels in the two-stage sampling setup. For concreteness, we consider KMEs for the Hilbertian embeddings, but analogous results can be derived in a similar manner for other embeddings.

Theorem 18. *Consider the situation of Proposition 12 with $\ell = \ell_{\mathrm{h}}$ and the Gaussian kernel on $\mathcal{X}$, and let in addition Assumption 16 hold for $P = P_{\Pi_\kappa}$. If $(M_N)_N \subseteq \mathbb{N}_+$ grows at least as $N^{\frac{2}{\alpha}}$, $(\lambda_N)_N \subseteq \mathbb{R}_{>0}$ decays as $N^{-\frac{1}{2}}$, $(\gamma_N)_N \subseteq \mathbb{R}_{>0}$ decays as $N^{-\mu}$ for some $\mu \in \mathbb{R}_{>0}$, and Assumption 10 holds for all H_{γ_N}, then $(\mathcal{S}^{M_N} \times \mathcal{Y})^N \ni \mathcal{D}^{(N)} \mapsto \bar{f}_{\mathcal{D}_{\Pi_\kappa}^{(N)}, \lambda_N}^{H_{\gamma_N}} \circ \Pi_\kappa$ achieves a learning rate of $N^{-\min\{2\mu\alpha_Q, \frac{1}{2}\}}$.*

This result follows directly from Theorems 9 and 17.

6 Conclusion

We considered kernel-based statistical learning with distributional inputs in the two-stage sampling setup, for which we established consistency and learning rates for rather general loss functions, covering the important case of binary classification. Furthermore, using a novel variant of an establish geometric margin exponent assumption, we were able to prove learning rates for distributional classification with SVMs using the hinge loss and Gaussian kernels. In particular, we could establish a learning rate without relying on an explicit smoothness assumption as common in regression, since this can be inappropriate for a classification setting. While we focused primarily on KMEs as Hilbertian embeddings, our results can be easily adapted to other embeddings. Our work opens up a multitude of interesting directions. First, by using a supremum bound in an oracle inequality like Theorem 9, our rates can be further refined, as in the case of single-stage sampling, cf. (Steinwart and Christmann 2008, Chapter 7). Second, using appropriate discretizations, for example via entropy number estimates, is another avenue for refinement of our rates, which requires dealing with a delicate interplay of the embedding map, the (in general infinite-dimensional) embedding Hilbert space, and the kernel used in the SVM. Third, a closer investigation of Assumption 16 and potential refinements is another interesting aspect for future work. Finally, extensions to related learning problems like multiclass classification are another line of interesting future work.

Acknowledgements

The author would like to thank Pierre-François Massiani, Oleksii Kachaiev, and Ingo Steinwart for very helpful discussions, Alessandro Scagliotti for a careful reading of the manuscript, Mattes Mollenhauer for insightful comments on Section 5, and the anonymous reviewers for helpful comments. The author acknowledges funding from DFG Project FO 767/10-2 (eBer-24-32734) "Implicit Bias in Adversarial Training".

References

Bach, F. 2024. *Learning theory from first principles*. MIT press.

Bonnier, P.; Oberhauser, H.; and Szabó, Z. 2023. Kernelized Cumulants: Beyond Kernel Mean Embeddings. *Advances in Neural Information Processing Systems*.

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

Christmann, A.; and Steinwart, I. 2010. Universal kernels on non-standard input spaces. *Advances in neural information processing systems*, 23.

Da Prato, G. 2006. *An introduction to infinite-dimensional analysis*. Springer Science & Business Media.

Da Prato, G.; and Zabczyk, J. 2002. *Second order partial differential equations in Hilbert spaces*, volume 293. Cambridge University Press.

Fang, Z.; Guo, Z.-C.; and Zhou, D.-X. 2020. Optimal learning rates for distribution regression. *Journal of complexity*, 56: 101426.

Fiedler, C. 2023. Lipschitz and Hölder Continuity in Reproducing Kernel Hilbert Spaces. *arXiv preprint arXiv:2310.18078*.

Fiedler, C.; Massiani, P.-F.; Solowjow, F.; and Trimpe, S. 2024. On statistical learning theory for distributional inputs. In *Forty-first International Conference on Machine Learning*.

Kachaiev, O.; and Recanatesi, S. 2024. Learning to embed distributions via maximum kernel entropy. *Advances in Neural Information Processing Systems*, 37: 44710–44734.

Liu, P.; and Zhou, D.-X. 2025. Generalization Analysis of Transformers in Distribution Regression. *Neural Computation*, 37(2): 260–293.

Lopez-Paz, D.; Muandet, K.; Schölkopf, B.; and Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, 1452–1461. PMLR.

Massiani, P. F.; Haverbeck, L.; Thesing, C.; et al. 2025. Robust screening of atrial fibrillation with distribution classification. *Scientific Reports*, 15: 26582.

Meunier, D.; Pontil, M.; and Ciliberto, C. 2022. Distribution regression with sliced Wasserstein kernels. In *International Conference on Machine Learning*, 15501–15523. PMLR.

Muandet, K.; Fukumizu, K.; Dinuzzo, F.; and Schölkopf, B. 2012. Learning from distributions via support measure machines. *Advances in neural information processing systems*, 25.

Mücke, N. 2021. Stochastic gradient descent meets distribution regression. In *International Conference on Artificial Intelligence and Statistics*, 2143–2151. PMLR.

Póczos, B.; Singh, A.; Rinaldo, A.; and Wasserman, L. 2013. Distribution-free distribution regression. In *artificial intelligence and statistics*, 507–515. PMLR.

Steinwart, I.; and Christmann, A. 2008. *Support vector machines*. Springer Science & Business Media.

Steinwart, I.; Hush, D. R.; Scovel, C.; et al. 2009. Optimal Rates for Regularized Least Squares Regression. In *COLT*, 79–93.

Steinwart, I.; and Scovel, C. 2007. Fast rates for support vector machines using Gaussian kernels.

Szabó, Z.; Gretton, A.; Póczos, B.; and Sriperumbudur, B. 2015. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, 948–957. PMLR.

Szabó, Z.; Sriperumbudur, B. K.; Póczos, B.; and Gretton, A. 2016. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1): 5272–5311.

Yu, Z.; Ho, D. W.; Shi, Z.; and Zhou, D.-X. 2021. Robust kernel-based distribution regression. *Inverse Problems*, 37(10): 105014.