# On Statistical Learning Theory for Distributional Inputs

**Christian Fiedler** [1] **Pierre-François Massiani** [1] **Friedrich Solowjow** [1] **Sebastian Trimpe** [1]

## Abstract

Kernel-based statistical learning on distributional inputs appears in many relevant applications, from medical diagnostics to causal inference, and poses intriguing theoretical questions. While this learning scenario received considerable attention from the machine learning community recently, many gaps in the theory remain. In particular, most works consider only the distributional regression setting, and focus on the regularized least-squares algorithm for this problem. In this work, we start to fill these gaps. We prove two oracle inequalities for kernel machines in general distributional learning scenarios, as well as a generalization result based on algorithmic stability. Our main results are formulated in great generality, utilizing general Hilbertian embeddings, which makes them applicable to a wide array of approaches to distributional learning. Additionally, we specialize our results to the cases of kernel mean embeddings and of the recently introduced Hilbertian embeddings based on sliced Wasserstein distances, providing concrete instances of the general setup. Our results considerably enlarge the scope of theoretically grounded distributional learning, and provide many interesting avenues for future work.

## 1. Introduction

Supervised statistical learning with distributional inputs has received significant attention lately, cf. (Szabó et al., 2016; Fang et al., 2020; Meunier et al., 2022), both from practical and theoretical perspectives. The goal is to learn a relation between inputs and outputs from data, where the inputs are probability distributions on some measurable space. Furthermore, the inputs (probability distributions) are not directly accessible, but the data contains only samples thereof. A classic example is the prediction of some health indicator of a patient from several clinical measurements (Szabó et al., 2015), which we recall now. Let $\mathcal{S}$ be the set of outcomes of some diagnosis tools (e.g., electrocardiogram characteristics, or the blood serum concentration of some substance). Since these measurements will vary even when coming from the same patient, it is reasonable to assume that an individual patient with a specific health status has a certain distribution $Q$ on $\mathcal{S}$ that generates the measurements and that can be a predictor for some health indicator $y \in \mathcal{Y}$ (e.g., healthy or not). However, during training, $Q$ is not directly accessible, but rather through independent and identically distributed (i.i.d.) samples $S_1, \dots, S_M \overset{\text{i.i.d.}}{\sim} Q$. For example, this could correspond to daily blood measurements of a patient during a week-long hospital stay, assuming the patient's distribution $Q$ has not changed during the week (e.g., when the health status has not changed). The training data consists of such data from $N$ different patients, so the data set is not of the form $\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\dots,N}$, but rather $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\dots,N}$, where $S_1^{(n)}, \dots, S_{M_n}^{(n)} \overset{\text{i.i.d.}}{\sim} Q_n$. The goal is to learn a map $f_{\mathcal{D}}$ from distributions $Q$ over $\mathcal{S}$ to outcomes $\mathcal{Y}$ (e.g., from distributions over diagnostic measurements to health status).

Among such learning problems, the focus of previous theoretical investigations has been on *distributional regression*. In this setting, one is interested in predicting a real-valued quantity from a distributional input, so $\mathcal{Y} = \mathbb{R}$. While the early work (Póczos et al., 2013) relied on density estimation, starting with Szabó et al. (2015), kernel mean embeddings (KMEs) together with kernel ridge regression (KRR) have been used. For concreteness, let us review this latter approach. Consider a data set $\mathcal{D}$ as introduced above. A single input item $S^{(n)}$ is first interpreted as an empirical measure $\hat{\mu}[S^{(n)}] = \frac{1}{M_n} \sum_{m=1}^{M_n} \delta_{S_m^{(n)}}$, where $\delta_s$ is the Dirac measure with atom on $s$, which is then embedded into a reproducing kernel Hilbert space (RKHS) $H_\kappa$ using the KME, $\hat{\mu}[S^{(n)}] \mapsto \Pi_\kappa \hat{\mu}[S^{(n)}]$. Assuming access to a (second) kernel $k$ on the RKHS $H_\kappa$, one then performs KRR on the transformed data set $\mathcal{D}_{\hat{\Pi}_\kappa} = ((\Pi_\kappa \hat{\mu}[S^{(n)}], y_n))_{n=1,\dots,N}$. The resulting learned function $f_{\mathcal{D}_{\hat{\Pi}_\kappa}}$ can then be used for prediction by composing it with the KME map. A distribution $Q$ on $\mathcal{S}$ would therefore lead to prediction $f_{\mathcal{D}_{\hat{\Pi}_\kappa}}(\Pi_\kappa Q)$. This approach has been thoroughly analyzed (Szabó et al., 2015; 2016; Fang et al., 2020). All of these investigations rely on the seminal analysis (Caponnetto & De Vito, 2007)

[1]Institute for Data Science in Mechanical Engineering (DSME), RWTH Aachen University, Aachen, Germany. Correspondence to: Christian Fiedler <fiedler@dsme.rwth-aachen.de>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

of the regularized least-squares algorithm for regression.

Recently, (Meunier et al., 2022) developed a much more general perspective on this problem. Instead of KMEs, they consider suitable embeddings $\Pi$ of probability distributions into some Hilbert space $\mathcal{H}$, and then utilize distance substitution kernels (Haasdonk & Bahlmann, 2004) with the induced Hilbertian (semi)metric $(P, Q) \mapsto \|\Pi P - \Pi Q\|_{\mathcal{H}}$ on probability distributions. In particular, they apply this construction to sliced Wasserstein (SW) distances (Bonneel et al., 2015) and construct corresponding SW kernels. The resulting method has been theoretically analyzed, building again on (Caponnetto & De Vito, 2007).

Despite this multitude of activity, many interesting and practically relevant problems in this area are still open. In this work, we focus on two theoretical aspects. *First*, most theoretically grounded works in the context of distributional learning methods have focused almost exclusively on the distributional regression problem. However, other learning scenarios are also highly relevant, in particular, distributional classification. For example, in the medical example outlined above, a natural task is to predict a binary health status of a patient (e.g., having a certain disease or not), corresponding to (distributional) binary classification. As another example, in Lopez-Paz et al. (2015), distributional classification is applied to the problem of predicting causal directions and causal graphs from data. KMEs are also used there, though empirical risk minization (ERM) is then applied on the transformed data set. To the best of our knowledge, this reference is also the only one investigating distributional classification with KMEs from a theoretical perspective. While they establish some generalization bounds based on margin theory, no consistency results or oracle inequalities in the sense of (Steinwart & Christmann, 2008) are provided. Recall that an oracle inequality in this context is a high probability bound on the excess risk of the learned hypothesis over what an oracle, that has access to the true data-generating distribution, can achieve. In turn, these inequalities allow derivation of consistency results, and under suitable distributional assumptions also of learning rates, and it is hence highly desirable that such inequalities are also available in the distributional learning setting. *Second*, the theoretical analyses of distributional learning have been restricted to rather specific settings. Even in the context of distributional regression, the learning setups considered have been fairly specific. In particular, in the context of KME-based distributional regression, to the best of our knowledge only KRR has been considered so far, and analyzed exclusively using (Caponnetto & De Vito, 2007). This technique is inherently limited to KRR, and hence cannot be used to analyze inter alia support vector regression (SVR) with the $\epsilon$-insensitive loss. It is furthermore also not suitable to analyze classification using support vector machines (SVMs), or more general regularized empirical risk approaches.

**Contributions** In this work, we tackle these open issues. First, for the distributional learning setting outlined above, we provide two oracle inequalities for the risk of SVMs (in the sense of regularized risk minimization over RKHSs) that cover a multitude of learning scenarios. To the best of our knowledge, both of these results are completely new in the context of learning on distributional inputs. Second, we establish a generalization bound for distributional learning based on algorithmic stability, and apply it to SVMs. Third, inspired by (Meunier et al., 2022), we formulate all of this for kernel-based methods that rely on a general Hilbertian embedding of probability distributions. In this manner, our results apply to the case of KMEs and SW kernels, and *any future method that provides Hilbertian embeddings*. Fourth, we specialize our results to KMEs and SW distances for the Hilbertian embedding.

**Outline** In Section 2, we collect necessary background, in particular, on statistical learning theory and the theory of RKHSs. Furthermore, we formalize the distributional learning setup, and provide details on kernel-based methods using Hilbertian embeddings. In Section 3, we present our two oracle inequalities, and specialize them to the case of KMEs. In Section 4, we present a generalization result for distributional learning based on algorithmic stability. We use this to prove a corresponding generalization result for SVMs in the distributional setting, and specialize the latter to the case of KMEs again. Section 5 closes with a summary and an outlook. In the appendix we collect some technical background, and proofs of the oracle inequalities and the main generalization result. Furthermore, in the appendix we also provide specializations of our results to the case of using the sliced 2-Wasserstein distance for the Hilbertian embedding.

## 2. Distributional Learning Setup

In this section, we introduce necessary technical background, and formalize the learning setup that we consider in the following.

**Preliminaries** For a measurable space $(\mathcal{Z}, \mathcal{A}_{\mathcal{Z}})$, we denote the set of all probability distributions on it by $\mathcal{M}_1(\mathcal{Z})$, suppressing the $\sigma$-algebra if no confusion can arise, and the set of measurable real-valued functions is denoted by $\mathcal{L}^0(\mathcal{Z})$. If $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$ are measurable spaces, $f : \mathcal{X} \to \mathcal{Y}$ is a measurable map, and $\mu \in \mathcal{M}_1(\mathcal{X})$ is a probability measure, then the *pushforward of $\mu$ along $f$* is defined as $f\sharp\mu(A') = \mu(f^{-1}(A'))$ for all $A' \in \mathcal{A}_{\mathcal{Y}}$. For a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$, we denote the associated Borel $\sigma$-algebra by $\mathcal{B}(\tau_{\mathcal{X}})$. Given $\mu_n, \mu \in \mathcal{M}_1(\mathcal{X})$, $n \in \mathbb{N}_+$, we say that $(\mu_n)_n$ converges weakly[1] to $\mu$, if for all bounded and continuous $f : \mathcal{X} \to \mathbb{R}$, we have $\int_{\mathcal{X}} f(x)\mathrm{d}\mu_n(x) \to \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x)$.

[1] In the sense of probability theory, not functional analysis.

This induces a topology $\tau_w$ on $\mathcal{M}_1(\mathcal{X})$, called the topology of weak convergence. If $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a normed space, we define $\mathcal{B}(\mathcal{X})$ as the Borel $\sigma$-algebra generated by the open sets w.r.t. to the topology induced by the norm.

We briefly review some basics of kernels, and refer to (Steinwart & Christmann, 2008) for more background and pointers to the vast literature. Let $\mathcal{X} \neq \emptyset$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ some function. We call $k$ a *kernel* on $\mathcal{X}$, if for all $x_1, \ldots, x_N \in \mathcal{X}$ and $N \in \mathbb{N}_+$, the matrices $(k(x_i, x_j))_{i,j}$ are (symmetric) positive semidefinite. Let $(H, \langle\cdot,\cdot\rangle_H)$ be a Hilbert space of real-valued functions on $\mathcal{X}$. We call $k$ a *reproducing kernel of* $H$, if $k(\cdot, x) \in H$ for all $x \in \mathcal{X}$, and $f(x) = \langle f, k(\cdot, x)\rangle_H$ for all $f \in H$ and $x \in \mathcal{X}$. If $H$ has a reproducing kernel, the latter is unique, and we call $H$ a reproducing kernel Hilbert space (RKHS). For every kernel $k$, there is a unique RKHS for which $k$ is the reproducing kernel, and we denote this RKHS by $(H_k, \langle\cdot,\cdot\rangle_k)$, and the induced norm by $\|\cdot\|_k$. We also define $\|k\|_\infty = \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}$, and call $k$ bounded if it is bounded as a map on $\mathcal{X} \times \mathcal{X}$, which is the case if and only if $\|k\|_\infty < \infty$. Furthermore, the *canonical feature map* of the kernel $k$ is given by $\Phi_k : \mathcal{X} \to H_k$, $\Phi_k(x) = k(\cdot, x)$. Finally, a kernel $k$ on a compact metric space $\mathcal{X}$ is called universal if $H_k$ is dense (w.r.t. the supremum norm) in the set of continuous functions on $\mathcal{X}$.

Furthermore, in order to balance generality and simplicity of notation, we use *comparison functions*, a very successful formalism in control theory (Kellett, 2014). Define $\mathcal{K}$ as the set of continuous functions $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\alpha(0) = 0$ and $\alpha$ is strictly increasing. Operations and relations are declared pointwise on $\mathcal{K}$, so for $\alpha_1, \alpha_2 \in \mathcal{K}$, $\alpha_1 \leq \alpha_2$ means that $\alpha_1(s) \leq \alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$. Note that $\mathcal{K}$ is closed under addition and scalar multiplication with positive real numbers. Finally, we call $(\alpha_B)_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ a *nondecreasing family*, if $\alpha_a \leq \alpha_b$ for all $0 < a \leq b < \infty$.

**Statistical learning theory** We now introduce the statistical learning theory setup, closely following Chapters 2 and 5 in (Steinwart & Christmann, 2008). Let $\mathcal{X}$ be a measurable space, and let $\emptyset \neq \mathcal{Y} \subseteq \mathbb{R}$ be closed. A *loss function* $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a measurable function. We call $\ell$ convex, differentiable, etc., if for all $x \in \mathcal{X}, y \in \mathcal{Y}$ the function $\ell(x, y, \cdot)$ has the corresponding property. If $\ell$ is locally Lipschitz continuous, we define for all $B \in \mathbb{R}_{>0}$

$$|\ell|_{1,B} = \sup_{\substack{t_1,t_2\in[-B,B]\\ t_1\neq t_2\\ x\in\mathcal{X},y\in\mathcal{Y}}} \frac{|\ell(x,y,t_1) - \ell(x,y,t_2)|}{|t_1 - t_2|}. \quad (1)$$

Given $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $f : \mathcal{X} \to \mathbb{R}$ measurable, we define the *risk* $\mathcal{R}_{\ell,P}(f) = \int \ell(x, y, f(x))\mathrm{d}P(x, y)$ and the *Bayes risk* $\mathcal{R}^*_{\ell,P} = \inf_{f\in\mathcal{L}^0(\mathcal{X})} \mathcal{R}_{\ell,P}(f)$. Let $k$ be a kernel on $\mathcal{X}$ such that all functions in $H_k$ are measurable. For $f \in H_k$ and a *regularization parameter* $\lambda \in \mathbb{R}_{>0}$, we define the *regularized risk* $\mathcal{R}_{\ell,P,\lambda}(f) = \mathcal{R}_{\ell,P}(f) + \lambda\|f\|_k^2$, as well as $\mathcal{R}^{H_k*}_{\ell,P,\lambda} = \inf_{f\in H_k} \mathcal{R}_{\ell,P,\lambda}(f)$ and $\mathcal{R}^{H_k*}_{\ell,P} = \inf_{f\in H_k} \mathcal{R}_{\ell,P}(f)$. Additionally, if $\mathcal{R}^{H_k*}_{\ell,P} < \infty$, define the *approximation error function* $A^{(2)}_{\ell,P} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ by

$$A^{(2)}_{\ell,P}(\lambda) = \mathcal{R}^{H_k*}_{\ell,P,\lambda} - \mathcal{R}^{H_k*}_{\ell,P}. \quad (2)$$

Furthermore, define the *empirical risk* of a function $f \in H_k$ w.r.t. data $D = ((x_n, y_n))_{n=1,\ldots,N} \in (\mathcal{X} \times \mathcal{Y})^N$ by $\mathcal{R}_{\ell,D}(f) = \frac{1}{N}\sum_{n=1}^N \ell(x_n, y_n, f(x_n))$, and the *regularized empirical risk* $\mathcal{R}_{\ell,D,\lambda}(f) = \mathcal{R}_{\ell,D}(f) + \lambda\|f\|_k^2$. If it exists, a solution to the optimization problem

$$\min_{f\in H_k} \mathcal{R}_{\ell,D,\lambda}(f) \quad (3)$$

is called an *(empirical) SVM solution* and we denote it by $f_{D,\lambda}$. Similarly, if a solution to the optimization problem

$$\min_{f\in H_k} \mathcal{R}_{\ell,P,\lambda}(f) \quad (4)$$

exists, we called it an *infinite-sample SVM solution* or just *SVM solution*, and denote it by $f_{P,\lambda}$.

**Two-stage sampling setup** We now introduce the concrete distributional learning setup that we consider, roughly following (Szabó et al., 2015) and (Meunier et al., 2022). *Unless noted otherwise, this will be the setup that we use in the remainder of this work.* Let $(\mathcal{S}, \tau_{\mathcal{S}})$ be a topological space and consider the set of Borel probability measures $\mathcal{M}_1(\mathcal{S})$ on $\mathcal{S}$ . Let $\tau_w$ be the topology induced by weak convergence in $\mathcal{M}_1(\mathcal{S})$, and consider the measurable space $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$.

Let $\mathcal{H}$ be a Hilbert space, which we endow with the Borel $\sigma$-algebra $\mathcal{B}(\mathcal{H})$, let $\Pi : \mathcal{M}_1(\mathcal{S}) \to \mathcal{H}$ be some map, and define the Hilbertian semimetric $\mathrm{d}_{\mathcal{H}}(P, Q) = \|\Pi(P) - \Pi(Q)\|_{\mathcal{H}}$, and the set $\mathcal{X} = \Pi(\mathcal{M}_1(\mathcal{S}))$. Additionally, we assume access to a family of maps $(\hat{\Pi}_M)_{M\in\mathbb{N}_+}$ with $\hat{\Pi}_M : \mathcal{S}^M \to \mathcal{X}$, and we define $\mathcal{S}^* = \bigcup_{M\in\mathbb{N}_+} \mathcal{S}^M$ and $\hat{\Pi} : \mathcal{S}^* \to \mathcal{X}$ by $\hat{\Pi}(S) = \hat{\Pi}_M(S)$ for all $S \in \mathcal{S}^M$ and $M \in \mathbb{N}_+$. The usual example is $\hat{\Pi}(S) = \Pi\left(\frac{1}{M}\sum_{m=1}^M \delta_{S_m}\right)$ for $S \in \mathcal{S}^M$ and all $M \in \mathbb{N}_+$. However, our setup allows also different choices. For the analysis of this setting, measurability of various components needs to be ensured, for which the following assumption can be invoked.

**Assumption 2.1.** $\mathcal{H}$ is separable, $\Pi$ 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 technical result now ensures that we can apply the usual statistical learning theory setup.

**Lemma 2.2.** *Under Assumption 2.1, the map* $\Pi$ *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 $P_\Pi$ 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 remainder of this subsection, we work under Assumption 2.1. Let $P$ be a probability distribution on $\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y}$ (often called *meta-distribution*), *and to ease the notational load, in the following we will use $P$ also for the pushforward*[2] *$P_\Pi$, if no confusion can arise.* Furthermore, we assume the following data-generating model. We sample $(Q_1, y_1), \ldots, (Q_N, y_N)$ i.i.d. from $P$, and for each $n = 1, \ldots, N$, we assume that $S^{(n)} \sim Q_n^{\otimes M_n}$ for some $M_n \in \mathbb{N}_+$, and that $S^{(1)}, \ldots, S^{(N)}$ are also independent. The data sets used for training are then of the form $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\ldots,N} \in (\mathcal{S}^* \times \mathcal{Y})^N$. Furthermore, we define

$$\mathcal{D}_{\hat{\Pi}} = ((\hat{\Pi}(S^{(n)}), y_n))_{n=1,\ldots,N}$$

and for $\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\ldots,N} \in (\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})^N$, define

$$\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\ldots,N} \in (\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})^N$$
$$\bar{\mathcal{D}}_\Pi = ((\Pi Q_n, y_n))_{n=1,\ldots,N} \in (\mathcal{X} \times \mathcal{Y})^N.$$

To summarize, we have to deal with two sampling stages. First, sampling input-output pairs $(Q, y) \sim P$, and then sampling from the distributions $Q$. Let now $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function and $k$ a kernel on $\mathcal{X}$. Given a data set $\mathcal{D}$ as above, consider the regularized empirical risk minimization problem

$$\min_{f \in H_k} \mathcal{R}_{\ell, \mathcal{D}_{\hat{\Pi}}, \lambda}(f) \tag{5}$$

where $\lambda \in \mathbb{R}_{>0}$ is the regularization parameter. If a solution $f_{\mathcal{D}_{\hat{\Pi}}, \lambda}$ to (5) exists, it can be used for a prediction task with distributional inputs by composing it with the map $\Pi$, so given input $Q \in \mathcal{M}_1(\mathcal{S})$, it leads to prediction $f_{\mathcal{D}_{\hat{\Pi}}, \lambda}(\Pi Q)$. Using Assumption 2.1 and Lemma 2.2, we can now consider various risks[3] like $\mathcal{R}_{\ell, P, \lambda}(f_{\mathcal{D}_{\hat{\Pi}}, \lambda})$.

*Remark* 2.3. Note that $\mathcal{X}$ is a subset of a Hilbert space $\mathcal{H}$, so in order to implement the approach just described, we need kernels $k$ on (subsets of) Hilbert spaces. On the one hand, *any* such kernel can in principle be used for this task, cf. (Christmann & Steinwart, 2010) for some examples. On the other hand, constructing kernels on (potentially infinite-dimensional) Hilbert spaces can be challenging. To tackle this, (Meunier et al., 2022) suggested a general framework based on *distance substitution kernels* (Haasdonk & Bahlmann, 2004). The Hilbertian embedding $(\mathcal{H}, \Pi)$ is used to construct a kernel on probability distributions by defining $k(P, Q) = \phi(\|\Pi P - \Pi Q\|_{\mathcal{H}})$, where $\phi$ is a function that induces a radial kernel. Note that all of our general results immediately apply to this framework, covering for example sliced 1- and 2-Wasserstein distances and the induced distance substitution kernels. For details and concrete examples, we refer to (Meunier et al., 2022).

**Special case: Kernel mean embeddings** The first works on distributional learning using Hilbertian embeddings relied on kernel mean embeddings (KMEs). We summarize the necessary background in the following result.

**Proposition 2.4.** *Let $(\mathcal{S}, \mathcal{A}_\mathcal{S})$ be a measurable space, and $\kappa$ a measurable and bounded kernel on $\mathcal{S}$.*

1. *The map*

$$\Pi_\kappa : \mathcal{M}_1(\mathcal{S}) \to H_\kappa,\ \Pi_\kappa Q = \int \kappa(\cdot, s) \mathrm{d}Q(s) \tag{6}$$

*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. $\kappa$.*

2. *Define $\hat{\Pi}_\kappa : \mathcal{S}^* \to H_\kappa$ by*

$$\hat{\Pi}_\kappa((s_1, \ldots, s_M)) = \frac{1}{M} \sum_{m=1}^{M} \kappa(\cdot, s_m). \tag{7}$$

*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*

$$\|\hat{\Pi}_\kappa S - \Pi_\kappa Q\|_\kappa \leq 2\sqrt{\frac{\|\kappa\|_\infty^2}{M}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(1/\delta)}{M}} \tag{8}$$

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

3. *Let $(\mathcal{S}, \tau_\mathcal{S})$ be a separable topological space, choose $\mathcal{A}_\mathcal{S} = \mathcal{B}(\tau_\mathcal{S})$, and assume that $\kappa$ is continuous. Then $\Pi_\kappa$ is $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$-$(H_\kappa, \mathcal{B}(H_\kappa))$-measurable, where $\tau_w$ is the topology induced by weak convergence in $\mathcal{M}_1(\mathcal{S})$.*

A proof can be found in Appendix A. This result allows to use KMEs as the Hilbertian embedding, i.e., setting $\mathcal{H} = H_\kappa$, $\Pi = \Pi_\kappa$ and $\hat{\Pi} = \hat{\Pi}_\kappa$.

[2] Formally, $P_\Pi = g\sharp P$, where the measurable map $g : \mathcal{M}_1(\mathcal{S}) \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ is defined by $g(Q, y) = (\Pi Q, y)$.

[3] Note that we tacitly assume that the learning methods induced by the regularized (empirical) risk minimization problems are *measurable* learning methods. In the setting we consider, this does not pose a problem, cf. the thorough discussion in Chapter 6 in (Steinwart & Christmann, 2008).

## 3. Oracle Inequalities

Oracle inequalities are important tools in modern statistical learning theory (Steinwart & Christmann, 2008). Roughly speaking, they are concentration inequalities for the excess risk of the learning outcome over the risk that is achieved by an oracle which has access to the true underlying distribution. In particular, oracle inequalities provide finite-sample guarantees, and can be used to derive consistency of a learning method, as well as (under additional assumptions on the data-generating distribution) learning rates. We now present our two oracle inequalities for the risk of SVMs in the distributional learning setting. The first one is based on a form of Lipschitz-continuity of SVMs, and can be interpreted as a distributional analogon of Theorem 6.24 in (Steinwart & Christmann, 2008).

**Theorem 3.1.** *Let Assumption 2.1 hold. Assume that $\ell$ is convex, differentiable, and that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x,y,0) \leq B_\ell$. Furthermore, assume that there exists $B'_\ell \in \mathbb{R}_{\geq 0}$ such that $|\ell'(x,y,0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and that there exist $\gamma_1 \in \mathcal{K}$ and a nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that $|\ell'(x_1,y,t_1) - \ell'(x_2,y,t_2)| \leq \gamma_1(\|x_1 - x_2\|) + \gamma_{3,B}(|t_1 - t_2|)$ for all $B \in \mathbb{R}_{>0}$, $x_1, x_2 \in \mathcal{X}$, $y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$. Let $k$ be a kernel on $\mathcal{H}$ that is measurable, bounded, has a separable RKHS $H_k$, and assume that there exists[4] $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. Finally, assume that for all $n = 1, \ldots, N$, there exists $B_n : (0,1) \to \mathbb{R}_{\geq 0}$ such that $\mathbb{P}[\|\hat{\Pi}(S^{(n)}) - \Pi(Q_n)\|_{\mathcal{H}} > B_n(\delta)] < \delta$ for all $\delta \in (0,1)$. We then have for all $\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0,1)$ that with probability at least $1-\delta$*

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) \\
&+ \frac{2\sqrt{B_\ell/\lambda} + |\ell|_{1,B_f}\|k\|_\infty/\lambda}{N} \sum_{n=1}^{N} \alpha_\lambda(B_n(\delta/(2N))) \\
&+ 2\frac{|\ell|_{1,B_f}\|k\|^2_\infty}{\lambda}\left(B'_\ell + \gamma_{3,B_f}\left(\|k\|_\infty\sqrt{B_\ell/\lambda}\right)\right) \\
&\times \left(\sqrt{\frac{2\ln(2/\delta)+1}{N}} + \frac{4\ln(2/\delta)}{3N}\right),
\end{aligned}$$

*where we defined $B_f = \|k\|_\infty\sqrt{B_\ell/\lambda}$ and $\alpha_\lambda = \|k\|_\infty(\gamma_1 + \gamma_{3,B_f} \circ \left(\sqrt{B_\ell/\lambda}\alpha_k\right)) + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\alpha_k$.*

The functions $B_n$ in the statement of the result are used to provide estimation bounds of the Hilbertian embeddings of the input distributions, i.e., how close $\hat{\Pi}S^{(n)}$ (which can be computed from data) is to $\Pi Q_n$ (which cannot be computed from data). In particular, the functions $B_n$ depend on $M_n$, but we suppressed this dependency to ease the notation. Similarly, $\alpha_\lambda \in \mathcal{K}$ describes (up to a multiplicative factor) how the estimation error of the Hilbertian embedding that arises from a single data set item, influences the risk. Using this abstraction allows us to formulate our results for *any* Hilbertian embedding approach. Specializing to a concrete embedding approach then boils down to checking the well-posedness assumptions (cf. Assumption 2.1), and replacing the $B_n$ by concrete estimation bounds. While this approach makes Theorem 3.1 (and similarly Theorem 3.4 presented below) broadly applicable to various Hilbertian embedding methods, as a result the bounds do not directly help in choosing an appropriate embedding.

[4]The latter condition implies that $\Phi_k$ is continuous, which implies that $k$ is continuous, which in turn implies that $k$ is measurable and has a separable RKHS. However, we kept these two conditions for emphasis.

*Proof sketch for Theorem 3.1.* The basic idea is to apply the proof strategy of Theorem 6.24 in (Steinwart & Christmann, 2008) to the ideal, but inaccessible data set $\bar{\mathcal{D}}_\Pi$, and then use estimation error bounds for the Hilbertian embeddings (encoded by the functions $B_n$) to translate this to the accessible data set $\mathcal{D}_{\hat{\Pi}}$. To do so, we use a known generalized Representer Theorem (recalled as Proposition B.4 in the appendix) together with the continuity property of the canonical feature map and the regularity and boundedness properties of the loss function, which allows us to propagate the estimation error through the bounds. A detailed proof is provided in Section B.2 in the appendix. □

**Example 3.2.** Let us provide some concrete examples for the ingredients of the preceding result. For instance, consider loss functions of the form $\ell(x,y,t) = \psi(y-t)$ (which are called *distance-based supervised losses* in (Steinwart & Christmann, 2008)), and assume that $\psi$ is continuously differentiable and that $\mathcal{Y} \subseteq [-M, M]$ for some $M \in \mathbb{R}_{>0}$. In this case, suitable constants $B_\ell$ and $B'_\ell$ exist, and we can choose an arbitrary $\gamma_1 \in \mathcal{K}$ (since $\ell$ does not depend on the first argument) and $\gamma_{3,B}(s) = C_B s$ for suitable constants $C_B \in \mathbb{R}_{>0}$. An example of the condition on $\Phi_k$ is given by Hölder-continuity of the canonical feature map $\Phi_k$, which has been used in previous works like (Szabó et al., 2015). This means that there exist $C_k \in \mathbb{R}_{>0}$, $\alpha \in (0,1]$, such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_{\mathcal{H}} \leq C_k\|x_1 - x_2\|^\alpha$ for all $x_1, x_2 \in \mathcal{X}$, and we can set $\alpha_k(s) = C_k s^\alpha$.

When using KMEs for the Hilbertian embedding, we get the following oracle inequality as a special case.

**Corollary 3.3.** *Let $\mathcal{S}$ be a compact metric space, $\kappa$ be a measurable, bounded, continuous and universal kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_\kappa$, and $\hat{\Pi} = \hat{\Pi}_\kappa$. Assume that $\ell$ is convex, differentiable, $\ell'$ is locally Lipschitz continuous, and that there exists $B_\ell, B'_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x,y,0) \leq B_\ell$ and $|\ell'(x,y,0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k$ be a universal kernel on $\mathcal{H}$ that is measurable and bounded, and that there exists $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. We then have for all*

*$\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0,1)$ that with probability at least $1-\delta$*

$$\begin{aligned}\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^*_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) \\ &+ \frac{2\sqrt{\lambda B_\ell} + |\ell|_{1,B_f}\|k\|_\infty}{N} \\ &\times \sum_{n=1}^{N} \alpha_\lambda \left(2\sqrt{\frac{\|\kappa\|^2_\infty}{M_n}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(2N/\delta)}{M_n}}\right) \\ &+ 2|\ell|_{1,B_f}\|k\|_\infty \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \\ &\times \left(\sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N}\right),\end{aligned}$$

*with $B_f$ and $\alpha_\lambda$ as in Theorem 3.1.*

Theorem 3.1 puts strong regularity requirements on the loss function, but needs only mild assumptions for the kernel used in the empirical risk minimization. The following oracle inequality, a distributional analogon of Theorem 6.25 from (Steinwart & Christmann, 2008), is complementary, putting only mild requirements on the loss function, but strong structural results on the RKHS are used.

**Theorem 3.4.** *Assume that $\ell$ is convex, that there exist $\gamma_1 \in \mathcal{K}$ and a nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}$, $y \in \mathcal{Y}$, and all $B \in \mathbb{R}_{>0}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$, it holds that $|\ell(x_1, y, t_1) - \ell(x_2, y, t_2)| \leq \gamma_1(\|x_1 - x_2\|_{\mathcal{H}}) + \gamma_{3,B}(|t_1 - t_2|)$, and that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x, y, 0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k$ be a kernel on $\mathcal{X}$ that is measurable, bounded, and has a separable RKHS $H_k$. Assume that there exists an nondecreasing family $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for $B \in \mathbb{R}_{>0}$ and $f \in H_k$ with $\|f\|_k \leq B$, we have $|f(x_1) - f(x_2)| \leq \alpha_{f,B}(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. Furthermore, let $\epsilon, \lambda \in \mathbb{R}_{>0}$, and let $\mathcal{F} \subseteq H_k$ be a finite set such that for all $f \in H_k$ with $\|f\|_k \leq \sqrt{B_\ell/\lambda}$ there exists $\tilde{f} \in \mathcal{F}$ with $\|f - \tilde{f}\|_\infty \leq \epsilon$. Finally, assume that for all $n = 1, \ldots, N$, there exists $B_n : (0,1) \to \mathbb{R}_{\geq 0}$ such that $\mathbb{P}[\|\hat{\Pi}(S^{(n)}) - \Pi(Q_n)\|_{\mathcal{H}} > B_n(\delta)] < \delta$ for all $\delta \in (0,1)$. We then have for all $\delta \in (0,1)$ that with probability at least $1-\delta$ it holds that*

$$\begin{aligned}\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) + 4\gamma_{3,\tilde{B}_f}(\epsilon) \\ &+ \frac{2}{N}\sum_{n=1}^{N} \alpha_\lambda \left(B_n\left(\frac{\delta}{N+|\mathcal{F}|}\right)\right) \\ &+ 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f)\right)\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}},\end{aligned}$$

*where we defined $\tilde{B}_f = \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} + \epsilon$ and $\alpha_\lambda = \gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}$.*

The central assumption of Theorem 3.4 is the existence of a suitable discretization $\mathcal{F}$ of $\bar{B}^{H_k}_{\sqrt{B_\ell/\lambda}}$, the closed centered ball with radius $\sqrt{B_\ell/\lambda}$ in the RKHS $H_k$. Under suitable assumptions, a *finite* $\mathcal{F}$ exists, and one can set $|\mathcal{F}| = \mathcal{N}(\bar{B}^{H_k}_{\sqrt{B_\ell/\lambda}}, \|\cdot\|_\infty, \epsilon)$, where $\mathcal{N}(T, d, \epsilon)$ is the $\epsilon$-*covering number* of a metric space $(T, d)$. For more details and pointers to the literature, we refer to Chapters 6, 7 in (Steinwart & Christmann, 2008).

*Proof sketch for Theorem 3.4.* Similarly to the proof of Theorem 3.1, we apply the proof strategy of Theorem 6.25 in (Steinwart & Christmann, 2008) to the ideal, but inaccessible data set $\bar{\mathcal{D}}_\Pi$, and translate the result to the accessible data set $\mathcal{D}_{\hat{\Pi}}$ by the estimation bounds described by the functions $B_n$, using the continuity and boundedness properties of the loss function (which can be milder now, since we do not use Proposition B.4 anymore) and the canonical feature map. A detailed proof is provided in Section B.2 in the appendix. □

**Example 3.5.** A sufficient condition for the existence of the nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ is Hölder-continuity. If $\mathrm{d}_{\mathcal{H}}(\mu,\nu) = \|(\Phi_k \circ \Pi)(\mu) - (\Phi_k \circ \Pi)(\nu)\|_k$, then it is well-known that one can choose $\alpha_{f,B}(s) = Bs$. If there exists $C_k, \alpha_k \in \mathbb{R}_{>0}$ such that $|k(x_1,x) - k(x_2,x)| \leq C_k\|x_1 - x_2\|^{\alpha_k}_{\mathcal{H}}$ for all $x_1, x_2 \in \mathcal{X}$, then one can choose $\alpha_{f,B}(s) = \sqrt{2C_k}s^{\alpha_k/2}$. For proofs of these facts and more general conditions, we refer to (Fiedler, 2023).

We can immediately specialize Theorem 3.4 to the case of KMEs for the Hilbertian embedding.

**Corollary 3.6.** *Consider the situation of Theorem 3.4. Additionally, let $\mathcal{S}$ be a compact metric space, $\kappa$ be a measurable, bounded, continuous and universal kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_\kappa$, and $\hat{\Pi} = \hat{\Pi}_\kappa$. We then have for all $\delta \in (0,1)$, that*

$$\begin{aligned}\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) + 4\gamma_{3,\tilde{B}_f}(\epsilon) \\ &+ 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f)\right)\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} \\ &+ \frac{2}{N}\sum_{n=1}^{N} \alpha_\lambda \left(2\sqrt{\frac{\|k\|^2_\infty}{M}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(\frac{N+|\mathcal{F}|}{\delta})}{M}}\right),\end{aligned}$$

*holds with probability at least $1-\delta$, with $\tilde{B}_f$ and $\alpha_\lambda$ as in Theorem 3.4.*

The proof is completely analogous to the one of Corollary 3.3.

## 4. Stability-based Generalization Bound

The oracle inequalities from the previous section allow us to compare the risk of the learned hypothesis (i.e., of the empirical SVM solution) to the minimum risk that could be achieved by an oracle (having access to the true underlying

meta-distribution). We now consider a slightly different question: How accurate is the empirical risk of the learned hypothesis (which can be computed from data) as an estimate of the true risk of the learned hypothesis (which cannot be computed, since we do not know the true underlying data-generating distribution)? In other words, how well does the learned hypothesis generalize from the training data to the population, as measured by its risk?

We investigate this using a variation of our basic setup. Let $(Q, y) \sim P$ as before, but now assume that the number of samples from $Q$ (collected in $S$) is also random. Denote the joint distribution of $(Q, S, y)$ by $\bar{P}$, the marginal distribution of $(S, y)$ by $\tilde{P}$, and the number of samples in $S$ by $M$, an $\mathbb{N}_+$-valued random variables. A special case covered by this setup is a constant $M$, a setting which is often considered in related works like (Szabó et al., 2015) or (Meunier et al., 2022). The data set $\mathcal{D}$ is therefore now generated by sampling $(Q_1, S^{(1)}, y), \ldots, (Q_N, S^{(N)}, y_N)$ i.i.d. from $\bar{P}$, and then setting $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\ldots,N}$, hence $\mathcal{D} \sim \tilde{P}^{\otimes N}$.

The generalization bounds that follow are based on the concept of algorithmic stability (Bousquet & Elisseeff, 2002), which applies to very general learning methods. A learning method is a map[5] $\bigcup_{N\in\mathbb{N}_+} (\mathcal{X} \times \mathcal{Y})^N \ni D \mapsto \mathcal{L}_D$, where $\mathcal{L}_D : \mathcal{X} \to \mathbb{R}$ is measurable. We call $\mathcal{L}$ $\beta$-stable (w.r.t. the loss function $\ell$) if there exists $(\beta_N)_{N\in\mathbb{N}_+}$, $\beta_N \in \mathbb{R}_{\geq 0}$, such that for all $N \in \mathbb{N}_+$, $x \in \mathcal{X}, y \in \mathcal{Y}$ we have

$$|\ell(x, y, \mathcal{L}_D(x)) - \ell(x, y, \mathcal{L}_{\tilde{D}}(x))| \leq \beta_N, \tag{9}$$

for all $D, \tilde{D} \in (\mathcal{X} \times \mathcal{Y})^N$ such that there exists $1 \leq i \leq N$ with $D_n = \tilde{D}_n$, $n \in \{1, \ldots, N\} \setminus \{i\}$. In other words, a learning method is $\beta$-stable, if changing just one sample in a data set of size $N \in \mathbb{N}_+$, changes the loss of the resulting hypothesis by at most $\beta_N$. We are now ready to present the announced generalization result. It is a distributional-input analogon of Theorem 14.2 in (Mohri et al., 2018).

**Theorem 4.1.** *Consider a $\beta$-stable learning method $\mathcal{L}$. Assume that there exists a concave $\alpha \in \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$ and all $D \in (\mathcal{S}^* \times \mathcal{Y})^N$ we have $|\ell(x_1, y, \mathcal{L}_D(x_1)) - \ell(x_2, y, \mathcal{L}_D(x_2))| \leq \alpha(\|x_1 - x_2\|_{\mathcal{H}})$. We then have for all $\delta \in (0, 1)$ that with probability at least $1 - \delta$, the bound*

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}) \leq \mathcal{R}_{\ell,\mathcal{D}_{\Pi}}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}) + (2N\beta_N + B)\sqrt{\frac{\ln(1/\delta)}{2N}} + \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right]\right) + \beta_N$$

*holds.*

The proof of this result can be found in Appendix C.2. We now present and prove a generalization bound for SVMs in the two-stage sampling setup, which is based on Theorem 4.1.

**Theorem 4.2.** *Let $\ell$ be convex, locally Lipschitz continuous, and assume that there exists $\gamma_1 \in \mathcal{K}$ such that $|\ell(x_1, y, t) - \ell(x_2, y, t)| \leq \gamma_1(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$ and $t \in \mathbb{R}$. Furthermore, assume that there exists $B_\ell \in \mathbb{R}_{>0}$ such that $\ell(x, y, 0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k$ be measurable and bounded, and assume that there exists a nondecreasing family $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}$, $B \in \mathbb{R}_{>0}$, and all $f \in H_k$ with $\|f\|_k \leq B$ we have $|f(x_1) - f(x_2)| \leq \alpha_{f,B}(\|x_1 - x_2\|_{\mathcal{H}})$. Assume that for $\lambda \in \mathbb{R}_{>0}$, there exists a concave $\alpha_\lambda \in \mathcal{K}$ with $\gamma_1 + |\ell|_{1,B_f}\alpha_{f,\sqrt{B_\ell/\lambda}} \leq \alpha_\lambda$, where we defined $B_f = \|k\|_\infty\sqrt{B_\ell/\lambda}$. We then have for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, that*

$$\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \frac{|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda N} + \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right]\right) + \left(\frac{2|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda} + B_\ell + |\ell|_{1,B_f}B_f\right)\sqrt{\frac{\ln(1/\delta)}{2N}}.$$

Before turning to the proof of Theorem 4.2, we describe two example classes of suitable $\ell$ and $\alpha_\lambda$.

**Example 4.3.** Assume that there exist $(C_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathbb{R}_{>0}$, $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq (0, 1]$ such that $|f(x_1) - f(x_2)| \leq C_{f,B}\|x_1 - x_2\|_\Pi^{\alpha_{f,B}}$ for all $B \in \mathbb{R}_{>0}$, $x_1, x_2 \in \mathcal{X}$, and $f \in H_k$ with $\|f\|_k \leq B$. Let us call a function $\phi : \mathbb{R} \to \mathbb{R}$ locally Hölder-continuous, if there exist $(C_{\phi,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathbb{R}_{>0}$, $(\alpha_{\phi,B})_{B\in\mathbb{R}_{>0}} \subseteq (0, 1]$, such that for all $B \in \mathbb{R}_{>0}$, $|\phi(s_1) - \phi(s_2)| \leq C_{\phi,B}|s_1 - s_2|^{\alpha_{\phi,B}}$ for all $s_1, s_2 \in [-B, B]$. We refer to (Fiedler, 2023) for a discussion of these properties, including characterizations of suitable $k$. (i) Assume that $\ell(x, y, t) = \psi(y - t)$, where $\psi$ is a nonnegative, locally Hölder-continuous function. Given $\lambda \in \mathbb{R}_{>0}$, we can then choose $\alpha_\lambda(s) = C_{\psi,\|k\|_\infty\sqrt{B_\ell/\lambda}}C^{\alpha_\psi}_{f,\sqrt{B_\ell/\lambda}}s^{\alpha_\psi\alpha_f}$ with $\alpha_\psi = \alpha_{\psi,\|k\|_\infty\sqrt{B_\ell/\lambda}}$ and $\alpha_f = \alpha_{f,\sqrt{B_\ell/\lambda}}$. (ii) Assume that $\ell(x, y, t) = \varphi(yt)$ (called a *margin-based loss function* in (Steinwart & Christmann, 2008)) for a nonnegative, locally Hölder-continuous function, and that $\mathcal{Y} \subseteq [-M, M]$ for some $M \in \mathbb{R}_{>0}$. Given $\lambda \in \mathbb{R}_{>0}$, we can then choose $\alpha_\lambda(s) = C_\varphi M^{\alpha_\varphi}C^{\alpha_\varphi}_{f,\sqrt{\varphi(0)/\lambda}}s^{\alpha_\varphi\alpha_f}$ with $C_\varphi = C_{\varphi,M\|k\|_\infty\sqrt{\varphi(0)/\lambda}}$, $\alpha_\varphi = \alpha_{\varphi,M\|k\|_\infty\sqrt{\varphi(0)/\lambda}}$, and $\alpha_f = \alpha_{f,\sqrt{\varphi(0)/\lambda}}$.

*Proof of Theorem 4.2.* Let $Q$ be a distribution on $\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y}$. From Lemma A.4 we have $|f_{Q,\lambda}(x)| \leq$

[5] In the present setting, it is safe to ignore measurability issues, cf. the discussion in Chapter 6 in (Steinwart & Christmann, 2008).

$\|k\|_\infty \|f_{Q,\lambda}\|_k \leq \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}} = B_f$, so

$$\begin{aligned}
&|\ell(x_1, y, f_{Q,\lambda}(x_1)) - \ell(x_2, y, f_{Q,\lambda}(x_2))| \\
&\leq \gamma_1(\|x_1 - x_2\|_{\mathcal{H}}) + |\ell|_{1,B_f} |f_{Q,\lambda}(x_1) - f_{Q,\lambda}(x_2)| \\
&\leq \left(\gamma_1 + |\ell|_{1,B_f} \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(\|x_1 - x_2\|_{\mathcal{H}}) \\
&\leq \alpha_\lambda(\|x_1 - x_2\|_{\mathcal{H}}),
\end{aligned}$$

hence $\alpha_\lambda$ fulfills the requirements of Theorem 4.1. Furthermore, as before we have $\ell(x, y, f_{Q,\lambda}(x)) \leq B_\ell + |\ell|_{1,B_f} B_f = B$.

Next[6], for all $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$ and all $x \in \mathcal{X}, y \in \mathcal{Y}$, we have $|\ell(x, y, f(x)) - \ell(x, y, g(x))| \leq |\ell|_{1,B_f} |f(x) - g(x)|$. An inspection of the proof of Proposition 14.4 in (Mohri et al., 2018) then shows that the learning method $(\mathcal{X} \times \mathcal{Y})^N \ni D \mapsto f_{\ell,D}\lambda \in H_k$ is $\beta_N = \frac{|\ell|^2_{1,B_f} \|k\|^2_\infty}{\lambda N}$ stable.

The result now follows from Theorem 4.1. □

*Remark* 4.4. An inspection of the proof of Theorem 4.2 and how Proposition 14.4 in (Mohri et al., 2018) is used there, reveals that instead of local Lipschitz continuity of $\ell$ the following continuity property is sufficient: There exists a family $(C_B, p_B)_{B \in \mathbb{R}_{>0}}$ with $C_B \in \mathbb{R}_{>0}$, $0 < p_B < 2$ for all $B \in \mathbb{R}_{>0}$, such that for all $x \in \mathcal{X}, y \in \mathcal{Y}$, $B \in \mathbb{R}_{>0}$ and all $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$ we have that

$$|\ell(x, y, t_1) - \ell(x, y, t_2)| \leq C_B |t_1 - t_2|^{p_B}.$$

Furthermore, we now need a concave $\alpha_\lambda \in \mathcal{K}$ such that $\gamma_1 + C_B \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}(\cdot)^{p_B} \leq \alpha_\lambda$ with $B = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}$. In this case, we have

$$\beta_N = C_B^{1+\frac{1}{2-p_B}} \|k\|_\infty^{p_B + \frac{p_B}{2-p_B}} \left(\frac{1}{N\lambda}\right)^{\frac{1}{2-p_B}}.$$

Once again, we can immediately specialize to the case of using KMEs for the Hilbertian embedding.

**Corollary 4.5.** *Consider the situation of Theorem 4.2. Additionally, let $\mathcal{S}$ be a compact metric space, $\kappa$ be a measurable, bounded, continuous and universal kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_\kappa$, and $\hat{\Pi} = \hat{\Pi}_\kappa$. We then have for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, that*

$$\begin{aligned}
\mathcal{R}_{\ell,P}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) &\leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) + \alpha_\lambda\left(\frac{\sqrt{2\|\kappa\|_\infty}}{\mathbb{E}[\sqrt{M}]}\right) \\
&+ \left(\frac{2|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda} + B_\ell + |\ell|_{1,B_f} B_f\right)\sqrt{\frac{\ln(1/\delta)}{2N}} \\
&+ \frac{|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda N},
\end{aligned}$$

*where we defined $B_f = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}$.*

*Proof.* First, as in the proof of Corollary 3.3, the KME setup fulfills Assumption 2.1. Let $Q \in \mathcal{M}_1(\mathcal{S})$, $M \in \mathbb{N}_+$, and $S \sim Q^{\otimes M}$. According to Lemma 4 in (Gretton et al., 2012), $\|\Pi_\kappa Q - \hat{\Pi}_\kappa S\|_\kappa$ is the maximum mean discrepancy between $Q$ and the empirical measure $\frac{1}{M}\sum_{m=1}^M \delta_{S_m}$, so we get from Equation (19) in the same reference that

$$\mathbb{E}_{S \sim Q^{\otimes M}}\left[\|\Pi_\kappa Q - \hat{\Pi}_\kappa S\|_\kappa\right] \leq \sqrt{\frac{2\|\kappa\|_\infty}{M}},$$

which implies that

$$\begin{aligned}
\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right] &\leq \mathbb{E}\left[\sqrt{\frac{2\|\kappa\|_\infty}{M}}\right] \\
&= \frac{\sqrt{2\|\kappa\|_\infty}}{\mathbb{E}[\sqrt{M}]}.
\end{aligned}$$

Combining this with Theorem 4.2 and using that $\alpha_\lambda$ is increasing, establishes the result. □

## 5. Conclusion

In this work, we continued the investigation of kernel-based statistical learning with distributional inputs from the perspective of modern statistical learning theory. To the best of our knowledge, we provided the first general oracle inequalities in this setting, complementing the existing excess risk bounds for distributional regression using kernel ridge regression. In particular, our analysis covers rather general loss functions encoding a multitude of learning scenarios. Additionally, we provided generalization bounds based on algorithmic stability, a result and setting which has not been analyzed at all in the distributional learning literature. We formulated all of this in a very general setup based on Hilbertian embeddings of probability distributions. On the one hand, in this manner the kernel construction approach from (Meunier et al., 2022) is applicable, and on the other hand, our main results apply directly to any existing and future embedding approach. For example, if appropriate estimation tools become available, our results will be directly applicable to the recently introduced kernel cumulant embeddings

[6] The following is a generalization of the property from Definition 14.3 in (Mohri et al., 2018).

(Bonnier et al., 2023). Finally, we provided specializations of our results to the important cases of KMEs as well as the recent sliced 2-Wasserstein distances.

Many relevant questions are still open, and our results form the starting point for a multitude of future investigations. First, while oracle inequalities can be used to derive consistency results, in order to guarantee learning rates, one needs suitable assumptions to derive bounds on the approximation error function. Finding such conditions in the present setting is an important open problem. Second, while the setting of our main results is rather general, we need various boundedness assumptions on the loss functions. Removing these assumptions, or replacing them by clippability (cf. Chapters 2 and 9 in (Steinwart & Christmann, 2008)), is another interesting problem. Third, both of our oracle inequalities are based on classic arguments, and it is known, cf. Chapter 7 in (Steinwart & Christmann, 2008), that using more advanced tools from empirical process theory, one can derive sharper oracle inequalities, which eventually can lead to better learning rates. We expect that this applies also in the distributional setting, and that the resulting analysis approach for kernel ridge regression from (Steinwart et al., 2009) then provides an alternative to the integral operator technique from (Caponnetto & De Vito, 2007), which so far was the main focus in the distributional regression literature.

## Acknowledgements

We would like to thank Antonia Holzapfel for a careful reading of the manuscript, and the anonymous reviewers for their detailed and helpful comments and questions.

## Impact statement

This paper presents work whose goal is to advance the field of Machine Learning, in particular, contributing to the theoretical analysis of kernel-based statistical learning with distributional inputs. There are many potential societal consequences of our work, but we do not think that any warrant a detailed discussion here. While the learning methods we analyze are highly relevant for several real-world applications, it is very unlikely that our theoretical work, which aims at a better understanding of statistical aspects of these methods, leads to negative societal impact.

## References

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Bonnier, P., Oberhauser, H., and Szabó, Z. Kernelized cumulants: Beyond kernel mean embeddings. *Advances in Neural Information Processing Systems*, 2023.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

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

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

Fiedler, C. Lipschitz and Hölder continuity in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2310.18078*, 2023.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Haasdonk, B. and Bahlmann, C. Learning with distance substitution kernels. In *Joint pattern recognition symposium*, pp. 220–227. Springer, 2004.

Kellett, C. M. A compendium of comparison function results. *Mathematics of Control, Signals, and Systems*, 26: 339–374, 2014.

Lin, T., Zheng, Z., Chen, E., Cuturi, M., and Jordan, M. I. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pp. 262–270. PMLR, 2021.

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

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2 edition, 2018.

Nietert, S., Goldfeld, Z., Sadhu, R., and Kato, K. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

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

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

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

Steinwart, I., Hush, D. R., Scovel, C., et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

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

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

## A. Additional Technical Background

**Comparison functions** In addition to $\mathcal{K}$, we define

$$\mathcal{L} = \{\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \mid \rho \text{ continuous, strictly decreasing, } \lim_{s\to\infty} \rho(s) = 0\}.$$

Observe that if $\rho \in \mathcal{L}$, then $\rho(s) > 0$ for all $s \in \mathbb{R}_{\geq 0}$, and its inverse is defined on its range, i.e., $\rho^{-1} : (0, \rho(0)] \to \mathbb{R}_{\geq 0}$. We define addition and scalar multiplication in $\mathcal{K}$ and $\mathcal{L}$ pointwise, i.e., if $\alpha_1, \alpha_2 \in \mathcal{K}$ (respectively, $\mathcal{L}$), then $\alpha_1 + \alpha_2$ is defined by $(\alpha_1 + \alpha_2)(s) = \alpha_1(s) + \alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$, and if $c_1 \in \mathbb{R}_{>0}$, then $c_1\alpha_1$ is defined by $(c_1\alpha_1)(s) = c_1\alpha_1(s)$. Note that $c_1\alpha_1 + \alpha_2 \in \mathcal{K}$ (respectively, in $\mathcal{L}$), so $\mathcal{K}$ and $\mathcal{L}$ form a cone. Furthermore, $\alpha_1 \circ \alpha_2 \in \mathcal{K}$. We also define comparison relations pointwise, e.g., if $\alpha_1, \alpha_1 \in \mathcal{K}$, then $\alpha_1 \leq \alpha_2$ means that $\alpha_1(s) \leq \alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$. For more background on comparison functions, including historical remarks and application examples, we refer to (Kellett, 2014).

**More on loss functions** For technical reasons, we need some additional concepts from (Steinwart & Christmann, 2008). A loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a *Nemitskii loss*, if there exists a measurable function $b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and an increasing function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and all $t \in \mathbb{R}$ we have

$$\ell(x, y, t) \leq b(x, y) + h(|t|).$$

Let $P$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. We call $\ell$ a *$P$-integrable Nemitskii loss*, if it is a Nemitskii loss, and the function $b$ from the definition of this concept is $P$-integrable.

**Boundedness in RKHSs** For convenience, we summarize some well-known results on boundedness of kernels and RKHS functions.

**Lemma A.1** (Boundedness in RKHSs)**.** *Let $\mathcal{X}$ be an arbitrary nonempty set and $k$ a kernel on it, and define*

$$\|k\|_\infty = \sup_{x\in\mathcal{X}} \sqrt{k(x, x)}. \tag{10}$$

1. *$k$ is bounded if and only if $\|k\|_\infty < \infty$.*
2. *All $f \in H_k$ are bounded if and only if $k$ is bounded.*
3. *For all $f \in H_k$ and $x \in \mathcal{X}$, $|f(x)| \leq \|f\|_k \|k\|_\infty$.*

*Proof.* For the first item, assume that $k$ is bounded, then obviously $\|k\|_\infty < \infty$. Conversely, if $\|k\|_\infty < \infty$, then we have for all $x, x' \in \mathcal{X}$

$$|k(x, x')| = |\langle k(\cdot, x'), k(\cdot, x)\rangle_k| \leq \|k(\cdot, x')\|_k \|k(\cdot, x)\|_k = \sqrt{k(x', x')}\sqrt{k(x, x)} \leq \|k\|_\infty^2 < \infty$$

so $k$ is indeed bounded.

The second statement is given by Lemma 4.23 in (Steinwart & Christmann, 2008).

For the last statement, let $f \in H_k$ and $x \in \mathcal{X}$ be arbitrary, then

$$|f(x)| = |\langle f, k(\cdot, x)\rangle_k| \leq \|f\| \|k(\cdot, x)\|_k = \|f\|_k \sqrt{k(x, x} \leq \|f\|_k \|k\|_\infty.$$

□

**Properties of loss functions and their risks** Next, we present two technical results on loss functions and their associated risks. These results are essentially known (cf. Chapter 2 in (Steinwart & Christmann, 2008)), however, we formulate them in greater generality using comparison functions.

**Lemma A.2** (Condition for $P$-integrable Nemitskii loss)**.** *Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function such that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ with $\ell(x, y, 0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and a nondecreasing family $(\alpha_{\ell,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ with $|\ell(x, y, t_1) - \ell(x, y, t_2)| \leq \alpha_{\ell,B}(|t_1 - t_2|)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$, then $\ell$ is a $P$-integrable Nemitskii loss for all distributions $P$ on $\mathcal{X} \times \mathcal{Y}$.*

In particular, this result applies to locally Lipschitz continuous functions, where $\alpha_{\ell,B}(t) = |\ell|_{1,|t|}|t|$.

*Proof.* Let $x \in \mathcal{X}, y \in \mathcal{Y}, t \in \mathbb{R}$ be arbitrary, then we have

$$\begin{aligned}\ell(x,y,t) &\leq \ell(x,y,0) + |\ell(x,y,t) - \ell(x,y,0)| \\ &\leq B_\ell + \alpha_{\ell,|t|}(|t|)\end{aligned}$$

Since $\int B_\ell \mathrm{d}P = B_\ell < \infty$ and $t \mapsto \alpha_{\ell,|t|}(|t|)$ is nondecreasing, the statement follows. □

**Lemma A.3** (Continuity of risk from continuity of loss function)**.** *Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function such that there exists a nondecreasing family $(\alpha_{\ell,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ with $|\ell(x,y,t_1) - \ell(x,y,t_2)| \leq \alpha_{\ell,B}(|t_1 - t_2|)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$. Let $P$ be a distribution such that $\ell$ is a $P$-integrable Nemitskii loss.*

1. *For all $B \in \mathbb{R}_{>0}$ and all measurable and bounded[7] $f, g$ with $\|f\|_\infty, \|g\|_\infty \leq B$, we have*
$$|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| \leq \alpha_{\ell,B}(\|f - g\|_\infty). \tag{11}$$
2. *Let $k$ be a measurable and bounded kernel on $\mathcal{X}$. For all $B \in \mathbb{R}_{>0}$ and $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq B$, we have*
$$|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| \leq \alpha_{\ell,\|k\|_\infty \cdot B}(\|f - g\|_k \|k\|_\infty). \tag{12}$$

*Proof.* For the first claim, let $B \in \mathbb{R}_{>0}$ and $f, g$ be measurable functions with $\|f\|_\infty, \|g\|_\infty \leq B$. We then have

$$\begin{aligned}|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| &\leq \int |\ell(x,y,f(x)) - \ell(x,y,g(x))| \mathrm{d}P(x,y) \\ &\leq \int \alpha_{\ell,B}(|f(x) - g(x)|) \mathrm{d}P(x,y) \\ &\leq \int \alpha_{\ell,B}(\|f - g\|_\infty) \mathrm{d}P(x,y) \\ &= \alpha_{\ell,B}(\|f - g\|_\infty),\end{aligned}$$

where we used the triangle inequality in the first step, the existence of $(\alpha_{\ell,B})_B$ in the second step, the fact that $\alpha_{\ell,B}$ is increasing in the third step, and finally that $P$ is a probability distribution.

For the second claim, let $B \in \mathbb{R}_{>0}$ and $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq B$. Since $k$ is measurable and bounded, also $f, g$ are measurable and bounded. From Lemma A.1 we get $\|f\|_\infty \leq \|f\|_k \|k\|_\infty \leq B\|k\|_\infty$, and similarly for $g$, as well as $\|f - g\|_\infty \leq \|f - g\|_k \|k\|_\infty$. The result now follows from the first claim. □

**Bound on norm of regularized risks minimizers** Finally, we recall a well-known result providing a bound on the norm of minimizers of regularized risks minimization problems, cf. the beginning of Section 5.1 in (Steinwart & Christmann, 2008).

**Lemma A.4** (Regularized risk minimization over RKHSs)**.** *Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex, locally Lipschitz continuous loss function, such that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ with $\ell(x,y,0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k$ be a kernel on $\mathcal{X}$ that is measurable, bounded, and with separable $H_k$. For all distributions $P$ on $\mathcal{X} \times \mathcal{Y}$ and all $\lambda \in \mathbb{R}_{>0}$, there exists a unique solution $f_{P,\lambda}$ of*

$$\min_{f \in H_k} \mathcal{R}_{\ell,P}(f) + \lambda \|f\|_k^2,$$

*and $\|f_{P,\lambda}\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$.*

*Proof.* Lemma A.2 ensures that $\ell$ is a $P$-integrable Nemitskii loss, so Lemma 5.1 and Theorem 5.2 from (Steinwart & Christmann, 2008) are applicable and ensure that a unique solution $f_{P,\lambda}$ exists.

Additionally, we have

$$\begin{aligned}\lambda \|f_{P,\lambda}\|_k^2 &\leq \mathcal{R}_{\ell,P}(f_{P,\lambda}) + \lambda \|f_{P,\lambda}\|_k^2 = \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) \\ &\leq \mathcal{R}_{\ell,P,\lambda}(0) = \mathcal{R}_{\ell,P}(0) + \lambda \|0\|_k^2 \\ &= \int \ell(x,y,0) \mathrm{d}P(x,y) \\ &\leq B_\ell,\end{aligned}$$

[7] Measurably essentially bounded would be enough.

where we first used the nonnegativity of $\ell$ (and monotonicity of the integral) in the first step, followed by the definition of $f_{P,\lambda}$, and finally the boundedness assumption of $\ell$ in zero. Rearranging shows that indeed $\|f_{P,\lambda}\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$. □

**Kernel mean embeddings**

*Proof of Proposition 2.4.* The first statement is contained in Theorem 2 and Proposition 2 in (Sriperumbudur et al., 2010), and the discussion following it.

The second statement follows from Theorem 1 in (Lopez-Paz et al., 2015), by two minor modifications. First, applying Lemma A.1 to $f \in H_\kappa$ with $\|f\|_\kappa \leq 1$ leads to $|f(s)| \leq \|f\|_\kappa \|\kappa\|_\infty \leq \|\kappa\|_\infty$ for all $s \in \mathcal{S}$, which shows that the constant of the bounded difference property in the proof of Theorem 1 in (Lopez-Paz et al., 2015) needs to be set to $2\|\kappa\|_\infty/M$. Second, we use $\int \kappa(s,s)\mathrm{d}Q(s) \leq \int \|\kappa\|_\infty^2 \mathrm{d}Q(s) \leq \|\kappa\|_\infty^2$ for all $Q \in \mathcal{M}_1(\mathcal{S})$.

The third statement is shown in Section A.1.1 in (Szabó et al., 2015). □

**Sliced Wasserstein distances** Let $\mathcal{S} = \mathbb{R}^d$ and denote by $\underline{\mathcal{W}}_2(\mu,\nu)$ the sliced 2-Wasserstein distance, cf. equation (13) in (Meunier et al., 2022). It has been shown in Proposition 5 in the same reference that there exists a Hilbert space $\mathcal{H}_2$ and a map $\Pi_2 : \mathcal{M}_1(\mathcal{S}) \to \mathcal{H}_2$ such that $\underline{\mathcal{W}}_2(\mu,\nu) = \|\Pi_2\mu - \Pi_2\nu\|_{\mathcal{H}_2}$. Setting $\hat{\Pi}_2 S = \Pi_2 \hat{\mu}[S]$ for all $S \in \mathcal{S}^M$ and $M \in \mathbb{N}_+$, where $\hat{\mu}[S] = \frac{1}{M}\sum_{m=1}^M \delta_{S_m}$ is the empirical measure having the components of $S$ as atoms, and assuming that Assumption 2.1 is fulfilled, our main results Theorems 3.1, 3.4 and 4.2 apply to the case of sliced 2-Wasserstein-based Hilbertian embeddings. For more details, as well as the case of sliced 1-Wasserstein-based Hilbertian embeddings, and concrete constructions of suitable kernels $k$ on $\mathcal{H}_2$, we refer to (Meunier et al., 2022).

## B. Oracle Inequalities

In this section, we present the proofs of our oracle inequalities Theorem 3.1 and Theorem 3.4. Furthermore, we state and prove specializations to the case of sliced 2-Wasserstein embeddings, analogous to the results for KMEs, cf. Corollary 3.3 and Corollary 3.6.

### B.1. Sliced Wasserstein Distances

Our specialization of the oracle inequalities to sliced 2-Wasserstein embeddings are based on the following error bound, which might be of independent interest.

**Proposition B.1.** *Let $P$ be a distribution on $\mathcal{M}_1(\mathbb{R}^d) \times \mathcal{Y}$ and $(Q,y) \sim P$. Assume that $P$-a.s. $Q$ is a log-concave distribution, and denote its ($P$-a.s. defined) covariance matrix by $\Sigma_Q$. Furthermore, assume that there exists $\rho_\Sigma \in \mathcal{L}$ such that for all $t \in \mathbb{R}_{\geq 0}$, $\mathbb{P}[\|\Sigma_Q\| \geq t] \leq \rho_\Sigma(t)$ $P$-a.s. Let $M \in \mathbb{N}_+$ and $S \sim Q^{\otimes M}$, then for all $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$, we have*

$$\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \frac{\rho_\Sigma^{-1}(\delta/2)}{\sqrt{M}}\left(C_d\sqrt{\ln(M)} + \tilde{C}_d \ln(4/\delta)\right)\right] \leq \delta,$$

*where $C_d$ and $\tilde{C}_d$ are universal constants that only depend on $d$.*

To simplify the notation in the following proof, we define $a \wedge b = \min\{a,b\}$ for $a,b \in \mathbb{R}$.

*Proof.* As shown in the proof of Proposition 7 in (Nietert et al., 2022), there exists a universal constant $c_d \in \mathbb{R}_{>0}$, depending only on $d \in \mathbb{N}_+$, such that $\frac{1}{P_\mu} \geq \frac{1}{c_d\|\Sigma_\mu\|}$ for all log-concave distributions $\mu$ on $\mathbb{R}^d$, where $P_\mu$ is the Poincare constant of $\mu$. Furthermore, according to Theorem 1 (choosing $p = 2$ there) in the same reference, there exists a universal constant $C_d \in \mathbb{R}_{>0}$, depending only on $d$, such that

$$\mathbb{E}[\underline{\mathcal{W}}_2(\mu,\hat{\mu}[X])] \leq \sqrt{\frac{\|\Sigma_\mu\|\ln(M)}{M}} \tag{13}$$

for all log-concave distributions $\mu$ on $\mathbb{R}^d$ and $X \sim \mu^{\otimes M}$.

Let $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$ be arbitrary, and $x, t \in \mathbb{R}_{>0}$ two constants to be chosen later. We start with

$$
\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t,\ \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad + \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t,\ \|\Sigma_Q\| > x \wedge \sqrt{x}\right] \\
&\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \mid \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \mathbb{P}[\|\Sigma_Q\| \leq x \wedge \sqrt{x}] \\
&\qquad + \mathbb{P}[\|\Sigma_Q\| > x \wedge \sqrt{x}] \\
&\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \mid \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] + \rho_\Sigma(x \wedge \sqrt{x}),
\end{aligned}
$$

where we used in the last step that probabilities are always from $[0, 1]$, and the assumption on $\|\Sigma_Q\|$.

We continue with the first term,

$$
\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq C_d \sqrt{\frac{x \ln(M)}{M}} + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq C_d \sqrt{\frac{\|\Sigma_Q\| \ln(M)}{M}} + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq \mathbb{E}[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S])] + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{P}\left[|\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) - \mathbb{E}[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S])]| \geq t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{E}\left[2 \exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\min\{2\sqrt{P_Q}, 6e^5 P_Q\}}\right) \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right],
\end{aligned}
$$

where we used Theorem 3.8 from (Lin et al., 2021), as used in the proof of Proposition 7 in (Nietert et al., 2022), and the last equality holds almost surely.

Conditional on $\|\Sigma_Q\| \leq x \wedge \sqrt{x}$, we get that

$$\begin{aligned}
\frac{1}{\min\{2\sqrt{P_Q}, 6e^5 P_Q\}} &= \max\left\{\frac{1}{2\sqrt{P_Q}}, \frac{1}{6e^5 P_Q}\right\} \\
&\geq \frac{1}{6e^5}\max\left\{\frac{1}{\sqrt{P_Q}}, \frac{1}{P_Q}\right\} \\
&\geq \frac{1}{6e^5}\max\left\{\frac{1}{\sqrt{c_d\|\Sigma_Q\|}}, \frac{1}{c_d\|\Sigma_Q\|}\right\} \\
&\geq \frac{1}{6e^5\max\{\sqrt{c_d}, c_d\}}\max\left\{\frac{1}{\sqrt{\|\Sigma_Q\|}}, \frac{1}{\|\Sigma_Q\|}\right\} \\
&= \frac{1}{6e^5\max\{\sqrt{c_d}, c_d\}}\frac{1}{\min\{\sqrt{\|\Sigma_Q\|}, \|\Sigma_Q\|\}} \\
&\leq \frac{1}{6e^5\max\{\sqrt{c_d}, c_d\}}\frac{1}{x \wedge \sqrt{x}} \\
&= \frac{1}{\tilde{C}_d(x \wedge \sqrt{x})}.
\end{aligned}$$

In the last inequality we used that

$$\min\{\sqrt{\|\Sigma_Q\|}, \|\Sigma_Q\|\} \leq \min\{\sqrt{x \wedge \sqrt{x}}, x \wedge \sqrt{x}\} \leq x \wedge \sqrt{x},$$

and in the last step we defined $\tilde{C}_d = 6e^5 \max\{\sqrt{c_d}, c_d\}$.

We therefore get (again almost surely) that

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x})C_d\sqrt{\frac{\ln(M)}{M}} + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad\leq \mathbb{E}\left[2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\min\{2\sqrt{P_Q}, 6e^5 P_Q\}}\right) \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad\leq \mathbb{E}\left[2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad= 2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right).
\end{aligned}$$

Observe now that

$$2\exp\left(-\frac{\sqrt{M}t^2 \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) = \frac{\delta}{2} \quad \Leftrightarrow \quad x \wedge \sqrt{x} = \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}$$

and since $\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} > 0$ (recall that we restricted $\delta$ to $(0, 1/4)$), we can choose $x \in \mathbb{R}_{>0}$ such that the last display holds.

With this choice of $x$, we are now at

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\quad = \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\quad \leq 2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) + \rho_\Sigma\left(x \wedge \sqrt{x}\right) \\
&\quad \leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}\right).
\end{aligned}$$

Note that this holds since the above computation works for any version of the conditional expectation.

Next, let $s > 1$ and set $t = \frac{\ln(4/\delta)}{\sqrt{M}} s$, then

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq s\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}} s\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \frac{\ln(4/\delta)s \wedge \ln(4/\delta)^2 s^2}{\tilde{C}_d \ln(4/\delta)} C_d \sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}} s\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}\right) \\
&= \frac{\delta}{2} + \rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right),
\end{aligned}$$

where we used that $\ln(4/\delta)s \wedge \ln(4/\delta)^2 s^2 = \ln(4/\delta)s$ since $\ln(4/\delta), s > 1$.

The condition $\mathbb{P}[\|\Sigma_Q\| \geq x] \leq \rho_\Sigma(x)$ for all $x \in \mathbb{R}_{\geq 0}$ implies that $\rho_\Sigma([0,\infty)) = (0,1]$, so we have

$$\rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right) = \frac{\delta}{2} \quad \Leftrightarrow \quad s = \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)$$

and since

$$s > 1 \quad \Leftrightarrow \quad \tilde{C}_d \rho_\Sigma^{-1}(\delta/2) > 1 \quad \Leftrightarrow \quad \delta < 2\rho_\Sigma(1/\tilde{C}_d),$$

our requirements on $\delta$ ensures that we can set $s = \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)$.

Altogether, we arrived at

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}} \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq s\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}} s\right] \\
&\leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right) \\
&= \delta
\end{aligned}$$

□

We can now formulate and prove the announced specializations of the oracle inequalities.

**Corollary B.2.** *Consider the situation of Theorem 3.1. Let $\mathcal{S} = \mathbb{R}^d$, set $\mathcal{H} = \mathcal{H}_2$, $\Pi = \Pi_2$, and $\hat{\Pi} = \hat{\Pi}_2$, and assume that Assumption 2.1 holds in this case. Furthermore, for $(Q, y) \sim P$, assume that $P$-a.s. $Q$ is a log-concave distribution, and denote its ($P$-a.s. defined) covariance matrix by $\Sigma_Q$. Assume that $\ell$ is convex, differentiable, $\ell'$ is locally Lipschitz continuous, and that there exists $B_\ell, B'_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x, y, 0) \leq B_\ell$ and $|\ell'(x, y, 0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k$ be a kernel on $\mathcal{H}$ that is measurable and bounded, and that there exists $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|)$. We then have for all $\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0, 1)$ that with probability at least $1 - \delta$*

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\vec{\pi}},\lambda}) - \mathcal{R}^*_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) \\
&+ \frac{2\sqrt{\lambda B_\ell} + |\ell|_{1,B_f}\|k\|_\infty}{N} \sum_{n=1}^{N} \alpha_\lambda \left( \frac{\rho_\Sigma^{-1}\left(\frac{\delta}{2(N+|\mathcal{F}|)}\right)}{\sqrt{M}} \left( C_d\sqrt{\ln(M)} + \tilde{C}_d \ln\left(\frac{4(N+|\mathcal{F}|)}{\delta}\right)\right)\right) \\
&+ 2|\ell|_{1,B_f}\|k\|_\infty \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \left( \sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N}\right),
\end{aligned}$$

*with $B_f$ and $\alpha_\lambda$ as in Theorem 3.1, and $C_d$ and $\tilde{C}_d$ are universal constants that only depend on $d$.*

*Proof.* The result follows immediately by combining Theorem 3.1 with Proposition B.1. □

**Corollary B.3.** *Consider the situation of Theorem 3.1. Let $\mathcal{S} = \mathbb{R}^d$, set $\mathcal{H} = \mathcal{H}_2$, $\Pi = \Pi_2$, and $\hat{\Pi} = \hat{\Pi}_2$, and assume that Assumption 2.1 holds in this case. Furthermore, for $(Q, y) \sim P$, assume that $P$-a.s. $Q$ is a log-concave distribution, and denote its ($P$-a.s. defined) covariance matrix by $\Sigma_Q$. Finally, assume that there exists $\rho_\Sigma \in \mathcal{L}$ such that for all $t \in \mathbb{R}_{\geq 0}$, $\mathbb{P}[\|\Sigma_Q\| \geq t] \leq \rho_\Sigma(t)$ $P$-a.s. We then have for all $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$ that with probability at least $1 - \delta$ it holds that*

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) + 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f)\right)\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + 4\gamma_{3,\tilde{B}_f}(\epsilon) \\
&+ \frac{2}{N}\sum_{n=1}^{N} \alpha_\lambda \left( \frac{\rho_\Sigma^{-1}\left(\frac{\delta}{2(N+|\mathcal{F}|)}\right)}{\sqrt{M}} \left( C_d\sqrt{\ln(M)} + \tilde{C}_d \ln\left(\frac{4(N+|\mathcal{F}|)}{\delta}\right)\right)\right),
\end{aligned}$$

*where we defined $\tilde{B}_f = \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} + \epsilon$, $\alpha_\lambda = \gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}$, and $C_d$ and $\tilde{C}_d$ are universal constants that only depend on $d$.*

*Proof.* The result follows immediately by combining Theorem 3.4 with Proposition B.1. □

### B.2. Proof of the Oracle Inequalities

We will need the following result, which is derived at the beginning of Section 5.2 in (Steinwart & Christmann, 2008), but not stated as a theorem there. For convenience, we repeat it here.

**Proposition B.4.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ a loss function that is convex, differentiable, and define $\ell' = \frac{\mathrm{d}}{\mathrm{d}t}\ell$. Let $k$ be a kernel on $\mathcal{X}$ that is measurable, bounded, and has a separable RKHS $H_k$. For all $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ such that $\ell$ and $|\ell'|$ are $P$-integrable Nemitskii losses, and for all $\lambda \in \mathbb{R}_{>0}$, there exists a unique solution $f_{P,\lambda}$ of*

$$\min_{f \in H_k} \mathcal{R}_{\ell,P}(f) + \lambda\|f\|_k^2, \tag{14}$$

*and this solution fulfills the equation*

$$f_{P,\lambda} = -\frac{1}{2\lambda}\int_{\mathcal{X}\times\mathcal{Y}} \ell'(x, y, f_{P,\lambda}(x))\Phi_k(x)\mathrm{d}P(x, y). \tag{15}$$

Note that in (15) a Bochner integral appears.

*Proof of Theorem 3.1.* Let $\lambda \in \mathbb{R}_{>0}$ be arbitrary and define $\bar{\mathcal{D}} = ((Q_n, y_n))_{n\in[N]}$. We then have

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P,\lambda} &= \mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) \\
&= \mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 + \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) \\
&\quad + \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) + \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) \\
&\quad - \mathcal{R}_{\ell,P}(f_{P,\lambda}) - \lambda\|f_{P,\lambda}\|_k^2 + \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 \\
&= \underbrace{\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})}_{I} + \underbrace{\mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,P}(f_{P,\lambda})}_{II} \\
&\quad + \underbrace{\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) + \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 - (\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) + \lambda\|f_{P,\lambda}\|_k^2)}_{=III} \\
&\quad + \underbrace{\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda})}_{=IV} + \underbrace{\lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2}_{=V}
\end{aligned}$$

We now upper bound terms I to V. First, by definition of $f_{\bar{\mathcal{D}}_\Pi,\lambda}$, term III is nonpositive, and hence can be discarded.

In order to bound the remaining terms, we need some preparations. Lemma A.2 ensures that for all distributions $Q$ on $\mathcal{X}\times\mathcal{Y}$, $\ell$ is a $Q$-integrable Nemitskii loss. Furthermore, repeating the proof of Lemma A.2 on $\ell'$ shows that also $|\ell'|$ is a $Q$-integrable Nemitskii loss. Altogether, we can apply Proposition B.4 to $\ell$ for any distribution $Q$ on $\mathcal{X}\times\mathcal{Y}$. An inspection of the proof of Theorem 5.9 in (Steinwart & Christmann, 2008) reveals that (5.14) in this reference applies to the present situation, so for all distributions $Q, \tilde{Q}$ on $\mathcal{X}\times\mathcal{Y}$, unique SVM solutions $f_{Q,\lambda}$ and $f_{\tilde{Q},\lambda}$ exist, and we have

$$\|f_{Q,\lambda} - f_{\tilde{Q},\lambda}\|_k \le \frac{1}{\lambda}\left\|\int h_Q(x,y)\Phi_k(x)\mathrm{d}Q(x,y) - \int h_Q(x,y)\Phi_k(x)\mathrm{d}\tilde{Q}(x,y)\right\|_k, \tag{16}$$

where we defined $h_Q(x,y) = \ell'(x,y,f_{Q,\lambda}(x))$.

Bounding I Using Lemma A.4, we have $\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k, \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \le \sqrt{\frac{B_\ell}{\lambda}}$, hence we get from Lemma A.1 that $|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(x)|, |f_{\bar{\mathcal{D}}_\Pi,\lambda}(x)| \le \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} =: B_f$. Define now for brevity $L_\ell := |\ell|_{1,B_f}$, then we get

$$\begin{aligned}
\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| &\le L_\ell\|k\|_\infty\|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \\
&\le \frac{L_\ell\|k\|_\infty}{\lambda}\left\|\frac{1}{N}\sum_{n=1}^N h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n)\Phi_k(\hat{\Pi}S^{(n)}) - \frac{1}{N}\sum_{n=1}^N h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)\right\|_k \\
&\le \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N\left\|h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n)\Phi_k(\hat{\Pi}S^{(n)}) - h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)\right\|_k \\
&\le \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N |h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n) - h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)|\|\Phi_k(\hat{\Pi}S^{(n)})\|_k \\
&\quad + |h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)|\|\Phi_k(\hat{\Pi}S^{(n)}) - \Phi_k(\Pi Q_n)\|_k
\end{aligned}$$

where we used Lemma A.3 in the first inequality, in the second step the bound (16), followed by using the triangle inequality twice. For each $n = 1,\ldots,N$, we have

$$\begin{aligned}
|h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n) - h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| &= |\ell'(\hat{\Pi}S^{(n)}, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\hat{\Pi}S^{(n)})) - \ell'(\Pi Q_n, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n))| \\
&\le \gamma_1(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}) + \gamma_{3,B_f}(|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\hat{\Pi}S^{(n)}) - f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)|) \\
&\le \left(\gamma_1 + \gamma_{3,B_f}\circ\left(\sqrt{B_\ell/\lambda}\cdot\alpha_k\right)\right)(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}),
\end{aligned}$$

where we used the definition of $h_{\mathcal{D}_{\hat{\Pi}}}$ in the first step, and in the following inequality we used the assumed continuity property of $\ell'$ (together with the previously derived bound $B_f$ on the values of $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$ and $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$). In the last inequality we used that for all $f \in H_k$ and $x_1, x_2 \in \mathcal{X}$,

$$|f(x_1) - f(x_2)| = |\langle f, \Phi_k(x_1) - \Phi_k(x_2)\rangle_k| \le \|f\|_k\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \le \|f\|_k\alpha_k(\|x_1 - x_2\|_{\mathcal{H}}).$$

Furthermore, we also have $\|\Phi_k(\hat{\Pi}S^{(n)}) - \Phi_k(\Pi Q_n)\|_k \leq \alpha_k(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}})$ and $\|\Phi_k(\hat{\Pi}S^{(n)})\|_k \leq \|k\|_\infty$.

Finally,

$$\begin{aligned}
|h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| &= |\ell'(\Pi Q_n, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n))| \\
&\leq |\ell'(\Pi Q_n, y_n, 0)| + |\ell'(\Pi Q_n, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)) - \ell'(\Pi Q_n, y_n, 0)| \\
&\leq B'_\ell + \gamma_{3,B_f}(|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)|) \\
&\leq B'_\ell + \gamma_{3,B_f}(B_f).
\end{aligned}$$

Altogether, we can continue with

$$\begin{aligned}
&\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| \\
&\leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N |h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n) - h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)|\|\Phi_k(\hat{\Pi}S^{(n)})\|_k \\
&\quad + |h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)|\|\Phi_k(\hat{\Pi}S^{(n)}) - \Phi_k(\Pi Q_n)\|_k \\
&\leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N \|k\|_\infty\left(\gamma_1 + \gamma_{3,B_f}\circ\left(\sqrt{B_\ell/\lambda}\cdot\alpha_k\right)\right)(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}) \\
&\quad + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\alpha_k(\|\hat{\Pi}S^{(n)} - \Pi Q_n)\|_{\mathcal{H}}) \\
&\leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N \left(\|k\|_\infty\left(\gamma_1 + \gamma_{3,B_f}\circ\left(\sqrt{B_\ell/\lambda}\cdot\alpha_k\right)\right) + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\alpha_k\right)(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}})
\end{aligned}$$

Defining $\alpha_\lambda = \|k\|_\infty(\gamma_1 + \gamma_{3,B_f}\circ\left(\sqrt{B_\ell/\lambda}\alpha_k\right)) + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\alpha_k$ and using a union bound, we finally get with probability at least $1-\delta/2$ that

$$\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| \leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N \alpha_\lambda(B_n(\delta/(2N))).$$

Bounding II and IV Let $Q = P$ or $\bar{\mathcal{D}}_\Pi$. We have

$$\begin{aligned}
\mathcal{R}_{\ell,Q}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,Q}(f_{P,\lambda}) &\leq L_\ell\|k\|_\infty\|f_{\bar{\mathcal{D}}_\Pi,\lambda} - f_{P,\lambda}\|_k \\
&\leq \frac{L_\ell\|k\|_\infty}{\lambda}\left\|\frac{1}{N}\sum_{n=1}^N h_{\bar{\mathcal{D}}_\Pi}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n) - \int h_{\bar{\mathcal{D}}_\Pi}(x,y)\Phi_k(x)\mathrm{d}P(x,y)\right\|_k \\
&= \frac{L_\ell\|k\|_\infty}{\lambda}\left\|\frac{1}{N}\sum_{n=1}^N \xi_n - \mathbb{E}[\xi_n]\right\|_k
\end{aligned}$$

where the first two steps are similar as in bounding I, and in the last step we defined $\xi_n = h_{\bar{\mathcal{D}}_\Pi}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)$. Since $(Q_1, y_1)\ldots,(Q_N, y_N) \overset{\text{i.i.d.}}{\sim} P$, also $\xi_1,\ldots,\xi_N$ are i.i.d. Furthermore,

$$\begin{aligned}
\|\xi_n\|_k &= \|h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)\|_k \\
&= |h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)|\|\Phi_k(\Pi Q_n)\|_k \\
&\leq |\ell'(\Pi Q_n, y_n, f_{\bar{\mathcal{D}}_\Pi,\lambda}(\Pi Q_n))|\|k\|_\infty \\
&\leq \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\|k\|_\infty,
\end{aligned}$$

so $\xi_1,\ldots,\xi_N$ are $H_k$-valued i.i.d. random variables bounded by $B_\xi := \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\|k\|_\infty$. Hoeffding's inequality for random variables in a separable Hilbert space, cf. Corollary 6.15 in (Steinwart & Christmann, 2008), now ensures that with probability at least $1-\delta/2$

$$\left\|\frac{1}{N}\sum_{n=1}^N \xi_n - \mathbb{E}[\xi_n]\right\|_k \leq B_\xi\left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N}\right).$$

This implies that with probability at least $1-\delta/2$

$$\mathcal{R}_{\ell,Q}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,Q}(f_{P,\lambda}) \leq \frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left( \sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right),$$

so with same probability the bound

$$II + IV \leq 2\frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left( \sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right)$$

holds.

Bounding V Using elementary computations, we get

$$\begin{aligned}
\lambda \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 &= \lambda \left( \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 \right) \\
&= \lambda \left( \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k + \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \right) \left( \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k - \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \right) \\
&\leq \lambda \left( \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k + \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \right) \|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \\
&\leq 2\lambda \sqrt{\frac{B_\ell}{\lambda}} \|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \\
&\leq 2\sqrt{\frac{B_\ell}{\lambda}} \frac{1}{N} \sum_{n=1}^N \alpha_\lambda(\|\hat{\Pi} S^{(n)} - \Pi Q_n\|_{\mathcal{H}}),
\end{aligned}$$

where we used Lemma A.4 in the second to last step, and the bound on $\|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k$ from bounding I. In particular, with probability at least $1-\delta/2$ we get that

$$\lambda \|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 \leq 2\sqrt{\frac{B_\ell}{\lambda}} \frac{1}{N} \sum_{n=1}^N \alpha_\lambda(B_n(\delta/(2N))).$$

Finishing Using again a union bound, we finally get that with probability at least $1-\delta$ we have

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}^{H_k *} &\leq \underbrace{\frac{L_\ell \|k\|_\infty}{\lambda} \frac{1}{N} \sum_{n=1}^N \alpha_\lambda(B_n(\delta/(2N)))}_{\text{from I}} \\
&\quad + \underbrace{2\frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left( \sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right)}_{\text{from II and IV}} \\
&\quad + \underbrace{2\sqrt{\frac{B_\ell}{\lambda}} \frac{1}{N} \sum_{n=1}^N \alpha_\lambda(B_n(\delta/(2N)))}_{\text{from V}} \\
&= \left( 2\sqrt{\frac{B_\ell}{\lambda}} + \frac{L_\ell \|k\|_\infty}{\lambda} \right) \frac{1}{N} \sum_{n=1}^N \alpha_\lambda(B_n(\delta/(2N))) \\
&\quad + 2\frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left( \sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right)
\end{aligned}$$

The result now follows from the definition of $A_{\ell,P}^{(2)}(\lambda)$. □

*Proof of Corollary 3.3.* Since $\mathcal{S}$ is compact, it is in particular separable, so Proposition 2.4 ensures that $\Pi_\kappa$ is $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$-$(H_\kappa, \mathcal{B}(H_\kappa))$-measurable. Furthermore, since $\mathcal{S}$ is a compact metric space, $\mathcal{M}_1(\mathcal{S})$ with the topology of weak convergence is compact. Since $\kappa$ is universal, $\Pi_\kappa$ is continuous, and hence $\mathcal{X} = \Pi_\kappa(\mathcal{M}_1(\mathcal{S}))$ is a compact

metric space. In particular, it is closed, and hence $\mathcal{X} \in \mathcal{B}(H_\kappa)$, and it is also separable. By definition, for all $M \in \mathbb{N}_+$ and $S \in \mathcal{S}^M$, $\hat{\Pi}_M(S) = \hat{\Pi}_\kappa(S) = \frac{1}{M}\sum_{m=1}^M k(\cdot, S_m)$, and hence measurable. Altogether, Assumption 2.1 is fulfilled.

Next, for all $x_1, x_2 \in \mathcal{X}$ we have $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_\mathcal{H})$, which shows that $\Phi_k$ is continuous, so according to Lemma 4.29 in (Steinwart & Christmann, 2008) also $k$ is continuous. Since $\mathcal{X}$ is separable, this shows that also $H_k$ is separable.

Using the KME estimation bound from Proposition 2.4 to find appropriate $B_n$, all assumptions of Theorem 3.1 are fulfilled, and we get

$$\begin{aligned}
&\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P,\lambda} \\
&\quad\leq (2\sqrt{\lambda B_\ell} + L_\ell\|k\|_\infty)\frac{1}{N}\sum_{n=1}^N \alpha_\lambda\left(2\sqrt{\frac{\|k\|_\infty^2}{M_n}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(2N/\delta)}{M_n}}\right) \\
&\qquad + 2L_\ell\|k\|_\infty\left(B'_\ell + L'_\ell\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}\right)\left(\sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N}\right),
\end{aligned}$$

where we defined

$$\alpha_\lambda = \|k\|_\infty L'_\ell \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}} + \left(B'_\ell + L'_\ell\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}\right)\alpha_k.$$

Finally, since $\ell$ is locally Lipschitz continuous, it is in particular continuous, and as shown by Lemma A.2, it is also a $P$-integrable Nemitskii loss. Together with the fact that $\mathcal{X}$ is a compact metric space and $k$ is universal, Corollary 5.29 in (Steinwart & Christmann, 2008) shows that $\mathcal{R}^{H_k*}_{\ell,P,\lambda} = \mathcal{R}^*_{\ell,P}$, and the result follows. □

The strategy of the following proof follows the one for Theorem 6.25 in (Steinwart & Christmann, 2008), however, several adaptions are necessary to deal with the two-stage sampling.

*Proof of Theorem 3.4.* Let $\lambda \in \mathbb{R}_{>0}$ be arbitrary. We start with

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) &= \mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) \\
&\quad + \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \left(\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{P,\lambda}) + \lambda\|f_{P,\lambda}\|_k^2\right) \\
&\quad + \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{P,\lambda}) - \mathcal{R}_{\ell,P}(f_{P,\lambda}) \\
&\leq 2 \sup_{\substack{f\in H_k \\ \|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}}} \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,P}(f)\right|,
\end{aligned}$$

where we used in the last step that $\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}},\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}},\lambda}(f_{P,\lambda})$ by definition of $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$, and we applied Lemma A.4 to $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$ and $f_{P,\lambda}$.

Let $f \in H_k$ with $\|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$, and choose $\tilde{f} \in \mathcal{F}$ such that $\|f - \tilde{f}\|_k \leq \epsilon$. Observe that $|\tilde{f}(x)| \leq |f(x)| + |\tilde{f}(x) - f(x)| \leq \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} + \epsilon = \tilde{B}_f$, where we used the choice of $\tilde{f}$ together with (the proof of) Lemma A.2. We then have

$$\begin{aligned}
\left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,P}(f)\right| &\leq \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f})\right| \\
&\quad + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| + \left|\mathcal{R}_{\ell,P}(\tilde{f}) - \mathcal{R}_{\ell,P}(f)\right| \\
&\leq \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| + 2\gamma_{3,\tilde{B}_f}(\epsilon),
\end{aligned}$$

where we used (a modified variant of) Lemma A.3 in the last step together with $|f(x)|, |\tilde{f}(x)| \leq \tilde{B}_f$.

We now bound the first two terms. First,

$$\begin{aligned}\left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f)-\mathcal{R}_{\ell,\bar{\mathcal{D}}_{\Pi}}(f)\right| &\leq \frac{1}{N}\sum_{n=1}^{N}\left|\ell(\hat{\Pi}S^{(n)},y_n,f(\hat{\Pi}S^{(n)}))-\ell(\Pi Q_n,y_n,f(\Pi Q_n))\right| \\ &\leq \frac{1}{N}\sum_{n=1}^{N}\gamma_1(\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}})+\gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\left(\left|f(\hat{\Pi}S^{(n)})-f(\Pi Q_n)\right|\right) \\ &\leq \frac{1}{N}\sum_{n=1}^{N}\left(\gamma_1+\gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\circ\alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}}),\end{aligned}$$

where we used the triangle inequality, then the continuity property of $\ell$, and then the continuity property of $f$.

Second,

$$\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f})-\mathcal{R}_{\ell,P}(\tilde{f})\right|=\left|\frac{1}{N}\sum_{n=1}^{N}\ell(\Pi Q_n,y_n,\tilde{f}(\Pi Q_n))-\int\ell(\Pi Q,y,\tilde{f}(\Pi Q))\mathrm{d}P(Q,y)\right|,$$

$\ell(\Pi Q_1,y_1,\tilde{f}(\Pi Q_1)),\ldots,\ell(\Pi Q_N,y_N,\tilde{f}(\Pi Q_N))$ are i.i.d. random variables (since $(Q_1,y_1),\ldots,(Q_N,y_N)$ are i.i.d.), and for all $n=1,\ldots,N$ we have $|\ell(\Pi Q_n,y_n,\tilde{f}(\Pi Q_n))|\leq B_\ell+\gamma_{3,\tilde{B}_f}(\tilde{B}_f)=B_\xi$ according to (the proof of) Lemma A.2. All of this means that we can use Hoeffding's inequality to bound this term.

Third, we can combine the previous two bounds. Using the union bound we have

$$\begin{aligned}&\mathbb{P}\left[\max_{n=1,\ldots,N}\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}}>B_n(\delta/(N+|\mathcal{F}|))\text{ or }\max_{\tilde{g}\in\mathcal{F}}\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{g})-\mathcal{R}_{\ell,P}(\tilde{g})\right|>B_\xi\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}\right] \\ &\leq\sum_{n=1}^{N}\mathbb{P}\left[\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}}>B_n(\delta/(N+|\mathcal{F}|))\right]+\sum_{\tilde{g}\in\mathcal{F}}\mathbb{P}\left[\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{g})-\mathcal{R}_{\ell,P}(\tilde{g})\right|>B_\xi\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}\right] \\ &\leq N\frac{\delta}{N+|\mathcal{F}|}+|\mathcal{F}|\frac{\delta}{N+|\mathcal{F}|}=\delta\end{aligned}$$

Together with our previous two bounds this implies that with probability at least $1-\delta$,

$$\begin{aligned}&\left|\mathcal{R}_{\ell,\mathcal{D}}(f)-\mathcal{R}_{\ell,\bar{\mathcal{D}}}(f)\right|+\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f})-\mathcal{R}_{\ell,P}(\tilde{f})\right| \\ &\leq\frac{1}{N}\sum_{n=1}^{N}\left(\gamma_1+\gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\circ\alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}})+\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f})-\mathcal{R}_{\ell,P}(\tilde{f})\right| \\ &\leq\frac{1}{N}\sum_{n=1}^{N}\left(\gamma_1+\gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\circ\alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)B_n(\delta/(N+|\mathcal{F}|))+B_\xi\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}.\end{aligned}$$

This also implies that with probability at least $1-\delta$,

$$\begin{aligned}|\mathcal{R}_{\ell,\mathcal{D}}(f)-\mathcal{R}_{\ell,P}(f)| &\leq\left|\mathcal{R}_{\ell,\mathcal{D}}(f)-\mathcal{R}_{\ell,\bar{\mathcal{D}}}(f)\right|+\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f})-\mathcal{R}_{\ell,P}(\tilde{f})\right|+2|\ell|_{1,\tilde{B}_f}\epsilon \\ &\leq\frac{1}{N}\sum_{n=1}^{N}\left(\gamma_1+\gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\circ\alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(B_n(\delta/(N+|\mathcal{F}|)) \\ &\quad+B_\xi\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}+2|\ell|_{1,\tilde{B}_f}\epsilon,\end{aligned}$$

and since $f \in H_k$ with $\|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$ was arbitrary, this in turn implies that

$$\sup_{\substack{f \in H_k \\ \|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}}} |\mathcal{R}_{\ell,\mathcal{D}}(f) - \mathcal{R}_{\ell,P}(f)| \leq \frac{1}{N}\sum_{n=1}^{N}\left(\gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(B_n(\delta/(N+|\mathcal{F}|)) + B_\xi\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + 2\gamma_{3,\tilde{B}_f}(\epsilon),$$

with probability at least $1-\delta$, and the result follows. □

## C. Generalization via Algorithmic Stability

### C.1. Sliced Wasserstein

**Corollary C.1.** *Consider the situation of Theorem 4.2. Additionally, assume $\mathcal{S} = \mathbb{R}^d$, let $(\mathcal{H}, \Pi)$ be the sliced 2-Wasserstein embedding, and assume that the support of $\pi_{\mathcal{S}}\sharp P$[8] is contained in the set of log-concave distributions, and for $(Q, y) \sim P$, denote by $\Sigma_Q$ the (a.s.) defined covariance matrix of $Q$. We then have for all $\delta \in (0,1)$, with probability at least $1-\delta$, that*

$$\mathcal{R}_{\ell,P}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) + \alpha_\lambda\left(C_d\mathbb{E}\left[\sqrt{\frac{\|\Sigma_Q\|\ln(M)}{M}}\right]\right) + \left(\frac{2|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda} + B_\ell + |\ell|_{1,B_f}B_f\right)\sqrt{\frac{\ln(1/\delta)}{2N}} + \frac{|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda N},$$

*where we defined $B_f = \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}$, and $C_d \in \mathbb{R}_{>0}$ is a universal constant that depends only on $d$.*

*Proof.* Let $Q \in \mathcal{M}_1(\mathcal{S})$ and $M \in \mathbb{N}_+$. According to Theorem 1 in (Nietert et al., 2022), we have

$$\mathbb{E}_{S\sim Q^{\otimes M}}[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S])] \leq C_d\sqrt{\frac{\|\Sigma_Q\|\ln(M)}{M}},$$

where $C_d \in \mathbb{R}_{>0}$ is a universal constant that depends only on $d$. This implies that

$$\alpha_\lambda\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}}\right]\right) = \alpha_\lambda\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S])\right]\right) \leq \alpha_\lambda\left(\mathbb{E}\left[C_d\sqrt{\frac{\|\Sigma_Q\|\ln(M)}{M}}\right]\right),$$

with $\alpha_\lambda$ defined in Theorem 4.2. This result now establishes the claim. □

### C.2. Proof of the general result

Our proof follows the one of Theorem 14.2 in (Mohri et al., 2018), adapted to the present distributional setting.

*Proof of Theorem 4.1.* Define $F : (\mathcal{X}\times\mathcal{Y})^N \to \mathbb{R}$ by

$$F(D) = \mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,D}(\mathcal{L}_D).$$

Let $N \in \mathbb{N}_+$, $D \in (\mathcal{X}\times\mathcal{Y})^N$, $1 \leq i \leq N$ and $(\tilde{x}, \tilde{y}) \in \mathcal{X}\times\mathcal{Y}$ be arbitrary. Define $\tilde{D} \in (\mathcal{X}\times\mathcal{Y})^N$ by

$$\tilde{D}_n = \begin{cases} D_n & \text{if } n \neq i \\ (\tilde{x}, \tilde{y}) & \text{if } n = i \end{cases}$$

[8] $\pi_{\mathcal{S}}$ is the usual coordinate projection onto $\mathcal{S}$.

and for all $1 \leq n \leq N$, define also $(\tilde{x}_n, \tilde{y}_n) = D_n$. We then have

$$\begin{aligned}
|F(D) - F(\tilde{D})| &= \left|\mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,D}(\mathcal{L}_D) - \left(\mathcal{R}_{\ell,P}(\mathcal{L}_{\tilde{D}}) - \mathcal{R}_{\ell,\tilde{D}}(\mathcal{L}_{\tilde{D}})\right)\right| \\
&\leq |\mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,P}(\mathcal{L}_{\tilde{D}})| + \left|\frac{1}{N}\sum_{n=1}^{N} \ell(x_n, y_n, \mathcal{L}_D(x_n)) - \frac{1}{N}\sum_{n=1}^{N} \ell(\tilde{x}_n, \tilde{y}_n, \mathcal{L}_{\tilde{D}}(\tilde{x}_n)\right| \\
&\leq \int |\ell(\Pi Q, y, \mathcal{L}_D(\Pi Q)) - \ell(\Pi Q, y, \mathcal{L}_{\tilde{D}}(\Pi Q))| \mathrm{d}P(Q, y) \\
&\quad + \frac{1}{N}\left|\ell(x_i, y_i, \mathcal{L}_D(x_i)) - \ell(\tilde{x}_i, \tilde{y}_i, \mathcal{L}_{\tilde{D}}(\tilde{x}_i)) + \sum_{\substack{n=1 \\ n\neq i}}^{N} \ell(x_n, y_n, \mathcal{L}_D(x_n)) - \ell(\tilde{x}_n, \tilde{y}_n, \mathcal{L}_{\tilde{D}}(\tilde{x}_n))\right| \\
&\leq \beta_N + \frac{1}{N}|\ell(x_i, y_i, \mathcal{L}_D(x_i)) - \ell(\tilde{x}, \tilde{y}, \mathcal{L}_{\tilde{D}}(\tilde{x}))| + \frac{1}{N}\sum_{\substack{n=1 \\ n\neq i}}^{N} |\ell(x_n, y_n, \mathcal{L}_D(x_n)) - \ell(x_n, y_n, \mathcal{L}_{\tilde{D}}(x_n))| \\
&\leq \beta_N + \frac{B}{N} + \frac{N-1}{N}\beta_N = \left(1 + \frac{N-1}{N}\right)\beta_N + \frac{B}{N} = C.
\end{aligned}$$

McDiarmid's bounded difference inequality then shows that for all $\delta \in (0,1)$, we have with probability at least $1-\delta$ that

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) \leq \mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right] + C\sqrt{\frac{N\ln(1/\delta)}{2}}$$

We now bound upper bound the expectation in the preceding display. We have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right] &= \underbrace{\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right]}_{=I} \\
&\quad + \underbrace{\mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right] - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right]}_{=II}
\end{aligned}$$

and bound the two terms separately. Observe that

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) = \mathbb{E}_{(Q,y)\sim P}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q))\right] = \mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q))\right],$$

so we have

$$\begin{aligned}
I &= \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q)) - \ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[|\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q)) - \ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))|\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\alpha(\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}})\right]\right] \\
&\leq \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}}\right]\right),
\end{aligned}$$

where we used Jensen's inequality together with the concavity of $\alpha$ in the last step.

We turn to term II. Let $(S^{(N+1)}, y_{N+1}) \sim \tilde{P}$ such that $(S^{(1)}, y_1), \ldots, (S^{(N+1)}, y_{N+1})$ are i.i.d., and define $\tilde{\mathcal{D}} =$

$((S^{(2)}, y_2), \ldots, (S^{(N+1)}, y_{N+1})$. Note that $\mathcal{D}$ and $\tilde{\mathcal{D}}$ have the same distribution. We then have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}})\right] = \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\frac{1}{N}\sum_{n=1}^{N}\ell(\hat{\Pi}S^{(n)}, y_n, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\ell(\hat{\Pi}S^{(n)}, y_n, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right] \\
&= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right] \\
&= \mathop{\mathbb{E}}_{\substack{(S^{(1)},y_1)\\ \vdots \\ (S^{(N+1)},y_{N+1})}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right] \\
&\leq \mathop{\mathbb{E}}_{\substack{(S^{(1)},y_1)\\ \vdots \\ (S^{(N+1)},y_{N+1})}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\tilde{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right] \\
&\quad + \mathop{\mathbb{E}}_{\substack{(S^{(1)},y_1)\\ \vdots \\ (S^{(N+1)},y_{N+1})}}\left[|\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)})) - \ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\tilde{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))|\right] \\
&\leq \mathop{\mathbb{E}}_{\substack{(S^{(1)},y_1)\\ \vdots \\ (S^{(N+1)},y_{N+1})}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\tilde{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right] + \beta_N \\
&= \mathop{\mathbb{E}}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right] + \beta_N.
\end{aligned}
$$

Furthermore, observe that

$$
\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right] = \mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right].
$$

We now get

$$
\begin{aligned}
II &= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right] - \frac{1}{N}\sum_{n=1}^{N}\ell(\hat{\Pi}S^{(n)}, y_n, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right] \\
&= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right] - \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}})\right] \\
&\leq \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right] - \mathop{\mathbb{E}}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right] + \beta_N \\
&= \mathbb{E}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S)) - \ell(\hat{\Pi}S, y, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right] + \beta_N \\
&= \beta_N.
\end{aligned}
$$

Altogether we have

$$
\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right] \leq \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}}\right]\right) + \beta_N,
$$

and the result follows. □