ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

IFA Automatic Control Laboratory

Semester Project

Distributional Robustness via Maximum Mean Discrepancy Metric

Alberto Loro
January 10, 2022

Advisors
Liviu Aolaritei
Prof. Dr. Florian Dörfler

Chapter 1

Abstract

This thesis constitutes a preliminary step in the study of distributionally robust optimization (DRO) problems, where the ambiguity set is a ball in the space of probability distributions defined using the maximum mean discrepancy (MMD) metric. We start by providing a self-contained introduction to the theory of reproducing kernel Hilbert spaces (RKHS) and the theory of kernel mean embedding (KME) of probability distributions. Here, we illustrate the fundamental results from the literature which will be essential for the comprehension of the results presented in this thesis. Armed with these results, we proceed to define and analyze the DRO problem. In this direction, our contribution can be summarized as follows. We first provide insights on the impact that the choice of the kernel has on the DRO problem. Then, we study the properties of the two protagonists of the DRO problem: the loss function and the ambiguity set. Finally, we focus on the data-driven scenario and study the regularization effect of the DRO problem. Specifically, we will prove that when the number of data points is large, and the center of the ambiguity set is taken as the empirical distribution, the DRO problem is equivalent to a RKHS norm regularization of the empirical loss.

Contents

Chapter 2

Introduction

Imagine a hypothetical scenario where you want to find the optimal route between two points in a map. Finding the shortest way is by no means a difficult task, but now suppose you have reports of some accidents that occured in the past in an unsafe area. Obviously you would like to avoid an accident and one way to do so is to choose a route that avoids entirely the unsafe area. This is what is called *worst-case robust optimization* (RO), it can be formulated as something like:

$$\min_{\theta} \sup_{\xi \in X} \ell(\theta, \xi) \tag{2.1}$$

where θ is your optimization variable , ξ is a variable representing the uncertainty and ℓ is some kind of loss function. However this approach leads to an overly conservative, and thus suboptimal solution, since the entire unsafe area might not be equally risky. From here the idea of constructing a family of probability distributions over the dangerous area representing the risk of accident. Now the problem is to minimize the expected loss assuming the worst case distribution, within the family defined above. This is the idea of *Distributionally Robust Optimization* (DRO):

$$\min_{\theta} \sup_{\mathbb{Q} \in \mathcal{K}} \int \ell(\theta, \xi) d\mathbb{Q}(\xi) \tag{2.2}$$

where $\mathcal{K}$ is the family of probability distributions considered and it is called the *ambiguity set*. The choice of the ambiguity set is crucial in the DRO framework since it directly affects the achieved robustness.

Recently, DRO ha received a lot of attention in the community since it has been proved to be a powerful tool for improving machine learning models. As a matter of fact, instead of seeking the parameter θ which minimizes the empirical risk $\frac{1}{n}\sum_i \ell(x_i, \theta) = \mathbb{E}_{x \sim \mathbb{P}_n}[\ell(x, \theta)]$ an adversary is allowed to perturb the sample distribution within a set $\mathcal{K}$, centered around the empirical distribution $\mathbb{P}_n$. The solution to the DRO problem is thus a model that performs well regardless of the perturbation and is therefore more robust and more likely to better generalize. Furthermore, assuming we are drawing data from a population distribution $\mathbb{P}$, and assuming the ambiguity set $\mathcal{K}$ is large enough to contain $\mathbb{P}$, then we are implicitly optimizing over $\mathbb{P}$ as well.

To correctly define the ambiguity set as a ball in the probability space, centered around the empirical distribution or some other given distribution, we need to establish some kind of metric between probability measures. In machine learning, the DRO ambiguity set $\mathcal{K}$ has so far always been defined as a f-divergence ball or Wasserstein ball around the empirical distribution $\mathbb{P}_n$. Even though these choices are convenient there are some drawbacks to these two kinds of ambiguity sets. For instance, any f-divergence ambiguity set $\mathcal{K}$ around $\mathbb{P}_n$ contains only distributions that have the same finite support of $\mathbb{P}_n$. This is an important limitation since it means that the "true" distribution is typically not in $\mathcal{K}$ and thus the DRO solution can not generalize well on new, unseen, samples. Wasserstein sets do not present this problem but they are more

computationally expensive and they require some nontrivial assumptions on the loss functions and on the ground metric used. We refer the interested reader to the discussion contained in [21].
In this thesis we will explore an alternative formulation of the DRO problem, where our ambiguity set is still a ball in the probability space, but the distance between probability measures is given by a particular instance of an integral probability metric (IPM), where we recall that, given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ the IPM between them is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_X f d\mathbb{P} - \int_X f d\mathbb{Q} \right|$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on X. Choosing $\mathcal{F}$ to be the unit ball in a reproducing kernel Hilbert space (RKHS) we obtain the Maximum Mean Discrepancy (MMD) pseudometric. MMD leverages kernel mean embeddings and it has been extensively used for two- and one-sample tests [6][8] and generative modeling [4][2]. Its main advantages are efficient estimation, fast convergence properties and an inherent flexibility which depends on the chosen kernel.
Before diving into the details, an extensive and self-contained introduction on the theory of reproducing kernel Hilbert spaces, kernel mean embeddings and maximum mean discrepancy distance will be provided, followed by the main results on the literature that are useful to our problem. Moreover, we will dig deeper into the DRO problem where the ambiguity set is defined with respect to the MMD distance. In particular, we will investigate the structure of the ambiguity set (lemma 5.0.1) and discuss the implications of changing center and radius. Further, we will ask ourselves under which conditions our minimax problem provides a coherent solution (theorem 5.0.5) and try to reformulate it in a more tractable way. Explicitly we will show that when the number of data points is large, and the center of the ambiguity set is taken as the empirical distribution, the DRO problem is equivalent to a RKHS norm regularization of the empirical loss (theorem 5.0.9). Finally we will expose the effects of the choice of the kernel in our framework and discuss some examples.

Chapter 3

Background

3.1 Notation and Basic Definitions

X denotes the input domain, i.e. the metric space where our data lives, which will be assumed to be compact. $\mathcal{P}(X)$, or sometimes just $\mathcal{P}$ denotes the set of all Borel probability measures on X. $\mathcal{M}(X)$ is the space of signed measures on X , $\mathcal{M}_b(X)$ is the space of finite signed measures on X while $\mathcal{M}_b^+(X)$ is the space of finite positive measures on X. $C_b(X)$ is the space of all real-valued continuous bounded functions over X. We call the support of a function $f : X \to \mathbb{R}$ the closure w.r.t. X of the subset of X where $f(x) \neq 0$. $C_c(X)$ is the space of all real-valued continuous functions with compact support over X, while $C_0(X)$ is the space of all real-valued continuous function over X which vanish at infinity, i.e. $C_0(X) = \{f \in C(X) \quad s.t. \quad \lim_{||x||\to\infty} f(x) = 0\}$ or equivalently the set of real-valued continuous functions $f \in C(X) \quad s.t. \quad \forall \varepsilon > 0$ the set $\{x \in x : ||f(x)|| \geq \varepsilon\}$ is compact. Note that in the case of X compact $C_b(X) = C_c(X)$. The *weak topology* on $\mathcal{P}(X)$ is the weakest topology such that the map $\mathbb{P} \mapsto \int_X f d\mathbb{P}$ is continuous for all $f \in C_b(X)$. A sequence of probability measures, $\{\mathbb{P}_n\}_{n\in\mathbb{N}}$ is said to *converge weakly* to $\mathbb{P}$, denoted as $\mathbb{P}_n \rightharpoonup \mathbb{P}$, if and only if $\int_X f d\mathbb{P}_n \to \int_X f d\mathbb{P}$ for all $C_b(X)$ as $n \to \infty$. A metric γ on $\mathcal{P}(X)$ is said to metrize the weak topology if the topology induced by γ coincides with the weak topology.

3.2 Reproducing Kernel Hilbert spaces

A Reproducing Kernel Hilbert Space (RKHS) is first of all a Hilbert space, that is, the most natural extension of the mathematical model for the actual space where everyday life takes place (the Euclidean space $\mathbb{R}^3$). Reproducing kernel Hilbert spaces are a particular instance of Hilbert spaces of functions where we have that if two function f and g are close in norm they are also pointwise close, i.e. $||f - g||_{\mathcal{H}}$ "small" $\implies$ $|f(x) - g(x)|$ "small". This will have many desirable implications as we shall see throughout this thesis.
Any discussion on RKHSs can not start without first understanding what a kernel is. A symmetric function $k : X \times X \to \mathbb{R}$ is called a kernel if $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0$ for any $n \in \mathbb{N}$, $\{x_i\}_{i=1}^n \subset X$ and $\{a_i\}_{i=1}^n \subset \mathbb{R}$. Given a kernel k, there exists a Hilbert space of functions $\mathcal{H}$ and a feature map $\phi : X \to \mathcal{H}$, for which $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$.
We are now ready to formally introduce the definition of reproducing kernel and reproducing kernel Hilbert space:

Definition 3.2.1 (Reproducing Kernel Hilbert Space)**.** A function

$$\begin{aligned} k : X \times X &\to \mathbb{R} \\ (x, y) &\mapsto k(x, y) \end{aligned}$$

is a reproducing kernel of the Hilbert space $\mathcal{H}$ if and only if

$$a) \forall x \in X, \quad k(.,x) \in \mathcal{H}$$
$$b) \forall x \in X, \quad \forall \phi \in \mathcal{H} \quad \langle \phi, k(.,x) \rangle_{\mathcal{H}} = \phi(x)$$

where the last condition is called the reproducing property. A Hilbert space of real-valued functions which possesses a reproducing kernel is called a *Reproducing kernel Hilbert space* (RKHS)

Using the Riesz representation theorem the above definition can be shown to be equivalent to defining $\mathcal{H}$ as a RKHS if for all $x \in X$, the evaluation functional, $e_x : \mathcal{H} \to \mathbb{R}$, $e_x(f) := f(x), f \in \mathcal{H}$ is continuous [Theorem 1 from [1]].
According to the Moore-Aronszajn Theorem the correspondence between a kernel and a Hilbert space is unique, moreover

$$\mathcal{H} = \overline{span\{k(x,.) \ \forall \ x \in X\}}$$

where the closure is taken w.r.t. to pointwise limits. This implies that any function $f \in \mathcal{H}$ can be written as $f(x) = \sum_{i=1}^{\infty} k(x, y_i)$ for all $x \in X$. As a side effect, we have that functions in the RKHS are shown to inherit properties from the corresponding reproducing kernel. Given their importance in later stages we will provide here a characterization for RKHSs of continuous functions.

Theorem 3.2.1 (Theorem 17 from [1])**.** *Let $\mathcal{H}$ be a Hilbert space of functions defined on a metric space (X, d) with reproducing kernel k. Then any element of $\mathcal{H}$ is continuous if and only if k satisfies the following conditions:*

- *a) $\forall y \in X, \quad k(.,y)$ is continuous*
- *b) $\forall x \in X, \quad \exists r > 0$, such that the function*

$$X \to \mathbb{R}^+$$
$$y \mapsto k(y,y)$$

  is bounded on the open ball $B(x,r)$.

Note that the second condition is less restrictive than requiring a bounded kernel on X , i.e. $\exists M > 0 \ s.t. \ \forall x \in X \ ||k(x,.)||_{\mathcal{H}} < M$. This implies that if the metric space X is compact, then a continuous kernel is sufficient to have a RKHS of continuous functions. Of course this does not mean that $\mathcal{H} = C_b(X)$, but under some special conditions it could happen that $\mathcal{H}$ is dense in $C_b(X)$ where X is compact, i.e. for any $f \in C_b(X)$ and $\varepsilon > 0 \ \exists g \in \mathcal{H} \ s.t. \ ||f - g||_{\infty} \leq \varepsilon$. In this case the continuous kernel associated to $\mathcal{H}$ is said to be *universal*. Universality is a useful property when working with RKHSs because it allows us to assume w.l.o.g. that any continuous function over X is in $\mathcal{H}$. Many common kernels such as Gaussian and Laplacian have been proved to be universal.
In the literature some alternative definitions of universality have been introduced to handle the case where X is not compact. For instance, [17] and [3] proposed the notion of c_0-universality, which is based on the concept of a c_0-kernel. A kernel k is said to be a c_0-kernel if it is bounded with $k(\cdot, x) \in \mathcal{C}_0(X), \forall x \in X$, where X is a locally compact Hausdorff (LCH) space . A c_0-kernel on a LCH space, X is said to be c_0-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $\mathcal{C}_0(X)$. Again, this means that any function in $\mathcal{C}_0(X)$ can be approximated arbitrarily well by functions in $\mathcal{H}$. Although this seems like an improvement to the standard definition of universality it does not really help us in our learning framework. As we will see later, we will be interested in assuming our loss function $\ell(\theta, \xi) \in \mathcal{H}$, or approximated by functions in $\mathcal{H}$, but this is not feasible since no loss function vanishes at infinity; on the contrary, loss functions are usually assumed to be convex and non-negative, which implies that $\lim_{\xi \to \infty} \ell(\theta, \xi) \neq 0$ for all $\theta \in \mathbb{R}^d$. That's the main reason we will assume X to be compact, a rather common assumption when working with kernel methods.

3.3 Kernel Mean Embedding and Maximum Mean Discrepancy

Kernels and RKHSs have been popularized by the kernelized support vector machine (SVM) for classification problems [12], but lately they have been extended to work with probability distributions via what is called kernel mean embedding [1][15][19]. The kernel mean embedding allows us to study probabilities through their representers in a Hilbert space. For this we need a "representer theorem" which allows us to move from the, hardly tractable and even nonlinear, set of probability measures to a Hilbert space of functions, where we can leverage some well know algebraic and topological properties and thus compare and manipulate distributions in a more simple and effective way, without the need of any density estimation. The choice of the kernel is crucial and directly influences what properties of the distributions are being kept and what kind of information is retained in the embedding. Understanding how this embedding works in the most general possible setting will be the goal of the next sections. Let's now formally introduce the kernel mean embedding of probability measures.

Definition 3.3.1. The kernel mean embedding of probability measures in $\mathcal{P}(X)$ into a RKHS $\mathcal{H}$ endowed with a reproducing kernel $k : X \times X \to \mathbb{R}$ is defined by a mapping

$$\mu : \mathcal{P}(X) \to \mathcal{H}, \quad \mathbb{P} \mapsto \int k(x, .)d\mathbb{P}(x)$$

We still have to clarify under which conditions the kernel mean embedding exists and belongs to $\mathcal{H}$; this lemma from [15] helps us:

Lemma 3.3.1. *If* $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X, X)}] < \infty$, *then* $\mu_{\mathbb{P}} \in \mathcal{H}$ *and* $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$

We include the proof and invite the reader to carefully look at it, since it is quite explanatory of the way we usually reason when working on RKHSs, moreover many intermediate results will be used later on in other derivations.

Proof. Let $T_{\mathbb{P}} : \mathcal{H} \to \mathbb{R}$ be the linear functional defined as $T_{\mathbb{P}}[f] := \int_X f(x)d\mathbb{P}(x)$. In other words, it is the expectation of $f(x)$ over $\mathbb{P}$. Its norm is defined by:

$$\|T_{\mathbb{P}}\| := \sup_{f\in\mathcal{H}, f\neq 0} \frac{|T_{\mathbb{P}}[f]|}{\|f\|_{\mathcal{H}}}$$

Now consider:

$$|T_{\mathbb{P}}[f]| = \left|\int_X f(x)d\mathbb{P}(x)\right| \leq \int_X |f(x)|d\mathbb{P}(x) = \int_X |\langle f, k(x, .)\rangle_{\mathcal{H}}|d\mathbb{P}(x) \leq \int_X \|f\|_{\mathcal{H}}\|k(x, .)\|_{\mathcal{H}}d\mathbb{P}(x) =$$
$$= \|f\|_{\mathcal{H}} \int_X \sqrt{\langle k(x, .), k(x, .)\rangle_{\mathcal{H}}}d\mathbb{P}(x) = \|f\|_{\mathcal{H}}\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X, X)}]$$

where the first inequality is Jensen's inequality and the second inequality is Cauchy-Schwarz inequality.
This implies that:

$$\frac{|T_{\mathbb{P}}[f]|}{\|f\|_{\mathcal{H}}} \leq \mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X, X)}] < \infty$$

Since this is true for any $f \in \mathcal{H}, f \neq 0$ we conclude that $\|T_{\mathbb{P}}\| < \infty$, i.e. $T_{\mathbb{P}}$ is a bounded linear functional. We can thus apply Riesz' theorem [5.2 from [11]]:

$$\exists g \in \mathcal{H} \; s.t. \; \langle g, f\rangle_{\mathcal{H}} = T_{\mathbb{P}}[f], \quad \forall f \in \mathcal{H} \text{ and } ||T_{\mathbb{P}}|| = ||g||_{\mathcal{H}}$$

Choose $f = k(x, .)$ for some $x \in X$, then

$$T_{\mathbb{P}}[k(x, .)] = \int_X k(x, y) d\mathbb{P}(y) = \langle g, k(x, .)\rangle_{\mathcal{H}} = g(x)$$

where the last equality is the reproducing property. In the end we have that:

$$g = \int_X k(., y) d\mathbb{P}(y) = \mu_{\mathbb{P}} \qquad \square$$

Note that $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$ can be viewed as the reproducing property of the expectation operation in the RKHS. In other words, the expectation of a function f in the RKHS w.r.t. the distribution $\mathbb{P}$ can be easily computed as an inner product between the function f and the embedding $\mu_{\mathbb{P}}$.
The kernel mean embedding can be used to define a metric for probability distributions and ultimately, this is the reason we were employing it in the first place. As already mentioned the metric defined in terms of embeddings can be seen as a particular instance of an integral probability metric (IPM) when the function class $\mathcal{F}$ is the unit ball in a RKHS $\mathcal{H}$, i.e. $\mathcal{F} := \{f \in \mathcal{H}\ s.t.\ ||f||_{\mathcal{H}} \leq 1\}$.
Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space X, the *maximum mean discrepancy* (MMD) distance, denoted by $\gamma_k(\mathbb{P}, \mathbb{Q})$, where k is the reproducing kernel, can be expressed as the distance in $\mathcal{H}$ between mean embeddings, provided their existence whose sufficient conditions were explained in lemma 3.3.1. That is,

$$\begin{aligned}
\gamma_k(\mathbb{P}, \mathbb{Q}) &= \sup_{||f||_{\mathcal{H}}\leq 1} \left| \int_X f(x) d\mathbb{P}(x) - \int_X f(x) d\mathbb{Q}(x) \right| && (3.1)\\
&= \sup_{||f||_{\mathcal{H}}\leq 1} |\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - E_{X\sim\mathbb{Q}}[f(X)]| && (3.2)\\
&= \sup_{||f||_{\mathcal{H}}\leq 1} |\langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}} - \langle f, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}| && (3.3)\\
&= \sup_{||f||_{\mathcal{H}}\leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} && (3.4)\\
&= ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}} && (3.5)
\end{aligned}$$

where we applied the definition of expected value in first equality, lemma 3.3.1 in the second equality, the linearity of the inner product in the third equality and the fact that in a Hilbert space $\sup_{||v||=1}\langle u, v\rangle = \|u\|$.
However this approach carries a practical problem, as in statistical learning applications it is impossible to check whether the conditions of lemma 3.3.1 are satisfied since the probability distribution $\mathbb{P}$ is unknown. A natural solution would be then to choose a kernel such that:

$$\int_X \sqrt{k(x, x)} d\mathbb{P}(x) < \infty,\ \forall \mathbb{P} \in \mathcal{P}(X) \qquad (3.6)$$

The following theorem proves that (3.6) is equivalent to have a measurable and bounded kernel.

Theorem 3.3.2 (Proposition 2 from [19])**.** *Let f be a measurable function on X. Then $\int_X f(x) d\mathbb{P}(x) < \infty$ for all $\mathbb{P} \in \mathcal{P}(X)$ if and only if f is bounded.*

Thus we can conclude that choosing a bounded kernel is a sufficient condition to correctly define a pseudometric γ_k between two probability distributions. Note that this is equivalent to require a continuous kernel if our input space X is compact.
Without any extra requirement, MMD is a pseudometric because it does not satisfy $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$. In order to have MMD as a metric we need to make sure that the embedding

$\mathbb{P} \mapsto \int_M k(x,.)d\mathbb{P}$ is injective. When this happens we say the kernel k is *characteristic* to the set of probability measures $\mathcal{P}(X)$. In other words if the kernel is characteristic there is no loss of information when mapping the distribution into the Hilbert space. In the rest of this summary when we write that the kernel is characteristic we implicitly assume that it is characteristic to $\mathcal{P}(X)$. We make this remark because in some applications one might encounter a kernel that is characteristic to $\mathcal{M}(X)$ or other sets of measures.
To improve our understanding of the MMD distance and its properties, we report here this equivalent representation where MMD is seen as nothing more than a straightforward sum of expectations of k:

$$\begin{aligned}
\gamma_k(\mathbb{P},\mathbb{Q})^2 &= \left\| \int_X k(.,x)d\mathbb{P}(x) - \int_X k(.,y)d\mathbb{Q}(y) \right\|_{\mathcal{H}}^2 \\
&= \left\langle \int_X k(.,x)d\mathbb{P}(x) - \int_X k(.,y)d\mathbb{Q}(y), \int_X k(.,x)d\mathbb{P}(x) - \int_X k(.,y)d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&= \left\langle \int_X k(.,x)d\mathbb{P}(x), \int_X k(.,x')d\mathbb{P}(x') \right\rangle_{\mathcal{H}} + \left\langle \int_X k(.,y)d\mathbb{Q}(y), \int_X k(.,y')d\mathbb{Q}(y') \right\rangle_{\mathcal{H}} \\
&\quad - 2\left\langle \int_X k(.,x)d\mathbb{P}(x), \int_X k(.,y)d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&= \iint_X k(x,x')d\mathbb{P}(x)d\mathbb{P}(x') + \iint_X k(y,y')d\mathbb{Q}(y)d\mathbb{Q}(y') \\
&\quad - 2\iint_X k(x,y)d\mathbb{P}(x)d\mathbb{Q}(y) \\
&= \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x') + \mathbb{E}_{y,y'\sim\mathbb{Q}}k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}k(x,y)
\end{aligned}$$

where the second-last equality follows the same reasoning we applied in the proof of lemma 3.3.1, where we defined a linear functional $T_{\mathbb{P}}[f] := \int_X f(x)d\mathbb{P}(x)$ over $\mathcal{H}$, proved that under hypothesis of lemma 3.3.1 it is bounded and thus Riesz' theorem is applicable. Then we found $\mu_{\mathbb{P}} = \int_X k(.,x)d\mathbb{P}(x) \in \mathcal{H}$ such that $\langle \mu_{\mathbb{P}}, f\rangle = T_{\mathbb{P}}[f]$ for any $f \in \mathcal{H}$.
Since MMD distance can be written as a sum of expectations of k, the plug-in estimator can be used for estimating the MMD from empirical data. This highlights one of the strengths of MMD when compared to other metrics, that is, the easiness and efficiency of calculating it when dealing with practical applications.

Chapter 4

Literature Review

In this section we will review some interesting results from the literature on kernel mean embeddings that we will leverage in our results. The reader should not expect an extensive literature on kernel mean embeddings (for this we refer to [10]), but rather a collection of few selected results that prepare the ground over which we can build the theory of the DRO problem via MMD metric. As a first step we want to clarify the relations between the different notions of universal and characteristic kernels in the framework of RKHS embedding of measures, in addition to clarifying their relations to other common notions of strictly positive definite and integrally strictly positive definite kernels. This will help us understand the influence of the kernel on the RKHS and will provide us an important tool to choose the right type of kernel for our applications. [18] has thoroughly investigated all these issues even in the non compact case, but we present here only the results for X compact.

4.1 Relations Between Characteristic and Universal Kernels to Strictly PD, Integrally Strictly PD

This section introduces the notions of strictly positive definite and integrally strictly positive definite, while finding the connections with the two ideas of characteristic and universal kernel previously discussed.
An *integrally strictly pd* kernel is a measurable and bounded kernel k that satisfies:

$$\int_X \int_X k(x,y)d\mu(x)d\mu(y) > 0 \tag{4.1}$$

for all finite non-zero signed Borel measures μ defined on X, where X is a topological space.
A *strictly pd* kernel is a kernel which for all $n \in \mathbb{N}$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$ and all $x_1, ..., x_n \in X$ we have:

$$\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) \geq 0$$

where for mutually distinct $x_1, ..., x_n \in X$ equality only holds for $\alpha_1 = ... = \alpha_n = 0$.
An integrally strictly pd kernel is always strictly pd, but not vice-versa.
Universal vs. characteristic: a universal kernel is always characteristic but the converse is not true.
Universal vs. Strictly pd: [Corollary 4.3 from [3]] showed that universal kernels are strictly pd.
Characteristic vs. Strictly pd: A characteristic kernel need not be characteristic and the converse also does not hold.
Characteristic vs. Integrally Strictly pd:[Theorem 4 from [20]] proved that integrally strictly pd kernels are characteristic, while the converse in general is not true.

4.2 The Special Case of Radial Kernels

For their importance in the literature a specific section is devoted to the family of radial kernels, which includes Gaussian, Laplacian and other commonly used kernels.
A bounded continuous kernel, k is said to be *radial* on $X \times X$ if there exists $\nu \in \mathcal{M}_b^+([0,\infty))$ such that

$$k(x,y) = \int_{[0,\infty)} e^{-t||x-y||_2^2} d\nu(t), \quad x,y \in X \tag{4.2}$$

For radial kernels on X the following conditions are equivalent:

1. $\text{supp}(\nu) \neq \{0\}$ [Propositions 14 & Theorem 17 from [9]]
2. k is integrally strictly pd
3. k is universal
4. k is strictly pd
5. k is characteristic

4.3 A unifying perspective

The various notions of universal, characteristic, integrally strictly pd and strictly pd kernel have been studied independently by many authors in the literature and, due to this, they might seem and, are in fact presented, as unrelated one to another. However this is not the case and the work of [14] has brought to light the general duality principle that connects these ideas, which will allow us to have a unifying perspective.
In order to do so we first need to unify the notions of strictly pd and integrally strictly pd kernels under one notion which we still call strictly pd but define it as follows:

Definition 4.3.1. Given $\mathcal{P}(X)$, the set of probability measures over X, the kernel k is strictly pd over $\mathcal{P}(X)$ if $\forall \mathbb{P} \in \mathcal{P}(X)$, $||\mu_{\mathbb{P}}||_{\mathcal{H}}^2 = 0 \implies \mathbb{P} = 0$.

Note how the previous notions of integrally strictly pd and strictly pd are a particular case of the above when the probability measures have continuous support and finite support respectively:

$$\begin{aligned} ||\mu_{\mathbb{P}}||_{\mathcal{H}}^2 &= \left\langle \int_X k(.,x)d\mathbb{P}(x), \int_X k(.,y)d\mathbb{P}(y) \right\rangle_{\mathcal{H}} \\ &= \int_X \int_X k(x,y)d\mathbb{P}(x)d\mathbb{P}(y) \end{aligned}$$

In the case of finite support the probability measure can be written as a finite convex combination of dirac measures, i.e. $\mathbb{P} = \sum_{i=1}^n \delta_{x_i}$ yielding:

$$\int_X \int_X k(x,y)d\mathbb{P}(x)d\mathbb{P}(y) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j)$$

We thus recovered the two definitions that were introduced in section 4.1 .
Furthermore we need to introduce the concept of a topological space being *continuously contained* in another topological space: we say $\mathcal{H}$ is continuously contained in $\mathcal{F}$ and we write it as $\mathcal{H} \hookrightarrow \mathcal{F}$, if $\mathcal{H} \subset \mathcal{F}$ and the topology on $\mathcal{H}$ is stronger than the topology on $\mathcal{F}$.

Theorem 4.3.1. *If $\mathcal{H} \hookrightarrow \mathcal{F}$, then the following statements are equivalent.*

1. *k is universal over $\mathcal{F}$ (w.r.t. the topology of $\mathcal{F}$)*

2. *k is characteristic to $\mathcal{F}'$ (dual of $\mathcal{F}$)*

3. *k is strictly pd over $\mathcal{F}'$*

This theorem allows us to have a unifying and more easily understandable view of the various concepts summarized in [18]. In fact it is enough to identify the so-called duality pairs $(\mathcal{F}, \mathcal{F}')$ such that $\mathcal{H} \hookrightarrow \mathcal{F}$. This pairs include for instance $C_b(X)$ and $\mathcal{M}(X)$, the space of signed measures over a LCH space.

4.4 Metrizing the Weak Topology

As already mentioned, MMDs have flourished in many areas of machine learning that require comparing probability distributions, such as two-sample tests, goodness-of-fit tests, generative model and many ohters. In order to correctly compare distributions while working on their embeddings, one would like to have the following behaviors:

- The MMD distance is zero if and only if the probability distributions are the same, i.e. $\gamma_k(\mathbb{P}, \mathbb{Q}) = ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. As the reader might recall this implies having a characteristic kernel.

- If the distance between two embeddings approaches zero, then also according to another standard notion of distance this distance should approach zero. A common standard notion used is the one of narrow or weak convergence. In mathematical terms we would like to have the following behavior: given a sequence of distributions $\{\mathbb{P}_n\}$ and a "target" distribution $\mathbb{P}$ belonging to $\mathcal{P}(X)$, as $\gamma_k(\mathbb{P}_n, \mathbb{P}) = ||\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}}||_{\mathcal{H}} \to 0$ also $\int_X f d\mathbb{P}_n \to \int_X f d\mathbb{P}$ for all $f \in C_b(X)$.

Building up on the results of [19] and [16], [13] investigates in what conditions (e.g. type of kernel, structure of underlying space) convergence in MMD (pseudo)-metric is equivalent to weak convergence on $\mathcal{P}(X)$. In that case we will say that the kernel k *metrizes* the weak convergence of probability measures. In all their results the authors make the following assumption:

(A1) The kernel k is bounded

In light of theorem 3.3.2 this assumption guarantees that the embeddings of the probability measures belong indeed to the RKHS induced by the kernel k and thus that the pseudometric MMD is well defined.
As a first step, the authors proved that integrally strictly pd, continuous and bounded kernels are sufficient to metrize weak topology on $\mathcal{P}(X)$.

Theorem 4.4.1. *Let k be and integrally strictly pd kernel such that $\mathcal{H} \subset \mathcal{C}_0$ and let $\{\mathbb{P}_\alpha\}$ (sequence) and $\mathbb{P}$ be probability measures. If k is continuous, then the following are equivalent.*

(i) $||\mathbb{P}_\alpha - \mathbb{P}||_k \to 0$ *(convergence in strong RKHS topology)*

(ii) $\int_X f d\mathbb{P}_\alpha \to \int_X f d\mathbb{P}$ *for all* $f \in \mathcal{H}$ *(convergence in weak RKHS topology)*

(iii) $\int_X f d\mathbb{P}_\alpha \to \int_X f d\mathbb{P}$ *for all* $f \in \mathcal{C}_0$ *(convergence in weak-* or vague topology)*

(iv) $\int_X f d\mathbb{P}_\alpha \to \int_X f d\mathbb{P}$ *for all* $f \in \mathcal{C}_b$ *(convergence in weak topology)*

Conversely, if (iv) implies (i) for any probability measures $\mathbb{P}_\alpha$ and $\mathbb{P}$, then k is continuous.

When (i) and (iv) are equivalent for all sequences of probability measures, we say that k metrizes the weak convergence of probability measures. Having found the sufficient conditions to metrize weak convergence the authors investigate whether they are also necessary and in order to do so, they distinguish between the two cases of input space X compact or LCH. We report only the result in the compact case:

Theorem 4.4.2. *On a compact Hausdorff space, a bounded, measurable kernel metrizes the weak convergence of probability measures if and only if it is continuous and characteristic to* $\mathcal{P}$

Since, as explained, the metrization of the weak topology is a highly desirable feature of MMD, the choice of characteristic and continuous kernel will be the most used in the applications. Note that this excludes some common kernels such as polynomial kernels, while including kernels such as Gaussian and Laplacian.

Chapter 5

MMD Distributionally Robust Optimization

With all the knowledge acquired in the previous sections at hand, we can finally tackle the problem we most care of: the DRO problem with an ambiguity set based on the MMD metric. Research in this field is still at an early stage and there are just a few papers and established results, nevertheless they constitute a solid base to start and to potentially answer some open questions.
First, let's recall how the DRO problem (here for simplicity considered in its parameterized version) looks:

$$\min_{\theta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in B_\varepsilon(\hat{\mathbb{P}})} \int \ell(\theta,\xi) d\mathbb{Q}(\xi) \tag{5.1}$$

where $B_\varepsilon(\hat{\mathbb{P}}) = \{\mathbb{Q} \in \mathcal{P}(X) \; s.t. \; \gamma_k(\hat{\mathbb{P}}, \mathbb{Q}) \leq \varepsilon\}$ defines our ambiguity set.
For now, we do not know anything about that exotic object defined above which we call the ambiguity set. But in order to work with it and eventually find the supremum over it, it would be interesting to understand its structure. To do so, let's assume for now that the kernel is characteristic and continuous over the compact domain X. As we have seen this provides us some nice properties: injectivity of the embedding, i.e. no loss of information, metrization of the weak topology and guarantees that the embedding exists and belongs to $\mathcal{H}$ (see lemma 3.3.1). Furthermore having a continuous kernel on a compact space, we have $\mathcal{H} \subset C_b(X)$ as a consequence of theorem 3.2.1. In other words all functions in our RKHS are continuous bounded functions. Under these assumptions we can prove that the ambiguity set is compact and convex w.r.t. the weak (weak-* , since they are equivalent on compact spaces) topology. Note that these properties are highly desirable in our framework: convexity is a common requirement in optimization, while compactness together with continuity ensures the existence of the optimum (Weierstrass theorem).

Lemma 5.0.1. *Let X be a compact metric space, $k : X \times X \to \mathbb{R}$ be a continuous and characteristic kernel. Then $\mathbb{B}_\varepsilon(\hat{\mathbb{P}})$ is compact and convex w.r.t. the weak topology.*

Proof. Assuming X compact we have that $\mathcal{P}(X)$ is compact w.r.t. the weak topology [Proposition 8.27 from [5]]. To answer if $\mathbb{B}_\varepsilon(\hat{\mathbb{P}}) \subset \mathcal{P}(X)$ is compact w.r.t. the weak topology we could prove that it is a closed subset of the compact set $\mathcal{P}(X)$. Let $\{\mathbb{Q}_n\} \in \mathbb{B}_\varepsilon(\hat{\mathbb{P}})$ be a convergent sequence w.r.t. the weak/weak* topology, i.e. $\mathbb{Q}_n \rightharpoonup \mathbb{Q} \in \mathcal{P}(X)$. Since MMD metrizes the weak/weak* topology we have that $\mathbb{Q}_n \rightharpoonup \mathbb{Q}$ is equivalent to writing:

$$\forall \delta > 0 \quad \exists N \in \mathbb{N} \quad s.t. \quad \gamma_k(\mathbb{Q}_N, \mathbb{Q}) \leq \delta \tag{5.2}$$

However at the same time we have: $\gamma_k(\mathbb{Q}_N, \hat{\mathbb{P}}) \leq \varepsilon$ because $\mathbb{Q}_N \in \mathbb{B}_\varepsilon(\hat{\mathbb{P}})$. Thus:

$$\gamma_k(\mathbb{Q}, \hat{\mathbb{P}}) \leq \gamma_k(\mathbb{Q}, \mathbb{Q}_N) + \gamma_k(\mathbb{Q}_N, \hat{\mathbb{P}}) \leq \delta + \varepsilon \tag{5.3}$$

Since δ is arbitrary we can conclude that $\gamma_k(\mathbb{Q}, \hat{\mathbb{P}}) \leq \varepsilon$, i.e. $\mathbb{Q} \in \mathbb{B}_\varepsilon(\hat{\mathbb{P}})$. We can conclude that $\mathbb{B}_\varepsilon(\hat{\mathbb{P}})$ is closed and thus compact w.r.t. the weak topology.
Moreover $\mathbb{B}_\varepsilon(\hat{\mathbb{P}}) \subset \mathcal{P}(X)$ is convex because given $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathbb{B}_\varepsilon(\hat{\mathbb{P}})$ and $\lambda \in [0, 1]$ then

$$\begin{aligned}
\gamma_k(\lambda\mathbb{Q}_1 + (1-\lambda)\mathbb{Q}_2, \hat{\mathbb{P}}) &= ||\lambda\mu_{\mathbb{Q}_1} + (1-\lambda)\mu_{\mathbb{Q}_2} - \mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} \\
&= ||\lambda\mu_{\mathbb{Q}_1} + (1-\lambda)\mu_{\mathbb{Q}_2} - \lambda\mu_{\hat{\mathbb{P}}} - (1-\lambda)\mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} \\
&\leq ||\lambda\mu_{\mathbb{Q}_1} - \lambda\mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} + ||(1-\lambda)\mu_{\mathbb{Q}_2} - (1-\lambda)\mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} \\
&= \lambda||\mu_{\mathbb{Q}_1} - \mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} + (1-\lambda)||\mu_{\mathbb{Q}_2} - \mu_{\hat{\mathbb{P}}}||_{\mathcal{H}} \\
&\leq \lambda\varepsilon + (1-\lambda)\varepsilon \\
&= \varepsilon
\end{aligned}$$

□

Now that we have acquired a basic idea on its structure, let's now proceed into understanding how to choose the ambiguity set for a given problem and a given reproducing kernel (later we will discuss how to pick the kernel as well). This can be broken down into two questions: how to choose the center $\hat{\mathbb{P}}$ and the radius ε. For the first task, the more common approach is the data driven one, that is, to rely on the empirical distribution, $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$, where δ is a Dirac measure and $\{\xi\}_{i=1}^n$ are data samples. Nevertheless, for some applications it is standard to set the center to be a known distribution; for instance, in Kalman filters the center distribution is assumed to be Gaussian. The choice of the radius ε can be trickier. Ideally we would like to have it large enough so that, with high probability, the population distribution $\mathbb{P}$, i.e. the "true" distribution from which our data is drawn, belongs to the ambiguity set. At the same time, increasing the radius will lead to a more pessimistic DRO minimax problem, to the point where as the radius grows to include all probability distributions, our DRO problem would be reduced to a worst-case RO problem. To have a better understanding of this behavior, let's analyze an example where the reproducing kernel is assumed to be Gaussian, which we recall is universal and continuous.
Example[22]: For the RKHS associated with the Gaussian kernel, i.e. $k(x, y) = e^{-\sigma||x-y||^2}$, the diameter of the space can be computed:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P}(X),\ ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}} \leq ||\mu_{\mathbb{P}}||_{\mathcal{H}} + ||\mu_{\mathbb{Q}}||_{\mathcal{H}} \leq 2\sup_{x,y}\sqrt{k(x,y)} = 2$$

Thus having an $\varepsilon \geq 2$ would imply having all the distributions in our ambiguity set, reducing DRO to worst-case RO on domain X. Intuitively, $\varepsilon \geq 2$ when the kernel is Gaussian means that we can robustify against all probability distributions. The worst case one would be the probability distribution that assigns the value one to the ξ that maximizes $\ell(\theta, \xi)$ and zero everywhere else.
In practice, when working with a ball centered around the empirical distribution $\mathbb{P}_n$, the radius ε depends on how quickly $\gamma_k(\mathbb{P}_n, \mathbb{P})$ shrinks to zero. In other words, it depends on the empirical accuracy of the MMD distance. It has been shown that under the assumption of a bounded kernel, the MMD distance between empirical sample $\mathbb{P}_n$ and population $\mathbb{P}$ shrinks as $O(n^{-1/2})$:

Lemma 5.0.2 ([10])**.** *Suppose that $k(x, x) \leq M$ for all x. Let $\mathbb{P}_n$ be a n sample empirical approximation to $\mathbb{P}$. Then with probability $1 - \delta$,*

$$\gamma_k(\mathbb{P}_n, \mathbb{P}) \leq 2\sqrt{\frac{M}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{5.4}$$

This concentration inequality highlights one of the strenghts of MMD when compared to other notions of distance between probability measures. Namely that the concentration between empirical and true distribution is independent of the dimension of the data, as opposed to, for example, Wasserstein distance where it shrinks at the rate $O(n^{-1/d})$, where d is the dimension of the data.
Exploiting this lemma we can find a simple upper bound on the population risk just choosing as our radius the quantity on the r.h.s. :

Corollary 5.0.3 (Corollary 3.1 from [21]). *Suppose that $k(x,x) \leq M$ for all x. Set $\varepsilon = 2\sqrt{\frac{M}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$. Then with probability $1-\delta$, we have the following bound on the population risk:*

$$\mathbb{E}_{\xi\sim\mathbb{P}}[\ell(\theta,\xi)] \leq \sup_{\mathbb{Q}:\gamma_k(\mathbb{P}_n,\mathbb{Q})\leq\varepsilon} \mathbb{E}_{\xi\sim\mathbb{Q}}[\ell(\theta,\xi)] \tag{5.5}$$

The r.h.s. is precisely the inner part of the DRO problem exposed in (5.1). From the concentration inequality expressed in lemma 5.0.2 we note that as the number of samples n grows we are more likely to include the "true" population distribution in our ambiguity set. This is a rather remarkable achievement since it implies that while we optimize over all distributions on the ambiguity set $B_\varepsilon(\mathbb{P}_n)$, we are also implicitly optimizing over the population distribution $\mathbb{P}$, thus achieving the best possible degree of generalization.
Our goal now is to find a value of the r.h.s. or an upper bound of it, in order to reduce our problem to a more familiar single variable optimization problem over the parameter θ.
Since the papers that explored this matter were few and not very rigorous, we first explored if the DRO problem, as formulated in (5.1) is well-posed. For one thing a minimax problem might never converge to a solution. As a matter of fact, optimizing a function over two variables, the parameter θ and the distribution $\mathbb{Q}$, will be reduced to finding a saddle point of said function, where the optimal solution (θ^*, $\mathbb{Q}^*$) is a minimum w.r.t. θ and a maximum w.r.t. $\mathbb{Q}$. But we have no guarantees at all that our loss function admits such saddle point. Or in game-theoretic words, we do not know if our game admits an equilibrium. Thankfully we can exploit the framework provided by Sion's minimax theorem. Strong duality ensured by Sion's minimax theorem guarantees us that the minimizer θ^* is indeed the optimum in correspondence of the least favorable distribution $\mathbb{Q}^*$, thus remarking that the model we would obtain has actually a physically interpretable meaning, i.e. it is the best model when the distribution is the worst-case or least favorable one.

Theorem 5.0.4 (Sion's minimax theorem). *Let X be a compact convex subset of a linear topological space and Y a convex subset of a linear topological space. If f is a real-valued function on $X \times Y$ with*

- *f(x,.) upper semicontinuous and quasi-concave on Y, $\forall x \in X$, and*
- *f(.,y) lower semicontinuous and quasi-convex on X, $\forall y \in Y$*

then:

$$\min_{x\in X}\sup_{y\in Y} f(x,y) = \sup_{y\in Y}\min_{x\in X} f(x,y) \tag{5.6}$$

Let's now analyze under what assumptions Sion's minimax theorem can be applied to our problem (5.1). First recall that the ambiguity set $B_\varepsilon(\hat{\mathbb{P}})$ is compact. We will now prove that under the following assumptions on the function $\ell(\theta,\xi)$ the hypotheses of Sion's minimax theorem are satisfied.

Theorem 5.0.5. *Let X be a compact metric space and $\ell(\theta,\xi) : \mathbb{R}^d \times \mathcal{P}(X) \to \mathbb{R}$ be such that it satisfies the following:*

1. *For a fixed $\xi \in X$ $\ell(.,\xi)$ is lower semicontinuous w.r.t. θ, convex w.r.t. θ, measurable and non-negative.*
2. *$\ell(\theta,.) \in C_b(X)$ for all $\theta \in \mathbb{R}^d$.*

then

$$\min_{\theta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in B_\varepsilon(\hat{\mathbb{P}})} E_{\mathbb{Q}}[\ell(\theta,\xi)] = \sup_{\mathbb{Q}\in B_\varepsilon(\hat{\mathbb{P}})} \min_{\theta\in\mathbb{R}^d} E_{\mathbb{Q}}[\ell(\theta,\xi)]$$

Proof. 1. Assuming $\ell(.,\xi)$ lower semicontinuous w.r.t. θ we have $\theta \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ lower semicontinuous as well. To prove it, let $\{\theta_n\}$ be a sequence in $\mathbb{R}^d$ converging to θ_0 ,i.e. $\theta_n \to \theta_0$ as $n \to \infty$.
Assuming $\{\ell(\theta_n,\xi)\}$ to form a sequence of non-negative measurable functions allows us to apply Fatou's lemma. Note that this is a reasonable assumption since ℓ is our loss function, thus it is non-negative and measurable. Fatou's lemma yields:

$$\int \liminf_{n\to\infty} \ell(\theta_n,\xi) d\mathbb{Q} \leq \liminf_{n\to\infty} \int \ell(\theta_n,\xi) d\mathbb{Q}$$

Under the hypothesis of ℓ lower-semicontinuous at θ_0 we have:

$$\ell(\theta_0,\xi) \leq \liminf_{n\to\infty} \ell(\theta_n,\xi)$$

by monotonicity of the Lebesgue integral this implies:

$$\int \ell(\theta_0,\xi) d\mathbb{Q} \leq \int \liminf_{n\to\infty} \ell(\theta_n,\xi) d\mathbb{Q}$$

In conclusion:

$$\int \ell(\theta_0,\xi) d\mathbb{Q} \leq \int \liminf_{n\to\infty} \ell(\theta_n,\xi) d\mathbb{Q} \leq \liminf_{n\to\infty} \int \ell(\theta_n,\xi) d\mathbb{Q}$$

In other words, the function $\theta \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ is lower-semicontinuous at θ_0. Since this reasoning can be applied to any θ in the domain we conclude $\theta \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ is lower-semicontinuous over all its domain.
Moreover we can prove our loss function also inherits convexity from ℓ. Recall definition of convex function: f is convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall \lambda \in [0,1]$. Under the hypothesis of $\ell(.,\xi)$ convex w.r.t. θ for a fixed $\xi \in X$ we have that:

$$\mathbb{E}_{\mathbb{Q}}[\ell(\lambda\theta_1+(1-\lambda)\theta_2,\xi)] \leq \mathbb{E}_{\mathbb{Q}}[\lambda\ell(\theta_1,\xi)+(1-\lambda)l(\theta_2,\xi)] = \lambda\mathbb{E}_{\mathbb{Q}}[\ell(\theta_1,\xi)]+(1-\lambda)\mathbb{E}_{\mathbb{Q}}[\ell(\theta_2,\xi)]$$

where the first inequality follows from convexity of $\ell(.,\xi)$, while the equality follows from linearity of expected value.

2. Now let's study the case where $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ and $\ell(\theta,.) \in C_b(X)$ for all $\theta \in \mathbb{R}^d$.
Let $\{\mathbb{Q}_n\}$ be a sequence converging to $\mathbb{Q}_0$ w.r.t. the weak convergence, i.e. $\mathbb{Q}_n \rightharpoonup \mathbb{Q}_0$ as $n \to \infty$. By definition of weak convergence we have:

$$\int_X f d\mathbb{Q}_n \to \int_X f d\mathbb{Q}_0 \text{ as } n \to \infty \quad \text{for all } f \in C_b(X)$$

Since $\ell(\theta,.) \in C_b(X)$, we conclude that:

$$\mathbb{Q}_n \rightharpoonup \mathbb{Q}_0 \text{ as } n \to \infty \implies \int_X \ell(\theta,.) d\mathbb{Q}_n \to \int_X \ell(\theta,.) d\mathbb{Q}_0$$

In conclusion, $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ is (weakly) continuous over $\mathcal{P}(X)$.

□

Note that no assumption is required on the convexity or concavity of $\ell(\theta, .)$ since, no matter how $\ell(\theta, .)$ looks like, we have that $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\theta, .)]$ is the restriction of a linear function (defined over the space of signed measures $\mathcal{M}(X)$) to the convex space of probability measures $\mathcal{P}(X)$. A linear function is both concave and convex and thus no further assumption is needed in order to satisfy the hypotheses of Sion's minimax principle.
Strong of this theoretical basis, we can actually dig deeper into how to reformulate the inner part of the DRO problem (5.1) in a more tractable way, leveraging an upper bound or better an exact equality. In order to do this we follow the approach of [21], which proposes two simplifications, namely:

- Optimizing over functions in the RKHS instead of directly optimizing over distributions
- Assuming $\ell_\theta := \ell(\theta, .) \in \mathcal{H}$. Even though $\ell_\theta \notin \mathcal{H}$, this assumption is reasonable if $\mathcal{H}$ is dense in the space where ℓ_θ lies.

Since as we saw in (3.1), given two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$, we have $\gamma_k(\mathbb{P}, \mathbb{Q}) = ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}}$, the idea of optimizing over elements of $\mathcal{H}$ appears to make sense. However we have to understand the difference between optimizing over functions in $\mathcal{H}$ and over distributions in $\mathcal{P}(X)$. To do so, let's first understand better how does $\mu_{\mathcal{P}(X)} = \{\mu \in \mathcal{H} \; s.t. \; \mu = \int_X k(x, .) d\mathbb{P}(x)$ for some $\mathbb{P} \in \mathcal{P}(X)\}$, i.e. the space made of the embeddings of all distributions, looks like as a subspace of $\mathcal{H}$.

Lemma 5.0.6. *Let X be a compact metric space, $k : X \times X \to \mathbb{R}$ be a continuous and characteristic kernel. Then the subspace $\mu_{\mathcal{P}(X)} \subset \mathcal{H}$ is compact and convex w.r.t. the topology induced by the norm of $\mathcal{H}$.*

Proof. Recall that assuming X compact we have that $\mathcal{P}(X)$ is compact w.r.t. the weak topology [Proposition 8.27 from [5]]. Now if we can prove that the embedding $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is continuous for every $\mathbb{P} \in \mathcal{P}(X)$ we can conclude that also $\mu_{\mathcal{P}(X)}$ is a compact subset of $\mathcal{H}$.
The embedding is continuous if and only if given a sequence $\{\mathbb{P}_n\} \subset \mathcal{P}(X)$ and $\mathbb{P} \in \mathcal{P}(X)$ the following holds:

$$\mathbb{P}_n \rightharpoonup \mathbb{P} \quad \text{as} \quad n \to \infty \quad \implies \quad \mu_{\mathbb{P}_n} \to \mu_{\mathbb{P}} \quad \text{as} \quad n \to \infty$$

where the convergence on the r.h.s. is the convergence w.r.t the RKHS norm, i.e. $||\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}}||_{\mathcal{H}} \to 0$. Since the kernel is characteristic and continuous over a compact metric space we know MMD metrizes the weak topology (see theorem 4.4.2). This means that $\mathbb{P}_n \rightharpoonup \mathbb{P}$ as $n \to \infty$ is equivalent to $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$. Moreover as we saw in (3.1) $\gamma_k(\mathbb{P}_n, \mathbb{P}) = ||\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}}||_{\mathcal{H}}$. We can then conclude that $\mathbb{P}_n \rightharpoonup \mathbb{P}$ as $n \to \infty$ implies (actually it is equivalent) to $||\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}}||_{\mathcal{H}} \to 0$ as $n \to \infty$.
Moreover $\mathcal{P}(X)$ is clearly convex, i.e. $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X) \implies (1 - \lambda)\mathbb{P} + \lambda\mathbb{Q} \in \mathcal{P}(X) \quad \forall \lambda \in [0, 1]$. To see this recall we can rewrite $\mathcal{P}(X)$ as follows: $\mathcal{P}(X) = \{\mu \in \mathcal{M}(X) : \int_X d\mu = 1\} \cap \cap_{f \geq 0} \{\mu \in \mathcal{M}(X) : \int_X f d\mu \geq 0\}$, thus convexity follows from linearity of the integral and from the trivial fact that the intersection of convex sets is convex.
Now, we can prove this convexity is kept also for $\mu_{\mathcal{P}(X)}$. Let $\mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \in \mu_{\mathcal{P}(X)}$, then:

$$\begin{aligned} \lambda\mu_{\mathbb{P}} + (1 - \lambda)\mu_{\mathbb{Q}} &= \lambda \int_X k(x, .) d\mathbb{P}(x) + (1 - \lambda) \int_X k(x, .) d\mathbb{Q}(x) \\ &= \int_X k(x, .) d(\lambda\mathbb{P} + (1 - \lambda)\mathbb{Q})(x) \in \mu_{\mathcal{P}(X)} \end{aligned}$$

since $\lambda\mathbb{P} + (1 - \lambda)\mathbb{Q} \in \mathcal{P}(X)$. The second equality holds since $\mathbb{P}$ and $\mathbb{Q}$ are finite.
In summary, we have proved that $\mu_{\mathcal{P}(X)} \subset \mathcal{H}$ is compact and convex w.r.t. the topology induced by the RKHS norm, i.e. the strong topology. □

At this point one might be tempted to think that we could restrict our optimization problem over a closed set of functions in $\mathcal{H}$ that are contained in $\mu_{\mathcal{P}(X)}$, so that we are guaranteed to find a solution which is an embedding of some distribution, and moreover we can still keep the properties of convexity and compactness. Unfortunately, this is not feasible since it would mean we have to constrain any function in $\mathcal{H}$ to integrate to one and to be non negative: a rather difficult task. This implies we must do some extra work and further analysis.
Recall that given a characteristic kernel, the embedding $\mathbb{P} \mapsto \int_X k(x,.)d\mathbb{P}(x)$ is guaranteed to be injective, but we have no guarantees on it being surjective and in fact that is not the case. Put differently, when we optimize over functions in $\mathcal{H}$ and thus work on the ambiguity set defined as a RKHS ball, we are implicitly doing a relaxation to our optimization problem. However, we have to keep in mind that ultimately our solution should be an embedding of a distribution and not just an arbitrary function belonging to the RKHS. To make this clearer, if we consider $\mathcal{C} := \{\mu \in \mathcal{H} \; s.t. \; ||\mu_{\hat{\mathbb{P}}} - \mu||_{\mathcal{H}} \leq \varepsilon\}$ as our ambiguity set, as opposed to $B_\varepsilon(\hat{\mathbb{P}}) = \{\mathbb{Q} \in \mathcal{P}(X) \; s.t. \; \gamma_k(\hat{\mathbb{P}}, \mathbb{Q}) \leq \varepsilon\}$ then each element in $B_\varepsilon(\hat{\mathbb{P}})$ has a corresponding embedding in $\mathcal{C}$ but the opposite is not true. Moreover, we have to clearly understand with what topology we are working. As we saw before on $B_\varepsilon(\hat{\mathbb{P}}) \subset \mathcal{P}(X)$ we considered the weak or weak-* topology. In $\mathcal{C}$ we are obviously using the strong RKHS topology induced by the norm. These topologies are in fact equivalent under our assumption of characteristic and continuous kernel. However, while $B_\varepsilon(\hat{\mathbb{P}})$ is compact under the weak topology, $\mathcal{C}$ is not compact under the strong RKHS topology! As a matter of fact in an infinite dimensional normed vector space any ball is not compact under the strong topology [Theorem 2.26 from [11]].
Following the second suggestion of [21] we may assume $\ell_\theta \in \mathcal{H}$. Under our usual assumption of continuous kernel we are guaranteed to have any function in $\mathcal{H}$ continuous, including ℓ_θ. This is a desirable feature for a loss function, since most optimization algorithms require globally continuous loss functions.
In the case of $\ell_\theta \notin \mathcal{H}$ often k is a universal kernel, meaning that ℓ_θ can be approximated arbitrarily well by a member of $\mathcal{H}$. In some particular cases, the kernel need not be universal to have a good approximation of ℓ_θ. For instance let's consider a squared loss , then a having a polynomial kernel with degree two is enough to assume $\ell_\theta \in \mathcal{H}$, provided the domain is compact, since we can write $\ell_\theta(x)$ as an infinite linear combination of $k_2(x,y) := (\langle x, y\rangle)^2$ for some $y \in X$, i.e. $\ell_\theta(x) = \sum_{i=1}^{\infty} a_i k(x, y_i)$.
The additional assumption of $\ell_\theta \in \mathcal{H}$ allows us to leverage the reproducing property of expectation (see lemma 3.3.1) and rewrite the risk $\mathbb{E}_{\xi\sim\mathbb{P}}[\ell(\theta,\xi)]$ as $\langle \ell_\theta, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$. The mean embedding form of the DRO problem is much simpler and we obtain the following expression for the inner problem:

$$\sup_{\mathbb{Q}:\gamma_k(\mathbb{Q},\hat{\mathbb{P}})\leq\varepsilon} \mathbb{E}_{\xi\sim\mathbb{Q}}[\ell_\theta(\xi)] \leq \sup_{\mu\in\mathcal{H}:\|\mu-\mu_{\hat{\mathbb{P}}}\|_{\mathcal{H}}\leq\varepsilon} \langle \ell_\theta, \mu\rangle_{\mathcal{H}} \tag{5.7}$$

where we have an inequality because, as we said before, not every function of $\mathcal{H}$ is an embedding of some probability distribution. In other words, the relaxation we are making is not tight.
Having reformulated the DRO problem as a RKHS function problem rather than a distribution problem allows us to leverage some well known facts from Hilbert spaces and ultimately leads to the following interpretation, which first appeared in [21] and was proved exploiting duality. Here we provide a shorter proof leveraging functional analysis.

Theorem 5.0.7. *Let $\ell_\theta, \mu_{\hat{\mathbb{P}}} \in \mathcal{H}$. We have the following equality:*

$$\sup_{\mu\in\mathcal{H}:\|\mu-\mu_{\hat{\mathbb{P}}}\|_{\mathcal{H}}\leq\varepsilon} \langle \ell_\theta, \mu\rangle_{\mathcal{H}} = \langle \ell_\theta, \mu_{\hat{\mathbb{P}}}\rangle_{\mathcal{H}} + \varepsilon\|\ell_\theta\|_{\mathcal{H}}$$

In particular, the optimal solution is $\mu^* = \mu_{\hat{\mathbb{P}}} + \frac{\varepsilon}{\|\ell_\theta\|_{\mathcal{H}}}\ell_\theta$

Proof. Let's rewrite the constraint as follows:

$$\mu \in \mathcal{H} : \left\| \frac{\mu - \mu_{\hat{\mathbb{P}}}}{\varepsilon} \right\|_{\mathcal{H}} \leq 1$$

Also the objective can be rewritten in the following way:

$$\langle \ell_\theta, \mu \rangle_{\mathcal{H}} = \langle \ell_\theta, \mu - \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}} + \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}} = \varepsilon \left\langle \ell_\theta, \frac{\mu - \mu_{\hat{\mathbb{P}}}}{\varepsilon} \right\rangle_{\mathcal{H}} + \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}}$$

Combining the above we can reformulate our problem in the following manner:

$$\sup_{\mu \in \mathcal{H}: \left\| \frac{\mu - \mu_{\hat{\mathbb{P}}}}{\varepsilon} \right\|_{\mathcal{H}} \leq 1} \varepsilon \left\langle \ell_\theta, \frac{\mu - \mu_{\hat{\mathbb{P}}}}{\varepsilon} \right\rangle_{\mathcal{H}} + \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}} = \varepsilon \|\ell_\theta\|_{\mathcal{H}} + \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}}$$

where the equality follows from the fact that in a Hilbert space $\sup_{f:||f||\leq 1} \langle f, g \rangle = ||g||$.
At the optimum we have:

$$\langle \ell_\theta, \mu^* \rangle_{\mathcal{H}} = \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}} + \varepsilon ||\ell_\theta||_{\mathcal{H}} = \langle \ell_\theta, \mu_{\hat{\mathbb{P}}} \rangle_{\mathcal{H}} + \varepsilon \left\langle \ell_\theta, \frac{\ell_\theta}{||\ell_\theta||_{\mathcal{H}}} \right\rangle_{\mathcal{H}} = \left\langle \ell_\theta, \mu_{\hat{\mathbb{P}}} + \varepsilon \frac{\ell_\theta}{||\ell_\theta||_{\mathcal{H}}} \right\rangle_{\mathcal{H}}$$

In other words: $\mu^* = \mu_{\hat{\mathbb{P}}} + \frac{\varepsilon}{\|\ell_\theta\|_{\mathcal{H}}} \ell_\theta$ □

Combining theorem 5.0.7 with (5.7), [21] shows that minimizing the risk plus a norm on ℓ_θ leads to a high probability bound on out-of-sample performance:

$$\min_\theta \sup_{\mathbb{Q}: \gamma_k(\mathbb{Q}, \hat{\mathbb{P}}) \leq \varepsilon} \mathbb{E}_{\xi \sim \mathbb{Q}}[\ell_\theta(\xi)] \leq \min_\theta \left(\mathbb{E}_{\xi \sim \hat{\mathbb{P}}}[\ell_\theta(\xi)] + \varepsilon \|\ell_\theta\|_{\mathcal{H}} \right) \tag{5.8}$$

We have thus substituted the inner problem with an upper bound of it, which is a regularized problem, where the penalization is made w.r.t. the Hilbert norm of the loss. Note that this is quite different from the common penalties in kernel methods which penalize the norm of the model rather than the norm of the loss. It is not entirely understood yet what this penalty implies in our learning framework, investigating this issue can be a possible direction for future works.
A natural question would be now to ask if we can further tighten this bound to eventually reach an equality in some special conditions. [7] has investigated thoroughly this matter in the general contest of integral probability metrics, reaching the following result which we will leverage for MMD metric:

Theorem 5.0.8 (Theorem 1 from [7])**.** *Let $\mathcal{F} := \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset \mathcal{H}$ and $\mathbb{P} \in \mathcal{P}(X)$. For any $\ell \in \mathcal{H}$ and for all $\varepsilon > 0$*

$$\sup_{\mathbb{Q} \in \mathcal{P}(X): \gamma_k(\mathbb{Q}, \mathbb{P}) \leq \varepsilon} \langle \ell, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \langle \ell, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \Lambda_{\mathcal{F}, \varepsilon}(\ell)$$

where the penalty $\Lambda_{\mathcal{F}, \varepsilon}(\ell) = (J_{\mathbb{P}}(\ell) \,\bar{\star}\, \varepsilon ||\ell||_{\mathcal{H}})$ where $\bar{\star}$ is the infimal convolution operator and $J_{\mathbb{P}}(\ell) = \sup_{\mu_{\mathbb{Q}} \in \mu_{\mathcal{P}(X)}} \langle \ell, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - \langle \ell, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$.
This result allows us to turn the infinite-dimensional optimization problem on the l.h.s. to a more familiar and tractable penalty-based regularization objective, without resorting to an upper bound. But the penalty $\Lambda_{\mathcal{F}, \varepsilon}(\ell)$ is still obscure and without a clear interpretation. That's what drove us to identify the conditions under which the above penalty is only reduced to be $\varepsilon ||\ell||_{\mathcal{H}}$, in other words, when the statement of theorem 5.0.7 is reduced to equality. Our result is stated in the following theorem:

Theorem 5.0.9. *Consider the inner problem of DRO (5.1) with the center of the ambiguity set being equal to the empirical distribution, i.e.* $\hat{\mathbb{P}} = \mathbb{P}_n$. *Then there exists a number of data samples* $N \in \mathbb{N}$ *such that for all* $n > N$ *the following equality holds:*

$$\sup_{\mathbb{Q}:\gamma_k(\mathbb{Q},\mathbb{P}_n)\leq\varepsilon} \mathbb{E}_{\xi\sim\mathbb{Q}}[\ell_\theta(\xi)] = \langle \ell_\theta, \mu_{\mathbb{P}_n}\rangle_{\mathcal{H}} + \varepsilon\|\ell_\theta\|_{\mathcal{H}}$$

Proof. The consequence of lemma 2 of [7] is that if $\min(J_{\mathbb{P}_n}(\ell_\theta), \varepsilon||\ell_\theta||_{\mathcal{H}})$ is subadditive then $\Lambda_{\mathcal{F},\varepsilon}(\ell) = \min(J_{\mathbb{P}}(\ell), \varepsilon||\ell||_{\mathcal{H}})$. From the proof of lemma 2 of [7] we know that both $J_{\mathbb{P}_n}(\ell_\theta)$ and $\varepsilon||\ell_\theta||_{\mathcal{H}}$ are subadditive. This implies that, if $\min(J_{\mathbb{P}_n}(\ell_\theta), \varepsilon||\ell_\theta||_{\mathcal{H}}) = \varepsilon||\ell_\theta||_{\mathcal{H}}$ for some $\varepsilon > 0$, then $\Lambda_{\mathcal{F},\varepsilon}(\ell_\theta) = \varepsilon||\ell_\theta||_{\mathcal{H}}$.
Thus if we show that:

$$\varepsilon||\ell_\theta||_{\mathcal{H}} \leq J_{\mathbb{P}_n}(\ell_\theta)\ \forall \ell_\theta \in \mathcal{H} \tag{5.9}$$

for some $\varepsilon > 0$ we have proven our claim.
Recall that from the concentration inequality expressed in 5.0.2 it makes sense to set

$$\varepsilon = 2\sqrt{\frac{M}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} = \frac{\varepsilon_0}{\sqrt{n}}$$

Moreover we can rewrite $J_{\mathbb{P}_n}(\ell_\theta)$ as follows

$$J_{\mathbb{P}_n}(\ell_\theta) = \sup_{\mathbb{Q}\in\mathcal{P}(X)} \int_X \ell_\theta d\mathbb{Q} - \int_X \ell_\theta d\mathbb{P}_n = \max_{\xi\in X} \ell_\theta(\xi) - \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(\xi_i)$$

since among all the distributions, the one that maximizes $\int_X \ell_\theta d\mathbb{Q}$ is a Dirac delta which assigns all the mass to $\xi^* = \arg\max_{\xi\in X} \ell_\theta(\xi)$. Note that we are allowed to substitute $\sup_{\xi\in X} \ell_\theta(\xi)$ with $\max_{\xi\in X} \ell_\theta(\xi)$ because $\ell_\theta(.)$ is continuous and X is compact.
Now we are able to rewrite (5.9) as:

$$\frac{\varepsilon_0}{\sqrt{n}}||\ell_\theta||_{\mathcal{H}} \leq \max_{\xi\in X} \ell_\theta(\xi) - \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(\xi_i) \tag{5.10}$$

It is clear that as n grows the l.h.s. shrinks while the r.h.s. is not affected. This implies that there exists a $N \in \mathbb{N}$ such that for all $n > N$ the above inequality holds, which proves our claim. □

This result is telling us that as the number of samples grows, we can, with more certainty, reformulate the inner part of the data-driven DRO problem as a penalized problem. As we mentioned before it is not entirely understood how this penalty affects our model and this is still an open question, yet it is of much interest to have reformulated an infinite dimensional problem as a more familiar penalty based problem and to have provided some rather tangible conditions under which this is attained.
So far we have almost completely ignored one, rather crucial, component of the DRO problem: the reproducing kernel. In fact, we assumed the kernel to be characteristic and continuous and we did all our considerations according to this assumption. Finally it has come the point to address this issue: what happens to our optimization problem when we change the kernel? What properties are kept and what are lost?
Let's begin with understanding better what a continuous and characteristic kernel implies in our framework. Intuitively, a characteristic kernel guarantees that each probability distribution is mapped to a different function in the RKHS. This implicitly means that our problem remains infinite dimensional and that we can still separate all possible probability distributions. In other words our DRO problem is the least conservative it can be. Conversely having a reproducing

kernel that induces a small and potentially finite dimensional RKHS, e.g. a linear kernel, forces kernel DRO to robustify against a large set of distributions, resulting in conservativeness. At the extreme if RKHS $= \{0\}$ and thus $B_\varepsilon(\hat{\mathbb{P}}) = \{0\}$ then DRO would be reduced to worst-case RO. To better understand this behavior we provide some examples below:
Example (linear kernels) [22]: let $\hat{\mathbb{P}} = \mathcal{N}(0,1)$, and $\mathbb{Q}_v = \mathcal{N}(0, v^2)$ and $\mathcal{H}$ the RKHS induced by the linear kernel $k_1(x,y) := xy$. What is $\gamma_k(\hat{\mathbb{P}}, \mathbb{Q}_v)$?

$$\begin{aligned}\gamma_k(\hat{\mathbb{P}}, \mathbb{Q}_v) &= \|\mu_{\hat{\mathbb{P}}} - \mu_{\mathbb{Q}_v}\|_{\mathcal{H}} \\ &= \left| \int_X xt \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x\|}{2}} dx - \int_X xt \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{\|x\|}{2v^2}} dx \right| \\ &= \left| t\mathbb{E}_{X\sim\hat{\mathbb{P}}}[X] - t\mathbb{E}_{X\sim\mathbb{Q}_v}[X] \right| = 0 \quad \forall t\end{aligned}$$

A linear kernel can not separate two probability distributions that have the same mean, thus any ambiguity set containing $\hat{\mathbb{P}}$ contains also a $\mathbb{Q}_v$ for any $v \neq 0$.
As the next example shows a similar constraint holds for a polynomial kernel of arbitrary degree:
Example (polynomials kernels) [22]: kernel DRO with the second-order polynomial kernel $k_2(x,y) := (1 + \langle x, y\rangle)^2$ and a singleton ambiguity set $\mathcal{K} = \{\hat{\mathbb{P}}\}$ robustifies against all distributions sharing the first two moments with $\hat{\mathbb{P}}$. This is equivalent to DRO with known first two moments. More generally, the choice of the pth-order polynomial kernel $k_p(x,y) := (1 + \langle x, y\rangle)^p$ corresponds to DRO with known first p moments. This is a direct consequence of the information retained by the kernel mean embedding. [10] showed that, for a polynomial kernel $k_p(x,y)$, the embedding can be written explicitly as:

$$\begin{aligned}\mu_{\mathbb{P}}(t) &= \int_X (\langle x, t\rangle + 1)^p d\mathbb{P}(x) \\ &= 1 + \binom{p}{1}\langle m_{\mathbb{P}}(1), t\rangle + \binom{p}{2}\langle m_{\mathbb{P}}(2), t^{(2)}\rangle + \binom{p}{3}\langle m_{\mathbb{P}}(3), t^{(3)}\rangle + ...\end{aligned}$$

where $m_{\mathbb{P}}(i) := \int_X x^i d\mathbb{P}(x)$ denotes the ith moment of the distribution $\mathbb{P}$.
The choice of the kernel is really application dependent and so it is difficult to provide a ready-to-use recipe that works in any case. Nonetheless, we hope we provided the reader with some basic understanding and knowledge of problems one may incur while working with kernels. In the end, we can confidently say that a characteristic and continuous kernel is a good starting point and a universal kernel can enable some extra features, such as being able to assume that any continuous loss function belongs to the RKHS. Yet in some applications, simpler kernels could accomplish the job and effectively lower computation costs. This is still an active area of research and we hope we will be able to provide more insights into this topic in some future works.

Chapter 6

Conclusion

In summary, this thesis provided an introductory view on the topic of distributionally robust optimization when the ambiguity set is defined as a maximum mean discrepancy ball in the space of probability measures. Actually it provided a detailed study path that any researcher or student can follow while approaching this problem. Moreover, with the formal proofs contained especially in chapter 5, we hope we provided the reader a clear method of thinking and reasoning that one should hold while approaching this kind of problems. We also provided few new results and reviewed many of the existing ones through a different perspective. In the end, the goal of this short thesis was to unveil the connections between the elements involved in the DRO problem such as the ambiguity set, the kernel or the loss function and their practical implications in the learning framework, unifying under a unique umbrella results that were independently derived by different authors in the last years. Nevertheless many problems remain unresolved, including the major bottleneck (for now) of this approach which is computation. In fact, even though we were able to reformulate the DRO problem as a regularized problem we have not yet understood how to make this problem computational. This is why we will next concentrate on a fixed scenario where the chosen kernel is standard and then hopefully complete the investigation of this problem, reaching a computational formulation and providing applications to real-world examples.

Bibliography

[1] A Berlinet and C Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2004.

[2] M Binkowski et al. "Demystifying MMD GANs". In: *arXiv preprint arXiv:1801.01401* (2018).

[3] C Carmeli et al. "Vector valued reproducing kernel Hilbert spaces and universality". In: *Analysis and Applications, 8* (2010), pp. 19–61.

[4] GK Dziugaite, DM Roy, and Z Ghahramani. "Training generative neural networks via maximum mean discrepancy optimization". In: *arXiv preprint arXiv:1505.03906* (2015).

[5] Manfred Einsiedler and Thomas Ward. *Analysis, Spectral Theory, and Applications*. Springer, 2017.

[6] A Gretton et al. "A kernel method for the two sample problem". In: *Information Processing Systems 19* (2007), pp. 513–520.

[7] H Husain. "Distributional Robustness with IPMs and links to Regularization and GANs". In: *arXiv preprint arXiv:2006.04349* (2020).

[8] W Jitkrittum et al. "A Linear-Time Kernel Goodness-of-Fit Test". In: *Advances in Neural Information Processing Systems* (2017).

[9] CA Micchelli, Y Xu, and H Zhang. "Universal kernels". In: *Journal of Machine Learning Research, 7* (2006), pp. 2651–2667.

[10] K Muandet et al. "Kernel mean embedding of distributions: A review and beyond". In: *arXiv preprint arXiv:1605.09522* (2017).

[11] Bryan P Rynne and Martin A Youngson. *Linear Functional Analysis*. Springer Undergraduate Mathematics Series, 2008.

[12] B Scholkopf, A Smola, and F Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2002.

[13] CJ Simon-Gabriel, A Barp, and L Mackey. "Metrizing weak convergence with maximum mean discrepancies". In: *arXiv preprint arXiv:2006.09268* (2020).

[14] CJ Simon-Gabriel and B Scholkopf. "Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions". In: *Journal of Machine Learning Research 19 (44)* (2018), pp. 1–29.

[15] A Smola et al. "A Hilbert space embedding for distributions". In: *International Conference on Algorithmic Learning Theory* (2007), pp. 13–31.

[16] BK Sriperumbudur. "On the optimal estimation of probability measures in weak and strong topologies". In: *Bernoulli 22* (2016), pp. 1839–1893.

[17] BK Sriperumbudur, K Fukumizu, and GRG Lanckriet. "On the relation between universality, characteristic kernels and RKHS embedding of measures". In: *JMLR Workshop and Conference Proceedings 9* (2010), pp. 781–788.

[18] BK Sriperumbudur, K Fukumizu, and GRG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures". In: *Journal of Machine Learning Research 12* (2011), pp. 2389–2410.

[19] BK Sriperumbudur et al. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research 11* (2010), pp. 1517–1561.

[20] BK Sriperumbudur et al. "Kernel choice and classifiability for RKHS embeddings of probability distributions". In: *Advances in Neural Information Processing Systems 22* (2009), pp. 1750–1758.

[21] M Staib and S Jegelka. "Distributionally robust optimization and generalization in kernel methods". In: *Advances in Neural Information Processing Systems 32* (2019), pp. 9134–9144.

[22] JJ Zhu et al. "Kernel Distributionally Robust Optimization". In: *arXiv preprint arXiv:2006.06981* (2021).