Kernel Distributionally Robust Optimization

Jia-Jie Zhu
Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany
jia-jie.zhu@tuebingen.mpg.de

Wittawat Jitkrittum *
Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany
wittawatj@gmail.com

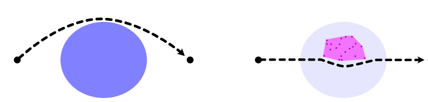
Moritz Diehl
Department of Microsystems Engineering
& Department of Mathematics
University of Freiburg
Freiburg, Germany
moritz.diehl@imtek.uni-freiburg.de

Bernhard Schölkopf
Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany
bernhard.schoelkopf@tuebingen.mpg.de

Abstract

This paper is an in-depth investigation of using kernel methods to immunize optimization solutions against distributional ambiguity. We propose *kernel distributionally robust optimization* (K-DRO) using insights from the robust optimization theory and functional analysis. Our method uses reproducing kernel Hilbert spaces (RKHS) to construct ambiguity sets. It can be reformulated as a tractable program by using the conic duality of moment problems and an extension of the RKHS representer theorem. Our insights reveal that universal RKHSs are large enough for K-DRO to be effective. This paper provides both theoretical analyses that extend the robustness properties of kernel methods, as well as practical algorithms that can be applied to general optimization problems, not limited to kernelized models.

1 Introduction



Imagine a hypothetical scenario in the figure where we want to arrive at a destination while avoiding unknown obstacles. A *worst-case robust optimization* approach is then to avoid the entire unsafe area (left, blue). Suppose we have historical locations of the obstacles (right, dots). We may choose to avoid only the convex polytope that contains all the samples (pink). This *data-driven robust decision-making* idea improves efficiency while retaining robustness.

The concept of distributional ambiguity concerns the uncertainty of uncertainty — the underlying probability measure is only partially known or subject to change. This idea is by no means a new one. The classical moment problem concerns itself with estimating the worst-case risk expressed by $\max_{P\in\mathcal{K}} \int l \, dP$ where l is some loss function. The constraint $P \in \mathcal{K}$ describes the *distribution ambiguity*, i.e., P is only known to live within a subset $\mathcal{K}$ of probability measures. The solution to the moment problem gives the risk under some worst-case distribution within $\mathcal{K}$. To make decisions that will minimize this worst-case risk is the idea of *distributionally robust optimization* (DRO) [14, 29].

Many of today's learning tasks suffer from various manifestations of distributional ambiguity — e.g., covariate shift, adversarial attacks, simulation to reality transfer — phenomena that are caused by the discrepancy between training and test distributions. Kernel methods are known to possess

*Now affiliated with Google Research.

robustness properties, e.g., [11]. However, this robustness only applies to kernelized models under light-tail contamination. This paper extends the robustness of kernel methods using the robust counterpart formulation techniques [4] as well as the principled conic duality theory [33]. We term our approach *kernel distributionally robust optimization* (K-DRO), which can robustify general optimization solutions not limited to kernelized models.

Contributions. The *main contributions* of this paper are, on the one hand, the theory of solving distributionally robust optimization via the combination of robust and stochastic optimization theory, functional analysis, and kernel methods in Section 3. On the other hand, we provide tractable algorithms in Section 3.2, and demonstrate how to apply K-DRO in practice in Section 4.

In addition, this paper contains several non-trivial insights not present in the literature.

- The K-DRO reformulation, a majorization-minimization algorithm, can be viewed as robustification by kernelization. Its applications are not limited to kernelized models.
- We give complete self-contained proofs that shed light on the connection between reproducing kernel Hilbert spaces (RKHSs) and DRO. Furthermore, we show that universal RKHSs are large enough for DRO from the perspective of functional analysis.
- This paper extends the robustness property of kernel-based learning methods by establishing a DRO counterpart of the RKHS representer theorem.
- Finally, we provide unifying perspectives of existing DRO methods: we discuss generalizing the K-DRO reformulation to using *integral probability metrics*, which includes the Wasserstein distance. Using the flexibility of choosing kernels, we show K-DRO is a generalization of DRO with moment constraints.

2 Background

Notation. $\mathcal{X} \subset \mathbb{R}^d$ denotes the input domain, which is assumed to be compact unless otherwise specified. $\mathcal{P} := \mathcal{P}(\mathcal{X})$ denotes the set of all Borel probability measures on $\mathcal{X}$. We use $\hat{P}$ to denote the empirical distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N}\delta_{\xi_i}$, where δ is a Dirac measure and $\{\xi_i\}_{i=1}^{N}$ are data samples. We refer to the function $\delta_{\mathcal{C}}(x) := 0$ if $x \in \mathcal{C}$, ∞ if $x \notin \mathcal{C}$, as the indicator function. $\delta^*_{\mathcal{C}}(f) := \sup_{\mu\in\mathcal{C}}\langle f, \mu\rangle_{\mathcal{H}}$ is the support function of $\mathcal{C}$. S_N denotes the N-dimensional simplex. ri denotes the relative interior of a set. A function f is upper semicontinuous on $\mathcal{X}$ if $\limsup_{x\to x_0} f(x) \leq f(x_0), \forall x_0 \in \mathcal{X}$; it is proper if it is not identically $-\infty$.

2.1 Robust and distributionally robust optimization

Robust optimization (RO) [4] studies mathematical decision-making under uncertainty. It solves the min-max problem (omitting constraints) $\min_\theta \sup_{\xi\in\mathcal{X}} l(\theta, \xi)$, where $l(\theta, \xi)$ denotes a general loss function, θ is the decision variable, and ξ is a variable representing the uncertainty. Intuitively, RO makes the decision assuming an adversarial scenario, as reflected in taking supremum w.r.t. ξ. For this reason, it is often referred to as the *worst-case RO*. A typical approach to solving RO is via reformulating the min-max program using duality to obtain a single minimization problem. The elegance of its theory and effective computation has made RO a powerful tool in applications where safety is essential, e.g., control, robotics. In contrast, distributionally robust optimization (DRO) minimizes the expected loss assuming the worst-case distribution:

$$\min_{\theta} \sup_{P\in\mathcal{K}} \left\{ \int l(\theta, \xi)\, dP(\xi) \right\}, \tag{1}$$

where $\mathcal{K} \subseteq \mathcal{P}$, called the *ambiguity set*, is a subset of distributions, e.g., all distributions with the given mean and variance. The goal of (1) is then to make the decision θ that will minimize the worst-case risk. Compared with RO, DRO only immunizes the solution against a subset $\mathcal{K}$ of distributions on $\mathcal{X}$ and is, therefore, less conservative (since $\sup_{P\in\mathcal{K}}\{\int l(\theta, \xi)\, dP(\xi)\} \leq \sup_{\xi\in\mathcal{X}} l(\theta, \xi)$).

The inner problem of (1), historically known as the *(generalized) moment problem*, estimates the worst-case risk under uncertainty in distributions. The modern approaches (e.g., [20] and [22, 33, 7, 26, 41, 40, 44]) typically seek a sharp (i.e., zero duality gap) upper bound via duality. This duality, rigorously justified in [33], is different from that in the Euclidean space because infinite-dimensional convex sets can become pathological. Using that methodology, we can reformulate DRO (1) into a single solvable minimization problem.

Existing DRO approaches can be grouped into three main categories by the type of ambiguity sets used: DRO with moment constraints, likelihood bounds, and Wasserstein distance. DRO with (finite-order) moment constraints has been studied in [14, 29, 45]. The authors of [2, 21, 24, 42] studied DRO using likelihood bounds as well as $\phi-$divergence. Wasserstein-distance-based DRO has been studied by the authors of [23, 43, 18, 8, 15], and applied in a large body of literature. Notably, the authors of [32] applied Wasserstein-DRO to a kernel-based learning task, which should not be confused with K-DRO in this paper.

2.2 Reproducing kernel Hilbert spaces

In this section, we give a brief introduction to reproducing kernel Hilbert spaces (RKHSs). Full coverage on the topic can be found in [5, 39, 31]. A symmetric function $k\colon \mathcal{X} \times \mathcal{X}$ is called a positive definite kernel if $\sum_{i=1}^{n}\sum_{i=1}^{n} a_i a_j k(x_i, x_j) \geq 0$ for any $n \in \mathbb{N}$, $\{x_i\}_{i=1}^{n} \subset \mathcal{X}$, and $\{a_i\}_{i=1}^{n} \subset \mathbb{R}$. Given a positive definite kernel k, there exists a feature map $\phi\colon \mathcal{X} \to \mathcal{H}$, for which $k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$, where $\mathcal{H}$ is a space of real-valued functions on $\mathcal{X}$. The space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS). It is equipped with the *reproducing property*: $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}, x \in \mathcal{X}$. By convention, we will interchangeably write $\phi(x)$ and $k(x,\cdot)$. Properties of the functions in $\mathcal{H}$ are inherited from the properties of k. For instance, if k is continuous, then any $f \in \mathcal{H}$ is continuous. A continuous kernel k on a compact metric space $\mathcal{X}$ is said to be *universal* if $\mathcal{H}$ is dense in $C(\mathcal{X})$ [39, Section 4.5]. A universal $\mathcal{H}$ can thus be considered a large RKHS since any continuous function can be approximated arbitrarily well by a function in $\mathcal{H}$. An example of a universal kernel is the Gaussian kernel $k(x,y) = \exp\left(-\frac{-\|x-y\|_2^2}{2\sigma^2}\right)$ defined on $\mathcal{X}$ where $\sigma > 0$ is the bandwidth parameter.

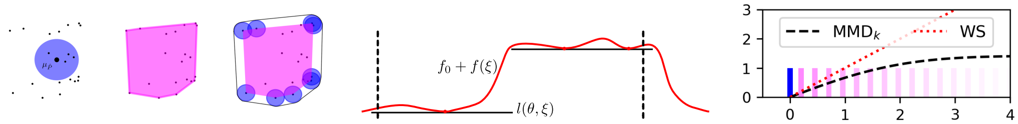
RKHSs first gained widespread attention following the advent of the kernelized support vector machine for classification problems [13]. More recently, the use of RKHSs has been extended to manipulating and comparing probability distributions via kernel mean embedding [35]. Given a distribution P, and a (positive definite) kernel k, the *kernel mean embedding* of P is defined as $\mu_P := \int k(x,\cdot)\, dP$. If $\mathbb{E}_{x\sim P}[k(x,x)] < \infty$, then $\mu_P \in \mathcal{H}$ [35, Section 1.2]. The reproducing property allows one to easily compute the expectation of any function $f \in \mathcal{H}$ since $\mathbb{E}_{x\sim P}[f(x)] = \langle f, \mu_P\rangle_{\mathcal{H}}$. Embedding distributions into $\mathcal{H}$ also allows one to measure the distance between distributions in $\mathcal{H}$. If k is universal, then the mean map $P \mapsto \mu_P$ is injective on $\mathcal{P}$ [19]. With a universal $\mathcal{H}$, given two distributions P, Q, $\|\mu_P - \mu_Q\|_{\mathcal{H}}$ defines a metric. This quantity is known as the maximum mean discrepancy (MMD) [19]. With $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f\rangle_{\mathcal{H}}}$ and the reproducing property, it can be shown that $\|\mu_P - \mu_Q\|_{\mathcal{H}}^2 = \mathbb{E}_{x,x'\sim P} k(x,x') + \mathbb{E}_{y,y'\sim Q} k(y,y') - 2\mathbb{E}_{x\sim P, y\sim Q} k(x,y)$, allowing the plug-in estimator to be used for estimating the MMD from empirical data. The MMD is an instance of the class of integral probability metrics (IPMs), and can equivalently be written as $\|\mu_P - \mu_Q\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}\leq 1} \int f\, d(P-Q)$, where f is a witness function [19, 36]. Figure 1c illustrates an example of MMD.

3 Main results

To solve the DRO problem (1), we need two essential elements: an appropriate ambiguity set that contains meaningful distributions and a reformulation that is both *sharp* and *tractable*. We now present the kernel distributionally robust optimization (K-DRO) and its reformulation, which we will show to satisfy those requirements:

$$\min_{\theta} \sup_{P,\mu} \left\{ \int l(\theta, \xi)\, dP(\xi) \colon \int \phi\, dP = \mu, P \in \mathcal{P}, \mu \in \mathcal{C} \right\}, \tag{2}$$

where $\mathcal{H}$ is an RKHS whose feature map is ϕ. Both sides of the constraint $\int \phi\, dP = \mu$ are functions in $\mathcal{H}$. Note μ can be viewed as a generalized moment vector, which is constrained to lie within the set $\mathcal{C} \subseteq \mathcal{H}$, referred to as an (RKHS) uncertainty set. Let us denote the set of all feasible distributions in (2) as $\mathcal{K}_{\mathcal{C}} = \{P\colon \int \phi\, dP = \mu, \mu \in \mathcal{C}, P \in \mathcal{P}\}$, i.e., $\mathcal{K}_{\mathcal{C}}$ is the ambiguity set. Intuitively, the set $\mathcal{C}$ restricts the RKHS embeddings of distributions in the ambiguity set $\mathcal{K}_{\mathcal{C}}$. In this paper, we take a geometric perspective to construct $\mathcal{C}$ using convex sets in $\mathcal{H}$. Given data samples $\{\xi_i\}_{i=1}^{N}$, we outline various choices for $\mathcal{C}$ in the left column of Table 1 (and 3 in the appendix), and illustrate our intuition in Figure 1 . We make the following regularity assumptions used in the proofs.



(a) RKHS uncertainty sets $\mathcal{C}$ (b) Interpretation of K-DRO (c) MMD and Wasserstein dist.

Figure 1: **(a)**: Geometric intuition for choosing uncertainty set $\mathcal{C}$ in $\mathcal{H}$ such as norm-ball, polytope, and Minkowski sum of sets. The scattered points are the embeddings of empirical samples. See Table 3 for more examples. **(b)**: Geometric interpretation of K-DRO (3). The (red) curve depicts f_0+f, which *majorizes* $l(\theta,\cdot)$ (black). The horizontal axis is ξ. The dashed lines denote the boundary of the domain $\mathcal{X}$. **(c)**: Comparing the distributions of δ_0 and $\delta_z, z>0$ using MMD with a Gaussian kernel and Wasserstein distance.

Table 1: Examples of support functions for K-DRO. See Table 3 for more details.

RKHS uncertainty set $\mathcal{C}$	Support function $\delta^*_{\mathcal{C}}(f)$
RKHS norm-ball $\mathcal{C}=\{\mu\colon \|\mu-\mu_{\hat{P}}\|_{\mathcal{H}}\le\epsilon\}$ $(\hat{P}=\sum_{i=1}^N \frac{1}{N}\delta_{\xi_i})$	$\frac{1}{N}\sum_{i=1}^N f(\xi_i)+\epsilon\|f\|_{\mathcal{H}}$
Polytope $\mathcal{C}=\operatorname{clconv}\{\phi(\xi_1),\ldots,\phi(\xi_N)\}$	$\max_i f(\xi_i)$ (equivalent to scenario approach [10])
Minkowski sum $\mathcal{C}=\sum_{i=1}^N \mathcal{C}_i$	$\sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f)$
Whole space $\mathcal{C}=\mathcal{H}$	Indicator function of $\{f=0\}$ (equivalent to RO)

Assumption 3.1. *$l(\theta,\xi)$ is proper, upper semicontinuous in ξ. $\mathcal{C}$ is closed convex.* $\operatorname{ri}(\mathcal{K}_{\mathcal{C}})\neq\emptyset$.

We first give the main result for solving K-DRO (2). All proofs are deferred to the appendix.

Theorem 3.1 (K-DRO reformulation)**.** *Under Assumption 3.1,* (2) *is equivalent to solving*

$$\min_{\theta, f_0\in\mathbb{R}, f\in\mathcal{H}} \; f_0+\delta^*_{\mathcal{C}}(f) \quad \text{subject to } l(\theta,\xi)\le f_0+f(\xi),\ \forall\xi\in\mathcal{X} \tag{3}$$

where $\delta^*_{\mathcal{C}}(f) := \sup_{\mu\in\mathcal{C}}\langle f,\mu\rangle_{\mathcal{H}}$ is the support function of $\mathcal{C}$, i.e., strong duality *holds for the inner moment problem for any θ point-wise.*

The theorem holds regardless of the dependency of l on θ, e.g., non-convexity. Formulation (3) has a clear geometric interpretation: we find a function f_0+f that *majorizes* $l(\theta,\cdot)$ and subsequently minimize a surrogate loss involving f_0 and f. This is illustrated in Figure 1b. Proposition 3.1 gives us a flexible tool to solve K-DRO for various uncertainty sets $\mathcal{C}$. We outline a few tractable formulations in Table 1, while more are given in Table 3. While we prove the results for general $\mathcal{C}$, we pay special attention to the RKHS norm-balls since they provide the elementary topological sets.

If $\mathcal{H}$ is small, the uncertainty set $\mathcal{C}$ (e.g., RKHS norm-ball) may fail to be meaningful. For example, if $\mathcal{H}=\{0\}$, then the MMD is too weak to separate any distinct distributions. This renders K-DRO (3) overly conservative since we can only choose $f=0$ in (3) — it is precisely reduced to worst-case RO. On the other extreme, if $\mathcal{H}$ is too large, e.g., the space of all bounded measurable functions, the metric induced by $\mathcal{H}$ becomes the *total variation* distance [37]. While the induced topology is strong, (3) has a trivial solution $f=l, f_0=0$, which renders the reformulation (3) intractable since it is reduced to (2). Let us consider a non-trivial small RKHS to understand the effect of size on K-DRO.

Example 3.1 (Non-universal kernels)**.** Consider distributions $\hat{P}=\mathcal{N}(0,1), Q_v=\mathcal{N}(0,v^2)$ and $\mathcal{H}_1$ induced by the linear kernel $k_1(x,y):=xy$. $\mathcal{H}_1$ is small since it only contains linear functions. $\operatorname{MMD}_{k_1}(\hat{P},Q_v)=0, \forall v\neq 0$ since they share the first moment. Therefore, any $\mathcal{C}$ that contains $\mu_{\hat{P}}$ also contains all the distributions in $\{\mu_{Q_v}, v\neq 0\}$. Then K-DRO must robustify against a large set of distributions, resulting in conservativeness.

Example 3.2 (Reduction to DRO with moment constraints)**.** K-DRO with the second-order polynomial kernel $k_2(x,y):=(1+x^\top y)^2$ and a singleton uncertainty set $\mathcal{C}=\{\mu_{\hat{P}}\}$ immunizes against all distributions sharing the first two moments with $\hat{P}$. This is equivalent to DRO with known first two moments, such as in [14, 29].

If $\mathcal{H}$ is associated with a universal kernel (e.g., Gaussian), it is large since $\mathcal{H}$ is dense in the space of continuous functions (cf. [37]). Then the induced topology (MMD) is strong enough to separate distinct probability measures. Meanwhile, RKHS allows for efficient computation using K-DRO. Therefore, our insight is that *universal RKHSs are large enough* for DRO.

We are now ready to derive the key result for the K-DRO reformulation in Theorem 3.1. Consider the inner moment problem of (2)

$$(P) := \sup_{P\in\mathcal{P},\mu\in\mathcal{C}} \int l \, dP \quad \text{subject to} \int \phi \, dP = \mu, \tag{4}$$

where we suppress θ in $l(\theta,\cdot)$ as we fix it for the moment. (4) generalizes the *(generalized) moment problem* in the sense that the constraint can be viewed as infinite-order moment constraints. Using conic duality, we obtain the strong duality of the inner moment problem.

Proposition 3.2 (Strong dual to (4))**.** *Under Assumption 3.1,* (4) *is equivalent to solving*

$$(D) := \min_{f_0\in\mathbb{R}, f\in\mathcal{H}} \delta^*_{\mathcal{C}}(f) + f_0 \quad \text{subject to } l(\xi) \le f_0 + f(\xi), \ \forall \xi \in \mathcal{X} \tag{5}$$

where $\delta^*_{\mathcal{C}}$ is the support function of $\mathcal{C}$, i.e., strong duality *holds* $(P) = (D)$.

Using Proposition 3.2, we can reformulate the inner moment problem in (2) to obtain Theorem 3.1.

Analyzing the properties of K-DRO. We now establish some theoretical results for K-DRO that will help us derive a tractable program. By the weak duality $(P) \le (D)$ of (4) and (5), we have $\int l \, dP \le f_0 + \delta^*_{\mathcal{C}}(f)$. Specifically, if $\mathcal{C}$ is the RKHS norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon, \hat{P} = \sum_{i=1}^N \frac{1}{N}\delta_{\xi_i}\}$, this inequality becomes $\int l \, dP \le f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$. This can be viewed as a generalization bound — learning on $\{\xi_i\}_{i=1}^N$ and generalizing to P. Using the *complementarity condition* (Section A.4), we can further extract the following insights.

Proposition 3.3 (Interpolation property)**.** *Let P^*, f^*, f_0^* be the optimal solutions to the primal and dual moment problem* (4) *and* (5)*. Then, $l(\theta,\xi) = f_0^* + f^*(\xi)$ holds $P^* -$ almost everywhere.*

Intuitively, this result states that $f_0^* + f^*$ interpolates the loss $l(\theta,\cdot)$ at the support points of P^*. This is illustrated in Figure 1 (b) and later empirically validated in Figure 3. We can also see that the size of RKHS $\mathcal{H}$ matters since, if $\mathcal{H}$ is small (e.g., $\mathcal{H} = \{0\}$), $f_0^* + f^*$ cannot interpolate the loss l well. The discussion so far motivates the following DRO version of the RKHS representer theorem [30].

Proposition 3.4 (Robust representer theorem)**.** *Given data $\{\xi_i\}_{i=1}^N$ and the uncertainty set chosen to be the RKHS norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon\}$. Suppose there exists a worst-case distribution P^* of* (4) *supported on $\{\zeta_j\}_{j=1}^M$. Then, in* (5)*, it suffices to consider the RKHS function of the form $f_s(\cdot) = \sum_{i=1}^N \beta_i k(\xi_i,\cdot) + \sum_{j=1}^M \gamma_j k(\zeta_j,\cdot)$ for some $\beta_i, \gamma_i \in \mathbb{R}, i = 1...N$.*

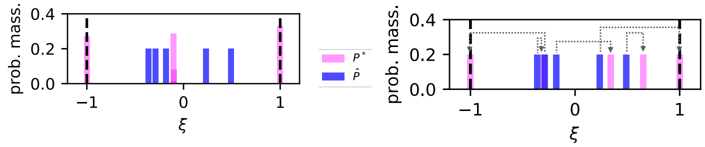
Proposition 3.4 states that the expansion points of the RKHS representer for K-DRO are exactly the support of the empirical and worst-case distributions. It extends the classical RKHS representer theorem [30], which uses the empirical samples as expansion points. The implication is that, to be distributionally robust, we should choose the representers as in Proposition 3.4 instead of only using empirical samples. This result also later helps us derive the tractable program for K-DRO. Our result is also consistent with [11], whose insight is that regularized kernel methods are not robust to heavy-tail contamination. To see this, we consider the K-DRO variant that relaxes the infinite constraint in (3) to only hold for $\xi \in \{\xi_i\}_{i=1}^N$,

$$\min_{\theta, f\in\mathcal{H}, f_0\in\mathbb{R}} f_0 + \delta^*_{\mathcal{C}}(f) \quad \text{subject to}\, l(\theta,\xi_i) \le f_0 + f(\xi_i), \ i = 1\ldots N. \tag{6}$$

Example 3.3 (Counterexample: (6) is not distributionally robust)**.** Let $\mathcal{H}$ be a Gaussian RKHS with the bandwidth $\sigma = \sqrt{2}$. (See Figure 1c for the associated metric.) Suppose our data set is $\{0\}$ and the uncertainty set is $\mathcal{C} := \{\mu \colon \|\mu - \phi(0)\|_{\mathcal{H}} \le \epsilon\}$. Suppose $\epsilon = \sqrt{2 - 2/e}$. We consider the loss function $l(\xi) = [|\theta + \xi| - 1]_+$. Under this setting, (6) reads

$$(d) := \min_{\theta, f\in\mathcal{H}, f_0\in\mathbb{R}} f_0 + f(0) + \epsilon\|f\|_{\mathcal{H}} \quad \text{subject to } [|\theta| - 1]_+ \le f_0 + f(0)$$

which admits an optimal solution $\theta^* = 0, f^* = 0, f_0^* = 0$ and the worst-case risk $(d) = 0$. However, let $P' = \frac{1}{2}\phi(0) + \frac{1}{2}\phi(2)$. It is straightforward to verify $P' \in \mathcal{C}, \int l(\theta^*,\xi) \, dP'(\xi) = \frac{1}{2} > (d)$, i.e., the solution θ^* is not robust against P'.



(a) Support sampling

(b) Perturbation

Figure 2: Computing the worst-case distribution P^* using **(a)**: (7) samples possible support ζ_j then optimizes w.r.t. weights α. **(b)**: (8) moves the empirical samples directly.

Remark. Nonetheless, (6) can be used as a finite-sample approximation to K-DRO (3) in practice. For Gaussian RKHS, $\forall p, q, \|\mu_p - \mu_q\|_{\mathcal{H}} \leq \|\mu_p\|_{\mathcal{H}} + \|\mu_q\|_{\mathcal{H}} \leq 2\sup_{x,y}\sqrt{k(x,y)} = 2$. Hence, if $\epsilon \geq 2$, $\mathcal{C}$ contains all probability distributions. Then, K-DRO is reduced to worst-case RO.

3.1 Worst-case distribution

Thus far, we have proposed K-DRO for making the decision θ via reformulation (3). In practice, it is often useful to find the worst-case distribution P^* (e.g., to study adversarial examples). We now propose two practical methods to compute P^* for a given θ, based on *support sampling* and *perturbation*, respectively. We illustrate the ideas in Figure 2. See the appendix for more results.

Support sampling. We consider the moment problem (4) where the distribution is restricted to discrete distributions supported on some sampled support $\{\zeta_j\}_{j=1}^M \subseteq \mathcal{X}$. [2] For any given θ,

$$\max_{\alpha \in S_M} \sum_{j=1}^M \alpha_i l(\theta, \zeta_j) \quad \text{subject to} \quad \left\| \sum_{j=1}^M \alpha_i \phi(\zeta_j) - \frac{1}{N}\sum_{i=1}^N \phi(\xi_i) \right\|_{\mathcal{H}} \leq \epsilon. \tag{7}$$

(7) can be written as a quadratically constrained program with linear objective, which admits a (strong) semidefinite program dual via what is historically known as the S-lemma [25] (cf. appendix). Alternatively, (7) can be directly handled by convex solvers for a given θ.

Perturbation. Alternatively, we search for worst-case distributions that are *perturbations* of the empirical distribution. Let $d_i \in \mathcal{X}$ be some perturbation vector, given θ,

$$\max_{\substack{d_i, i=1\ldots N,\\ \xi_i + d_i \in \mathcal{X}}} \quad \frac{1}{N}\sum_{i=1}^N l(\theta, \xi_i + d_i) \quad \text{subject to} \quad \left\| \frac{1}{N}\sum_{i=1}^N \left(\phi(\xi_i + d_i) - \phi(\xi_i)\right) \right\|_{\mathcal{H}} \leq \epsilon. \tag{8}$$

Compared with (7), (8) directly searches for the support of the worst-case distribution. It can be interpreted as transporting the probability mass from empirical samples ξ_i to form the worst-case distribution. Depending on the kernel used, (8) may become a nonlinear program. However, since (8) can always be initialized with a feasible solution $d_i = 0$, it can be handled by off-the-shelf solvers.

3.2 Convex programs for solving K-DRO

Finally, we derive the tractable K-DRO formulation for practical use. Since RKHS functions are not polynomial-representable in general, we can not use semi-infinite programming methods such as sum-of-squares. In the following, we consider a sample-based convex approximation to (3) (cf. [10],[34, Ch. 5]) by relaxing the infinite constraint. Given sampled support $\{\zeta_j\}_{j=1}^M \subseteq \mathcal{X}$ [2] , we solve

$$\min_{\theta, f_0 \in \mathbb{R}, f \in \mathcal{H}} \delta^*_{\mathcal{C}}(f) + f_0 \quad \text{subject to } l(\theta, \zeta_j) \leq f(\zeta_j) + f_0, \ j = 1 \ldots M \tag{9}$$

If $\{\zeta_j\}_{j=1}^M = \{\xi_i\}_{i=1}^N$, then (9) is reduced to the relaxed variant (6). (9) searches for worst-case distributions whose support is a subset of $\{\zeta_j\}_{j=1}^M$. Specifically, suppose the uncertainty set is the

[2] Note the sampled support $\{\zeta_j\}_{j=1}^M$ need not be real data; they are only the candidates for the worst-case support. The purpose is to make the the semi-infinite constraint approximately satisfied. See the appendix for more details.

RKHS norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon\}$. By the robust representer theorem Proposition 3.4, it suffices to solve

$$\min_{\theta,\beta,\gamma,f_0} f_0 + \frac{1}{N}\sum_{i=1}^{N} f_s(\xi_i) + \epsilon\|f_s\|_{\mathcal{H}} \quad \text{subject to } l(\theta,\zeta_j) \le f_0 + f_s(\zeta_j),\ j = 1\ldots M, \tag{10}$$

where the plug-in form is $f_s(\cdot) = \sum_{i=1}^{N}\beta_i k(\xi_i,\cdot) + \sum_{j=1}^{M}\gamma_j k(\zeta_j,\cdot)$, $\|f_s\|_{\mathcal{H}} = \sqrt{\alpha^\top K \alpha}$, where $\alpha = (\beta_1,\ldots,\beta_N,\gamma_1,\ldots,\gamma_M)^\top$, $K = [k(\eta_i,\eta_j)]$, $\eta = (\xi_1,\ldots,\xi_N,\zeta_1,\ldots,\zeta_M)^\top$. As an alternative, we can derive the following soft version of (10) as an unconstrained program using the conditional value-at-risk. Given α (e.g., $\alpha = 0.01$),

$$\min_{\theta,\beta,\gamma,f_0} \frac{1}{\alpha M}\sum_{j=1}^{M}[l(\theta,\zeta_j) - f_s(\zeta_j) - f_0]_+ + f_0 + \frac{1}{N}\sum_{i=1}^{N} f_s(\xi_i) + \epsilon\|f_s\|_{\mathcal{H}}. \tag{11}$$

Both practical formulations (10) and (11) are *convex programs* if $l(\theta,\xi)$ in convex in θ.

4 Numerical studies

This section demonstrates the theoretical insights of K-DRO in action. It is not a benchmark of state-of-art performances. See the appendix for more results. The code will be available online

4.1 Distributionally robust solution to uncertain least squares

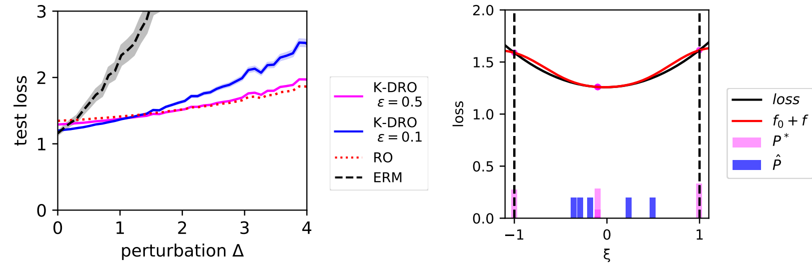
We first consider a robust least squares problem adapted from [17], which demonstrated an important application of RO to statistical learning historically. (See also [9, Ch. 6.4].) The task is to minimize the objective $\|A\theta - b\|_2^2$ w.r.t. θ. A is modeled by $A(\xi) = A_0 + \xi A_1$, where $\xi \in \mathcal{X}$ is uncertain, $\mathcal{X} = [-1,1]$, and $A_0, A_1 \in \mathbb{R}^{10\times 10}, b \in \mathbb{R}^{10}$ are given. We compare K-DRO against using *(a)* empirical risk minimization (ERM; also known as sample average approximation) that minimizes $\frac{1}{N}\sum_{i=1}^{N}\|A(\xi_i)\,\theta - b\|_2^2$, *(b)* worst-case RO via SDP from [17]. We consider a data-driven setting with given samples $\{\xi_i\}_{i=1}^N$. We formulate the K-DRO problem as $\min_\theta \max_{P\in\mathcal{P},\mu\in\mathcal{C}} \mathbb{E}_{\xi\sim P}\,\|A(\xi)\,\theta - b\|_2^2 \quad \text{subject to} \int \phi\, dP = \mu$, where we choose the uncertainty set to be $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon\}$, where $\mu_{\hat{P}} = \sum_{i=1}^{N}\frac{1}{N}\phi(\xi_i)$.

Empirical samples $\{\xi_i\}_{i=1}^N (N = 10)$ are generated uniformly from $[-0.5, 0.5]$. We then apply K-DRO formulation (10). To test the solution, we create a distribution shift by generating test samples from $[-0.5\cdot(1+\Delta), 0.5\cdot(1+\Delta)]$, where Δ is a perturbation varying within $[0,4]$. Figure 3a shows this comparison. As the perturbation increases, ERM quickly lost robustness. On the other hand, RO is the most robust with the trade-off of being conservative. As expected, K-DRO achieves some level of optimality while retaining robustness. We then ran K-DRO with fewer empirical samples ($N = 5$) to show the geometric interpretations. We plot the optimal dual solution $f_0^* + f^*$ in Figure 3b. Recall it is an over-estimator of the loss $l(\theta,\cdot)$. We solve (7) to obtain a worst-case distribution P^*. Comparing P^* with $\hat{P}$, we can observe the adversarial behavior of the worst-case distribution. See the caption for more description.

4.2 Distributionally robust classification

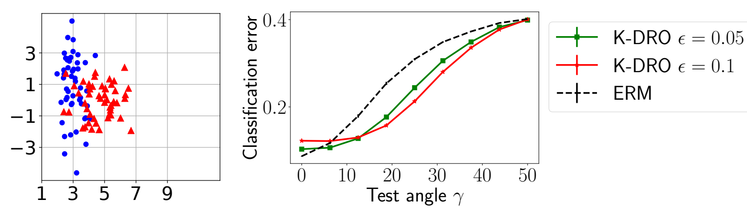
We now show how kernel-DRO can be applied to train a classification model $g_\theta \colon x \mapsto y$. We consider a two-dimensional ($x \in \mathbb{R}^2$), two-class classification problem ($y \in \{-1,1\}$). Samples from class 1 (red) are drawn from $p(x|y=1) = \mathcal{N}((5,0)^\top, I)$, while that from class -1 (blue) are generated from $\mathcal{N}\left((3,1)^\top, \mathrm{diag}(1/2,2)\right)$. The class prior probability is uniform i.e., $p(y=-1) = p(y=1) = 1/2$. The training samples are shown in Figure 4a. The model is trained by solving K-DRO (10), where $\xi_i := [x_i, y_i]$, with the hinge loss $l(\theta,\xi) := \max(0, 1 - g_\theta(x)y)$. We use a product kernel of the form $k((x,y),(x',y')) = k_X(x,x')k_Y(y,y')$, where both k_X, k_Y are Gaussian kernels. For simplicity, we use a linear classifier $g_\theta(x) := \mathrm{sign}(m^\top x + c)$ where $\theta := (m,c)$.

During test time, we create a distribution mismatch by rotating the class -1 data (blue) γ degrees clockwise about the point $(3,-1)$ (γ varies from 0 to 50 degrees). That is, $p(x|y=-1,\gamma) = \mathcal{N}\left(R_\gamma(3,1)^\top, R_\gamma \mathrm{diag}(1/2,2)R_\gamma^\top\right)$, where R_γ is a linear transformation performing the rotation.



(a) Robust least squares loss (b) Geometric interpretation

Figure 3: **(a)** This plot depicts the test loss of algorithms. All error bars are in standard error. We ran 10 independent trials. In each trial, we solved K-DRO to obtain θ^* and tested it on a test dataset of 500 samples. We then vary the perturbation Δ from 0 to 4. **(b)** (red) is the dual optimal solution $f_0^* + f^*$. (black) is the function $l(\theta^*, \cdot)$. The pink bars depict a worst-case distribution while the blue bars the empirical distribution. We can observe that $f_0^* + f^*$ touches loss $l(\theta^*, \cdot)$ at the support of the worst-case distribution P^* (pink dots). Note f^* (normalized) can be viewed as a witness function of the two distributions.



(a) Training samples (b) Classification error

Figure 4: Distributionally robust classification (Section 4.2). **(a)**: $N = 100$ training samples drawn from $p(x, y|\gamma = 0)$. **(b)**: average classification errors over 50 trials of a linear model trained with kernel DRO (K-DRO), and with the standard ERM of hinge loss without robustness guarantee.

The classification errors, averaged over 50 independent trials, are presented in Figure 4b. A test set of 50000 points is redrawn in each trial. We observe that the models learned with K-DRO provide robustness against changes of the angle at test time. In comparison, ERM achieves a lower error rate when there is no test-time rotation. However, it quickly worsens as the distribution shift takes place.

5 Other related work

This paper uses similar techniques of reformulating min-max programs as in data-driven RO, e.g., [3, 6], but our uncertainty set is constructed in an RKHS. Compared with the popular Wasserstein distance, related works using RKHS distances are less common. The authors of [16] proposed variational approximations to marginal DRO to treat covariate shift in supervised learning. One of their practical algorithms is similar (but not identical) to the soft version (11) of K-DRO. The authors of [44] used kernel mean embedding for the inner moment problem, which is equivalent to our formulation (7). The work of [38] used insights from DRO to motivate a regularizer for kernel ridge regression but did not contain reformulations that can solve DRO. To the best of our knowledge, K-DRO in this paper is the first tractable reformulation of DRO using RKHS distances.

6 Discussion

Our K-DRO reformulation (3) can be generalized to the larger class of integral probability metric. Suppose $d_{\mathcal{F}}$ is the IPM defined by some class of functions $\mathcal{F}$. Then, we can reformulate IPM-DRO $\min_\theta \sup_{d_{\mathcal{F}}(P,\hat{P}) \le \epsilon} \int l(\theta, \xi)\, dP(\xi)$ using the same technique in Proposition 3.2. (See Section A.9.) If we choose the class $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$, we recover K-DRO. Similarly, $\mathcal{F} = \{f : \mathrm{lip}(f) \le 1\}$ recovers the (type-1) Wasserstein-DRO. This puts Wasserstein-DRO and K-DRO into a unified perspective.

The compactness assumption on $\mathcal{X}$ may be further extended, just like universality can be extended to non-compact domains [37]. Choosing ϵ in this paper can be further motivated using kernel statistical testing [19] and domain adaptation. All the programs in this paper are solved using off-the-shelf solvers. A future direction is to explore tailored numerical methods for K-DRO, as well as alternative practical approaches to the support sampling method in Section 3.2.

Acknowledgments and Disclosure of Funding

The authors would like to thank Daniel Kuhn for the helpful discussion during a workshop at IPAM, UCLA. We also thank Yassine Nemmour and Simon Buchholz for sending us their feedback on the paper draft. During part of this project, Jia-Jie Zhu was supported by the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 798321. Moritz Diehl would like to acknowledge the funding support from the DFG via Project DI 905/3-1 (on robust model predictive control).

References

[1] Alexander Barvinok. *A Course in Convexity*, volume 54. American Mathematical Soc., 2002.

[2] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, February 2013.

[3] Aharon Ben-Tal, Dick den Hertog, and Jean-Philippe Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1):265–299, February 2015.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

[5] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[6] Dimitris Bertsimas, Nathan Kallus, and Vishal Gupta. *Data-Driven Robust Optimization*. Springer Berlin Heidelberg, 2017.

[7] Dimitris Bertsimas and Ioana Popescu. Optimal Inequalities in Probability Theory: A Convex Optimization Approach. *SIAM Journal on Optimization*, 15(3):780–804, January 2005.

[8] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein Profile Inference and Applications to Machine Learning. *Journal of Applied Probability*, 56(03):830–857, September 2019.

[9] Stephen Boyd, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[10] G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, May 2006.

[11] Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, August 2007.

[12] John B Conway. *A course in functional analysis*, volume 96. Springer, 2019.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[14] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58(3):595–612, June 2010.

[15] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[16] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses Against Mixture Covariate Shifts. page 39.

[17] Laurent El Ghaoui and Hervé Lebret. Robust Solutions to Least-Squares Problems with Uncertain Data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, October 1997.

[18] Rui Gao and Anton J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv:1604.02199 [math]*, July 2016.

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

[20] Keiiti Isii. On sharpness of tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, December 1962.

[21] Garud N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[22] Jean B. Lasserre. Bounds on measures satisfying moment conditions. *The Annals of Applied Probability*, 12(3):1114–1137, 2002.

[23] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, September 2018.

[24] Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, October 2005.

[25] Imre Pólik and Tamás Terlaky. A Survey of the S-Lemma. *SIAM Review*, 49(3):371–418, January 2007.

[26] Ioana Popescu. A Semidefinite Programming Approach to Optimal-Moment Bounds for Convex Classes of Distributions. *Mathematics of Operations Research*, 30(3):632–657, August 2005.

[27] R. Tyrrell Rockafellar. *Convex Analysis*. Number 28. Princeton university press, 1970.

[28] W W Rogosinski. Moments of Non-Negative Mass. page 28.

[29] Herbert Scarf. A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 1958.

[30] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and R. Williamson, editors, *Annual Conference on Computational Learning Theory*, number 2111 in Lecture Notes in Computer Science, pages 416–426, Berlin, 2001. Springer.

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

[32] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. page 68.

[33] Alexander Shapiro. On Duality Theory of Conic Linear Problems. In Panos Pardalos, Miguel Á. Goberna, and Marco A. López, editors, *Semi-Infinite Programming*, volume 57, pages 135–165. Springer US, Boston, MA, 2001.

[34] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

[35] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[36] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[37] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[38] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141, 2019.

[39] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[40] Bart P. G. Van Parys, Paul J. Goulart, and Daniel Kuhn. Generalized Gauss inequalities via semidefinite programming. *Mathematical Programming*, 156(1-2):271–302, March 2016.

[41] Lieven. Vandenberghe, Stephen. Boyd, and Katherine. Comanor. Generalized Chebyshev Bounds via Semidefinite Programming. *SIAM Review*, 49(1):52–64, January 2007.

[42] Zizhuo Wang, Peter W. Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, April 2016.

[43] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, March 2018.

[44] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Worst-Case Risk Quantification under Distributional Ambiguity using Kernel Mean Embedding in Moment Problem. *arXiv:2004.00166 [cs, eess, math]*, March 2020.

[45] Steve Zymler, Daniel Kuhn, and Berç Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, February 2013.

Appendix

A Proofs of theoretical results

Table 2 provides an overview to help readers navigate the theoretical results in this paper.

Table 2: List of the theoretical results in this paper

Theoretical results	K-DRO reformulation Theorem 3.1 Strong duality of the inner moment problem Proposition 3.2 Interpolation property Proposition 3.3 Complementarity condition Lemma A.1 Robust representer theorem Proposition 3.4 IPM-DRO reformulation Section A.9
Formulations	K-DRO (2), reformulation (3) Formulations for various RKHS uncertainty sets Table 1, 3 Program to compute worst-case distributions (7),(8),(20) Practical convex program to solve K-DRO (10), (11) (see also (6)) IPM-DRO reformulation (22)

In general, we refer to standard texts in optimization [9, 34, 4], convex analysis [27, 1], and functional analysis [12] for more mathematical background.

Notation. In the proofs, we use $\mathcal{M}$ to denote the space of signed measures on $\mathcal{X}$. The dual cone of a set of signed measures $\mathcal{K} \subseteq \mathcal{M}$ is defined as $\mathcal{K}^* := \{h\colon \int h dm \geq 0, \forall m \in \mathcal{K}, h \text{ measurable}\}$. Using the reproducing property, we have the identity $\int f\, dP = \langle f, \mu_P \rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$ and $P \in \mathcal{P}$, which we will frequently use in the proofs.

A.1 Proof of Proposition 3.2 strong duality of the inner moment problem in K-DRO

We prove the duality result for the inner moment problem in Proposition 3.2. We first derive the weak dual and then prove the strong duality.

Proof. We first relax the constraint $P \in \mathcal{P}$ to its conic hull $P \in \mathrm{co}(\mathcal{P})$. To constrain P to still be a probability measure, we impose $\int 1\, dP(x) = 1$, which results in the primal problem equivalent to (4)

$$(P) := \max_{P \in \mathrm{co}(\mathcal{K}), \mu \in \mathcal{C}} \int l\, dP \quad \text{subject to } \int \phi\, dP = \mu,\ \int 1\, dP = 1.$$

We construct the Lagrangian relaxation by associating the constraints with the dual variables $f \in \mathcal{H}, f_0 \in \mathbb{R}$, as well as adding the indicator function of $\mathcal{C}$. Note both sides of the constraint

$\int \phi \, dP = \mu$ are functions in $\mathcal{H}$, hence the multiplier f is an RKHS function.

$$
\begin{aligned}
\mathcal{L}(P, \mu;\ f, f_0) &= \int l \, dP - \delta_{\mathcal{C}}(\mu) + \langle \mu - \int \phi \, dP, f \rangle_{\mathcal{H}} + f_0 (1 - \int 1 \, dP) \\
&= \int l \, dP - \delta_{\mathcal{C}}(\mu) + \langle \mu, f \rangle_{\mathcal{H}} - \int f \, dP + f_0 - \int f_0 \, dP \\
&= \int l - f - f_0 \, dP + (\langle \mu, f \rangle_{\mathcal{H}} - \delta_{\mathcal{C}}(\mu)) + f_0. \qquad (12)
\end{aligned}
$$

The second equality is due to the reproducing property of RKHS. The dual function is given by

$$
g(f, f_0) = \sup_{P, \mu} \int l - f - f_0 \, dP + (\langle \mu, f \rangle_{\mathcal{H}} - \delta_{\mathcal{C}}(\mu)) + f_0.
$$

The first term is bounded above by 0 iff $l - f - f_0 \in -K^*$. By Lemma C.1, this conic constraint is equivalent to the constraint of (5), $l(\xi) \leq f_0 + f(\xi),\ \forall \xi \in \mathcal{X}$.

Finally, expressing the second term using convex conjugate $\delta^*_{\mathcal{C}}(f) = \sup_\mu \langle \mu, f \rangle_{\mathcal{H}} - \delta_{\mathcal{C}}(\mu)$ concludes the derivation. $\square$

Next, we prove strong duality.

A.1.1 Proof of strong duality

Strong duality can potentially be adapted from the strong duality result of moment problem, e.g., [33]. However, we give a self-contained proof with only elementary mathematics that sheds light on the connection between the RKHS theory and distributionally robust optimization. The proof is a generalization of the Euclidean space conic duality theorem ([4] Theorem A.2.1) to infinite dimensions. Figure 5 illustrates the idea of the proof.

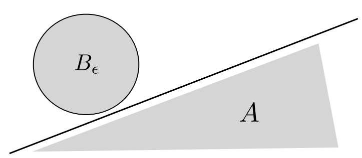


Figure 5: Illustration of the strong duality proof that uses a separating hyperplane. See the proof for detailed descriptions.

Proof. We assume the dual optimal value of (5) is finite $(D) < \infty$. Since the converse means that the dual problem is infeasible, which implies that the primal problem is unbounded. Due to the upper semicontinuity of l in Assumption 3.1, this can not happen on a compact $\mathcal{X}$.

Let us consider the Hilbert space $\mathbb{R} \times \mathcal{H} \times \mathbb{R}$ equipped with the inner product $\langle , \rangle_{\mathbb{R}} + \langle , \rangle_{\mathcal{H}} + \langle , \rangle_{\mathbb{R}}$. We construct a cone in $\mathbb{R} \times \mathcal{H} \times \mathbb{R}$

$$
A = \left\{ \left(\int 1 \, dP, \int \phi dP, \int l \, dP \right) : P \in \mathrm{co}(\mathcal{P}) \right\},
$$

where co again denotes conic hull.

Let $t = (P)$ denote the optimal primal value. $\forall \epsilon > 0$, we construct the set

$$
B_\epsilon = \left\{ (1, \mu, t + \epsilon) : \mu \in \mathcal{C} \right\},
$$

which is a closed convex set with non-empty relative interior by Assumption 3.1 (i.e., Slater condition is satisfied).

It is straightforward to verify that those two sets do not intersect. Suppose $x = (x_1, x_2, x_3) \in A \cap B_\epsilon$, this means $\exists \mu', P'$ such that $x_1 = 1 = \int 1 \, dP', x_2 = \mu' = \int \phi dP'$, i.e., μ', P' is a primal feasible solution. Then the third coordinate of x satisfies $x_3 = \int l \, dP' \leq (P) < t + \epsilon = x_3$, which is impossible. Hence, $A \cap B_\epsilon = \emptyset$.

In the rest of the proof, we will show that, $\forall \epsilon > 0$, the dual optimal value (D) satisfies

$$(D) \leq (P) + \epsilon.$$

Combining this with weak duality $(D) \geq (P)$ will result in strong duality. We now justify this inequality.

By the separation theorem, (see, e.g., [1] Theorem III.3.2, 3.4), there exists a closed hyperplane that strictly separates A and B_ϵ. The separation is strict because $t + \epsilon > t = \int l \, dP$. By the Riesz representation theorem, $\exists (f_0, f, \tau) \in \mathbb{R} \times \mathcal{H} \times \mathbb{R}, s \in \mathbb{R}$, such that

$$f_0 + \langle f, \mu \rangle_{\mathcal{H}} + \tau(t + \epsilon) < s,$$

$$f_0 \int 1 \, dP + \langle f, \int \phi dP \rangle_{\mathcal{H}} + \tau \int l \, dP > s.$$

Plugging in $P = 0$, we obtain $s < 0$. Since P lives in a cone, the left-hand side of the second inequality must be non-negative. Otherwise, we can scale P so that the separation will fail. In summary, we have

$$\begin{aligned} &f_0 + \langle f, \mu \rangle_{\mathcal{H}} + \tau(t + \epsilon) < 0, \forall \mu \in \mathcal{C}, \\ &f_0 \int 1 \, dP + \langle f, \int \phi dP \rangle_{\mathcal{H}} + \tau \int l \, dP \geq 0, \forall P \in \operatorname{co}(\mathcal{P}). \end{aligned} \tag{13}$$

By Assumption 3.1 (Slater condition), the primal problem has a non-empty solution set. Because l is proper and upper semi-continuous and the feasible solution set for the optimization problem is compact (see Section C) , the primal optimum is attained by the extreme value theorem. Suppose P^* is a primal optimal solution, from the second inequality of (13),

$$f_0 \int 1 \, dP^* + \langle f, \int \phi dP^* \rangle_{\mathcal{H}} + \tau \int l \, dP^* = f_0 + \langle f, \mu_{P^*} \rangle_{\mathcal{H}} + \tau t \geq 0.$$

Using this and the first inequality of (13), we obtain $\tau < 0$. Without loss of generality, we let $\tau = -1$.

From the second inequality of (13), we have

$$f_0 \int 1 \, dP + \langle f, \int \phi dP \rangle_{\mathcal{H}} - \int l \, dP = \int f_0 + f - l \, dP \geq 0, \forall P \in \operatorname{co}(\mathcal{P}).$$

This tells us that f_0, f is a feasible dual solution because it satisfies the semi-infinite constraint in (5).

By the first inequality of (13),

$$f_0 + \sup_{\mu \in \mathcal{C}} \langle f, \mu \rangle_{\mathcal{H}} \leq t + \epsilon, \forall \epsilon > 0,$$

where the left-hand side is precisely the dual objective in (5). This implies $(D) \leq (P) + \epsilon$. By weak duality, $(D) \geq (P)$. Therefore, strong duality holds. $\square$

This proof gives us the third interpretation of the dual variables f_0, f — they define a separating hyperplane of A and B_ϵ.

Remark. From the proof, we see that the Slater condition in Assumption 3.1 is stronger than needed be. If $\mathcal{C}$ is singleton, we can still find a convex neighborhood W_ϵ of the singleton B_ϵ since $\mathbb{R} \times \mathcal{H} \times \mathbb{R}$ is locally convex. Then W_ϵ and A can still be strictly separated using the same technique in the proof. Hence strong duality still holds when $\mathcal{C}$ is a singleton.

A.2 Proof of Theorem 3.1 K-DRO reformulation

Proof. Theorem (3.1) is obtained by reformulating the inner moment problem in (2) using the strong duality result in Proposition 3.2, i.e.,

$$\begin{aligned} &\min_\theta \sup_{P,\mu} \left\{ \int l(\theta, \xi) \, dP(\xi) \colon \int \phi \, dP = \mu, P \in \mathcal{P}, \mu \in \mathcal{C} \right\} \\ &\qquad = \min_\theta \min_{f_0 \in \mathbb{R}, f \in \mathcal{H}} \left\{ f_0 + \delta^*_{\mathcal{C}}(f) : l(\theta, \xi) \leq f_0 + f(\xi), \ \forall \xi \in \mathcal{X} \right\}, \end{aligned} \tag{14}$$

Table 3: Robust counterpart formulations of K-DRO.

RKHS uncertainty set $\mathcal{C}$	Robust counterpart formulation
norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \leq \epsilon\}$	$f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$
convex hull $\mathcal{C} = \operatorname{conv}\{\mathcal{C}_1, \dots, \mathcal{C}_N\}$ (same under closure $\operatorname{clconv}\{\mathcal{C}\}$)	$f_0 + \max_i \delta^*_{\mathcal{C}_i}(f)$
example: polytope $\mathcal{C} = \operatorname{conv}\{\phi(\xi_1), \dots, \phi(\xi_N)\}$	$f_0 + \max_i f(\xi_i)$ (equivalent to the scenario approach in [10])
Minkowski sum $\sum_{i=1}^N C_i$	$f_0 + \sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f)$
example: $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ $\mathcal{C}_1 = \{\mu \colon \|\mu\|_{\mathcal{H}} \leq \epsilon\}$ $\mathcal{C}_2 = \operatorname{conv}\{\phi(\xi_1), \dots, \phi(\xi_N)\}$	$f_0 + \max_i f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$
affine combination $\mathcal{C} = \sum_{i=1}^N \alpha_i \mathcal{C}_i, \sum_{i=1}^N \alpha_i = 1$	$f_0 + \sum_{i=1}^N \alpha_i \delta^*_{\mathcal{C}_i}(f)$
example: data contamination $\mathcal{C} = \{\alpha\mu_{\hat{P}} + (1-\alpha)\mu_Q : \mu_Q \in \mathcal{C}_Q\}$	$f_0 + \frac{\alpha}{N}\sum_{i=1}^N f(\xi_i) + (1-\alpha)\delta^*_{\mathcal{C}_Q}(f)$
multiple kernels $\mathcal{C}_i \subseteq \mathcal{H}_i$	$f_0 + \sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f_i)$ where $f_i \in \mathcal{H}$
example: $\mathcal{C}_i = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \leq \epsilon_i\}$	$f_0 + \frac{1}{N}\sum_{i=1}^N \sum_{i=j}^N f_i(\xi_j) + \epsilon \sum_{i=1}^N \|f_i\|_{\mathcal{H}_i}$
singleton $\mathcal{C} = \left\{\sum_{i=1}^N \frac{1}{N}\phi(\xi_i)\right\}$	$f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i)$ (equivalent to ERM/SAA)
entire RKHS $\mathcal{C} = \mathcal{H}$	$f_0 + \delta_0(f)$ (equivalent to worst-case RO [4])

which results in formulation (3). □

A.3 Table 3 deriving formulations for various choices of RKHS uncertainty set $\mathcal{C}$

We now derive the formulations of support functions for various RKHS uncertainty sets in Table 3.

(RKHS norm-ball) Let us consider the uncertainty set of $\mathcal{C} = \{\mu \colon \|\mu - \hat{\mu}\|_{\mathcal{H}} \leq \epsilon\}$, where $\hat{P} = \sum_{i=1}^N \frac{1}{N}\delta_{\xi_i}$. The support function is given by

$$\delta^*_{\mathcal{C}}(f) = \sup_{\mu \in \mathcal{C}} \langle f, \mu \rangle_{\mathcal{H}} = \langle f, \hat{\mu} \rangle_{\mathcal{H}} + \sup_{\|\mu - \hat{\mu}\|_{\mathcal{H}} \leq \epsilon} \langle f, \mu - \hat{\mu} \rangle_{\mathcal{H}} = \langle f, \hat{\mu} \rangle_{\mathcal{H}} + \epsilon\|f\|_{\mathcal{H}}$$

where the last equality is by the Cauchy-Schwarz inequality, or alternatively by the self-duality of Hilbert norms. (Note we assume there exists some $\mu \in \mathcal{H}$ such that $\|\mu - \hat{\mu}\|_{\mathcal{H}} = \epsilon$.)

(Polytope, convex hull of uncertainty set) The result for convex hull follows from standard support function calculus. If the uncertainty set $\mathcal{C}$ is described by the polytope $\operatorname{conv}\{\phi(\xi_1), \dots, \phi(\xi_N)\}$, then $\delta^*_{\mathcal{C}}(f) = \max_{1 \leq i \leq N} f(\xi_i)$. Furthermore, the support function value remains the same under closure operation [3] . The equivalence to the scenario approach in [10] can be seen by noticing that $\max_{1 \leq i \leq N} l(\xi_i) \leq f_0 + \max_{1 \leq i \leq N} f(\xi_i)$. If $\mathcal{H}$ is universal, then there exists f_0, f such that the equality is attained.

(Minkowski sum, affine combination) Those cases follow directly from the support function calculus; cf. [3].

(K-DRO with multiple kernels) Let us consider multiple uncertainty sets from different RKHSs. Suppose $\mathcal{H}_1, \dots, \mathcal{H}_{N_h}$ are RKHSs associated with feature maps $\phi_1, \dots, \phi_{N_h}$. Let $\mathcal{C}_1, \dots, \mathcal{C}_{N_h}$ be

[3]The convex hull can be replaced with its closure $\operatorname{clconv}(\cdot)$. Note convex hulls in infinite-dimensional spaces are not automatically closed; cf. Krein-Milman theorem.

the uncertainty sets in the respective RKHSs. K-DRO formulation with multiple kernels is given By

$$\min_{\theta}\sup_{P,\mu}\left\{\int l(\theta,\xi)\,dP(\xi)\colon \int \phi_i\,dP=\mu_i, P\in\mathcal{P}, \mu_i\in\mathcal{C}_i, i=1\dots N_h\right\}, \tag{15}$$

Using the same proof as Proposition 3.2, we have the K-DRO reformulation

$$\min_{\theta, f_0\in\mathbb{R}, f_i\in\mathcal{H}_i} f_0+\sum_{i=1}^{N}\delta^*_{\mathcal{C}_i}(f_i) \quad \text{subject to } l(\theta,\xi)\le f_0+\sum_{i=1}^{N} f_i(\xi),\ \forall\xi\in\mathcal{X}. \tag{16}$$

Hence we obtain the formulation in Table 3.

(Singleton uncertainty set $\mathcal{C} = \left\{\sum_{i=1}^N \frac{1}{N}\phi(\xi_i)\right\}$) By the reproducing property, the support function of the singleton uncertainty set is given by $\delta^*_{\mathcal{C}}(f)=\frac{1}{N}\sum_{i=1}^N f(\xi_i)$.

(If $\mathcal{C}=\mathcal{H}$, reduction to classical RO) $\delta^*_{\mathcal{H}}(f)\neq\infty$ iff $f=0$. Then (3) is reduced to

$$\min_{\theta, f_0\in\mathbb{R}} f_0 \quad \text{subject to } l(\theta,\xi)\le f_0,\ \forall\xi\in\mathcal{X} \tag{17}$$

which is the epigraphic form of the worst-case RO. [2]

A.4 Complementarity condition and proof

Lemma A.1 (Complementarity condition)**.** *Let P^*, f^*, f_0^* be a set of optimal primal-dual solutions of (P) and (D), then*

$$\int l-f^*-f_0^*\,dP^*=0,\quad \delta^*_{\mathcal{C}}(f^*)=\int f^*\,dP^* \tag{18}$$

If $\mathcal{C}=\{\mu\colon \|\mu-\mu_{\hat{P}}\|_{\mathcal{H}}\le\epsilon\}$, the second equality implies

$$\int \frac{f^*}{\|f^*\|_{\mathcal{H}}}\,d(P^*-\hat{P})=\mathrm{MMD}(P^*,\hat{P}), \tag{19}$$

which gives a second interpretation of the dual solution f^* as a witness function.

It is well known that complementarity condition holds iff strong duality holds in the moment problem; cf. [33]. The following is a straightforward proof.

Proof. Plug P^*, f^*, f_0^* into Lagrangian (12),

$$\int l\,dP^*\le\int l-f-f_0\,dP^*+\delta^*_{\mathcal{C}}(f^*)+f_0^*\le\delta^*_{\mathcal{C}}(f^*)+f_0^*.$$

By strong duality, all inequalities above are equalities. Therefore, the first equality gives the condition $\delta^*_{\mathcal{C}}(f^*)=\int f^*\,dP^*$ while the second yields $\int l-f^*-f_0^*\,dP^*=0$. □

A.5 Proof of Proposition 3.3 (Interpolation property)

Proof. Since f_0^*, f^* is a solution to the inner moment problem of K-DRO (5), we have $l(\theta,\xi)\le f_0^*+f^*(\xi)$, $\forall\xi\in\mathcal{X}$ for any given θ. By the first equation in the complementarity condition (18), we have $\int l-f^*-f_0^*\,dP^*=0$. Hence the integrand must be zero P^*-a.e. □

[2]Note that $\mathcal{C}=\mathcal{H}$ is no longer closed. However, the resulting ambiguity set becomes $\mathcal{P}$, which is still compact if $\mathcal{X}$ is compact.

A.6 Proof of Proposition 3.4 (Robust representer)

Proof. Let $\mu_{P^*} = \sum_{i=1}^M \alpha_i \phi(\zeta_i)$, be the embedding of the worst-case distribution, for some $\alpha \in S_N$. From (19) and the reproducing property, we have

$$\|\mu_{P^*} - \mu_{\hat{P}}\|_{\mathcal{H}} = \langle \frac{f^*}{\|f^*\|_{\mathcal{H}}}, \mu_{P^*} - \mu_{\hat{P}} \rangle_{\mathcal{H}}.$$

By the Cauchy-Schwarz inequality, we also have $\langle \frac{f^*}{\|f^*\|_{\mathcal{H}}}, \mu_{P^*} - \mu_{\hat{P}} \rangle_{\mathcal{H}} \leq \|\mu_{P^*} - \mu_{\hat{P}}\|_{\mathcal{H}}$. The equality is attained if $f^* = \frac{\|f^*\|_{\mathcal{H}}}{\|\mu_{P^*} - \mu_{\hat{P}}\|_{\mathcal{H}}}(\mu_{P^*} - \mu_{\hat{P}})$. Since both terms on the right-hand side of this equation admit finite-term expansion, so does f^*. The conclusion follows. □

Remark. In data-driven Wasserstein-DRO, it is known that there always exists worst-case a distribution supported on $N + 1$ points [18]. In this paper, we do not prove the existence of such minimal representations.

A.7 SDP dual via S-lemma

We consider a discretized version of the primal moment problem in (7) where the distribution is constrained to be a discrete distribution. We rewrite (7) as a quadratically constrained program using the plug-in estimator of MMD,

$$\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^M \alpha_i l(\zeta_i) \\
\text{subject to} \quad & \alpha^\top K_z \alpha - 2\frac{1}{N}\alpha^\top K_{zx}\mathbf{1} + \frac{1}{N^2}\mathbf{1}^\top K_x \mathbf{1} \leq \epsilon^2 \\
& \sum_{i=1}^M \alpha_i = 1, \alpha_i \geq 0, i = 1 \dots M.
\end{aligned}$$

This is a quadratically constrained linear objective convex optimization problem, where the Gram matrix K_z almost always has exponentially decaying eigenvalues. By applying S-lemma [25], this program can be reformulated as the following SDP,

$$\begin{aligned}
\min_{\lambda \geq 0, x, y \geq 0, t} \quad & t \\
\text{subject to} \quad & \begin{bmatrix} \lambda P & -\lambda q - \frac{1}{2}(l + x \cdot \mathbb{1} + y) \\ (-\lambda q - \frac{1}{2}(l + x \cdot \mathbb{1} + y))^\top & t - \lambda\epsilon^2 + x + \lambda r \end{bmatrix} \succeq 0,
\end{aligned} \tag{20}$$

where $P := K_z$, $q := \frac{1}{N}\mathbf{1}^\top K_{zx}$, $r := \frac{1}{N^2}\mathbf{1}^\top K_x \mathbf{1}$, and $K_z = [k(\zeta_i, \zeta_j)]_{ij}, K_{zx} = [k(\zeta_i, \hat{\xi}_j)]_{ij}, K_x = [k(\hat{\xi}_i, \hat{\xi}_j)]_{ij}, l = [l(\zeta_1), \dots, l(\zeta_M)]^\top$.

A.8 Deriving the unconstrained program (11) using conditional value-at-risk (CVaR)

We relax the infinite constraint in (3) by using the CVaR. Given the constraint satisfaction level α, the resulting approximate K-DRO program is

$$\min_{\theta, f_0 \in \mathbb{R}, f \in \mathcal{H}} \delta^*_{\mathcal{C}}(f) + f_0 \quad \text{subject to } P_\zeta\text{–CVaR}_\alpha(l(\theta, \zeta) - (\zeta) - f_0) \leq 0, \tag{21}$$

where $\mathcal{C}$ is the RKHS norm-ball uncertainty set. The CVaR constraint can be written as $P_\zeta\text{–CVaR}_\alpha(l(\theta, \zeta) - (\zeta)) \leq f_0$. Then

$$P_\zeta\text{–CVaR}_\alpha(l(\theta, \zeta) - (\zeta)) = \inf_{t \in \mathbb{R}} \{\frac{1}{\alpha}\mathbb{E}_{P_\zeta} [l - f - t]_+ + t\} \leq f_0.$$

By eliminating the epigraphic variable f_0 in (21), replacing t with f_0, and applying the plug-in form of the expectation, we obtain (11)

$$\min_{\theta, f_0 \in \mathbb{R}, \beta_i, \gamma_i \in \mathbb{R}} \frac{1}{\alpha M}\sum_{j=1}^M [l(\theta, \zeta_j) - f_s(\zeta_j) - f_0]_+ + f_0 + \frac{1}{N}\sum_{j=1}^N f_s(\xi_i) + \epsilon\|f_s\|_{\mathcal{H}}.$$

This program is also convex. Its effect is similar to (10) when α is small and less conservative when α increases.

A.9 Derivation of IPM-DRO reformulation

A reformulation of IPM-DRO is given by

$$\min_{\theta,\lambda\geq 0, f_0\in\mathbb{R}, f\in\mathcal{F}} f_0 + \frac{1}{N}\sum_{i=1}^{N}\lambda f(\xi_i) + \lambda\epsilon \quad \text{subject to } l(\theta,\xi) \leq f_0 + \lambda f(\xi),\ \forall \xi\in\mathcal{X}. \tag{22}$$

Below is a derivation using the technique alternative to the proof of Proposition 3.2.

Proof. We consider the Lagrangian

$$\begin{aligned}
\mathcal{L}(P;\lambda) = \int l\, dP - \lambda(d_{\mathcal{F}}(P,\hat{P}) - \epsilon) \\
&= \int l\, dP - \lambda \sup_{f\in\mathcal{F}} \int f d(P-\hat{P}) + \lambda\epsilon \\
&= \inf_{f\in\mathcal{F}} \int l - \lambda f\, dP + \lambda \int f d\hat{P} + \lambda\epsilon \\
&\leq \inf_{f\in\mathcal{F}} \sup_{\xi\in\mathcal{X}} [l(\xi) - \lambda f(\xi)] + \frac{\lambda}{N}\sum_{i=1}^{N} f(\xi_i) + \lambda\epsilon. \qquad (23)
\end{aligned}$$

The second equality above is due to the dual representation of IPM. The last inequality is due to that the expectation is always dominated by the supremum. This results in the reformulation

$$\min_{\theta,\lambda\geq 0, f\in\mathcal{F}} \sup_{\xi\in\mathcal{X}} [l(\xi) - \lambda f(\xi)] + \frac{\lambda}{N}\sum_{i=1}^{N} f(\xi_i) + \lambda\epsilon.$$

By introducing the epigraphic variable f_0, we obtain the reformulation (22). □

B Further numerical results

In this section, we carry out additional numerical experiments to study K-DRO.

B.1 Testing other variants of K-DRO

We empirically test the following proposed variants of K-DRO.

- Relaxed K-DRO formulation (6) (K-DRO-relaxed)
- Unconstrained K-DRO using CVaR (11) (K-DRO-CVaR)

Figure 6 plots the test error under the same setting as in Section 4.1.

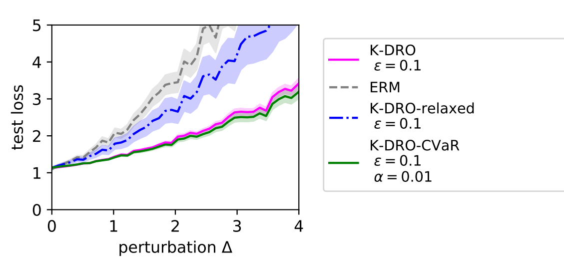


Figure 6: Comparing K-DRO-relaxed (6), K-DRO-CVaR (11), ERM, and regular K-DRO. y-axis limit is adjusted to show the plot. All error bars are in standard error.

First, let us consider formulation (6), which we will refer to as K-DRO-relaxed. We compare it with the ERM as well as the regular K-DRO. Compared with ERM, K-DRO-relaxed still possesses moderate robustness. In this case, we effectively proposed a way to apply RKHS regularization to general optimization problems, not limited to kernelized models. Hence, it may be used in practice as a finite-sample approximation to K-DRO.

We then test the K-DRO-CVaR formulation (11). As expected from the discussion in the main text, we observe no significant difference in performance between K-DRO-CVaR formulation (11) (with small chance constraint level α. cf. Section (A.8)) and regular K-DRO (10).

B.2 Analyzing the generalization bound

In the experiment in Section 4.1, an insight can be obtained by observing the plot of the MMD estimator between the training and test data in Figure 7 (left). As K-DRO with $\epsilon = 0.5$ robustified against perturbation less than the level $\mathrm{MMD} = 0.5$, we see this threshold was exceeded as we increase the perturbation in test data. Meanwhile, this is the same time ($\Delta \approx 1.5$) where K-DRO solutions start to exceed the generalization bound $\int l\, dP \leq f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$, see Figure 7 (right). This empirically validates our theoretical results for robustification.

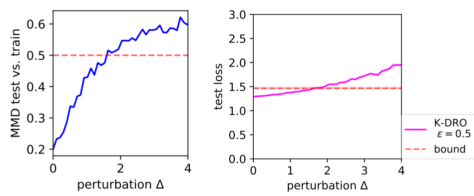
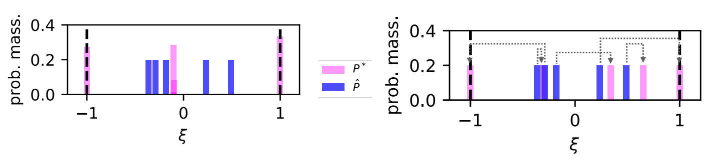


Figure 7: (Left) MMD estimator between the empirical samples and test samples. The level $\mathrm{MMD} = 0.5$ is marked in red. (Right) Loss compared to the generalization bound. As the test data falls outside the robustification level ϵ, the loss starts to exceed the generalization bound (red) $f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$.

B.3 Computing the worst-case distribution

We now empirically examine the *support sampling* method (7) and *perturbation* method (8) to recover the worst-case distribution. Since both programs (7) and (8) search for the worst-case distribution within a subset of all distributions, their optimal values lower-bound the true worst-case risk (P) in (4), i.e., with finite samples, they are optimistic bound.

Under the experimental setting as in Figure 3b, we ran K-DRO with fewer empirical samples ($N = 5$). After we obtain the K-DRO solution θ^*, we plug it into (7) and (8), respectively, to compute the worst-case distribution P^*. Figure 8 plots the results. Note (7) is a convex optimization problem, while (8) results in a nonlinear program (with Gaussian kernel). Nonetheless, we solve it with an always-feasible initialization $d_i = 0$.



(a) Support sampling

(b) Perturbation

Figure 8: Computing the worst-case distribution P^* using **(a)**: (7) samples possible support ζ_j then optimizes w.r.t. weights α. **(b)**: (8) moves the empirical samples directly. The figures are the same as in Figure 2.

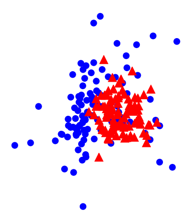
B.4 Miscellaneous details for experimental setup

Our experiments are implemented in Python. The convex optimization problems are solved using ECOS or MOSEK interfaced with CVXPY. The nonlinear program in (8) is solved using IPOPT.

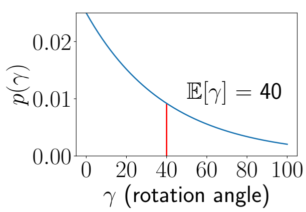
In all the experiments, we chose the bandwidth for the Gaussian kernel using the medium heuristic [19]. ϵ in this paper are fixed to constants below 2 for Gaussian kernels. Choosing ϵ can be further motivated by kernel statistical tests[19] and is left for future work.

Sampled support ζ_j In applying K-DRO, we need to choose sampled support ζ_j (except for the variant K-DRO-relaxed (6)). In practical applications, we may obtain support ζ_j by simply sampling in $\mathcal{X}$. $\{\zeta_j\}_i$ need not be real data, e.g., in stochastic control, they can be a grid of system states; in learning, they can be synthetic samples such as convex combinations of data $\zeta_j = \sum_{i=1}^{N} a_{ij}\xi_i, a_{\cdot j} \in S_N$ (simplex), or perturbations $\zeta_j = \hat{\xi}_i + \Delta_i$ where Δ_i can be a small perturbation, or they can be obtained by domain knowledge of the specific application. In the setting of supervised machine learning, there is a difference between this paper's approach of sampling ζ_j and commonly used data-augmentation techniques: ζ_j need not have the correct labels or targets. Directly training on them may have unforeseen consequences.

In the robust least squares experiment in Section 4.1, we sampled the support ζ_i uniformly random from $[-1, 1]$. In the distributionally robust classification experiment, the sampled candidate for support points $\zeta_j = [x_j, y_j]$ are plotted in (Figure 9). For simplicity, each possible support point ζ_j for $j = 1, \ldots, M := 200$ is drawn from $p(x, y|\gamma)$ where $\gamma \sim \mathrm{Exp}(\lambda = 1/40)$ (see Figure 9b). Note it is possible to use other sampling schemes such as randomly sampling ζ_j within the domain.



(a) Support points



(b) Rotation angle density

Figure 9: Distributionally robust classification (Section 4.2) **(b)**: probability density of the random angle γ for generating the support points.. $M = 200$ sampled support points, each drawn from $p(x, y|\gamma)$ with a random angle γ.

C Supporting lemmas

We establish a few technical results that are used in the proofs.

C.1 Reducing conic constraint to infinite constraint

To derive the semi-infinite constraint in (5), we need a standard result from the literature of the moment problem. We give a self-contained proof below.

Lemma C.1. *Let K^* be the dual cone to the probability simplex $\mathcal{P}$. The conic constraint $l - f - f_0 \in -K^*$ is equivalent to*

$$l(\theta, \xi) \le f_0 + f(\xi), \; \forall \xi \in \mathcal{X}. \tag{24}$$

Proof. " $\Longrightarrow$ ": Let us consider the set of all Dirac measures on $\mathcal{X}$, $\mathcal{D} := \{\delta_\xi : \xi \in \mathcal{X}\}$. For any $\xi \in \mathcal{X}$, we have

$$l(\xi) - f_0 - f(\xi) = \int l - f_0 - f d\delta_\xi \le 0.$$

Hence sufficiency.

"$\Longleftarrow$": Suppose there exists $P' \in \mathrm{co}(\mathcal{P})$ such that $\int l - f_0 - f dP' > 0$. Without loss of generality, we assume $P' \in \mathcal{P}$, or we can normalize it to be a probability measure. Then,

$$0 < \int l - f_0 - f dP' \leq \sup_{\xi \in \mathcal{X}} l(\xi) - f_0 - f(\xi) \leq 0.$$

The second inequality is due to that expectation is always less than or equal to the supremum. The last inequality holds because l is u.s.c. This double inequality is impossible, hence $l - f_0 - f \in -K^*$. $\square$

Note an extension of this result to generating classes other than all Dirac measures $\mathcal{D}$ can be proved using Choquet theory, cf. [34, Proposition 6.66] [26, Lemma 3.1], as well as in [33, 28].

C.2 Compactness of the ambiguity set

We now prove the compactness of the ambiguity set. We use the mean map notation $\mathcal{T} : P \mapsto \mu_P$ to denote a map between the space of $\mathcal{P}$ equipped with MMD, and $\mathcal{H}$ equipped with its norm. Let us denote the image of a subset $\mathcal{K}$ of measures under $\mathcal{T}$ by $\mathcal{T}(\mathcal{K}) := \{\mu_P \mid P \in \mathcal{K}\} \subseteq \mathcal{H}$. If $\mathcal{H}$ is universal, then MMD is a metric. By the definition of MMD, $\mathcal{T}$ is an isometry (i.e., distance-preserving map) between $\mathcal{P}$ and $\mathcal{H}$.

Lemma C.2. *$\mathcal{T}(\mathcal{P})$ is compact if $\mathcal{X}$ is compact.*

Proof. If $\mathcal{X}$ is compact, by Prokhorov's theorem $\mathcal{P}$ is compact. Since $\mathcal{T}$ is an isometry, $\mathcal{T}(\mathcal{P})$ is compact. $\square$

It is straightforward to verify that $\mathcal{T}(\mathcal{P})$ is convex.

Lemma C.3. *Let $C_P = \mathcal{C} \cap \mathcal{T}(\mathcal{P})$. If $\mathcal{X}$ is compact, under Assumption 3.1, C_p is compact.*

Proof. By the Krein-Milman theorem, the convexity and compactness of $\mathcal{T}(\mathcal{P})$ (proved in the previous lemma) imply that $\mathcal{T}(\mathcal{P})$ is closed. By Assumption 3.1, $\mathcal{C}$ is closed, which results in the closedness of C_p. Since C_p is a closed subset of a compact set $\mathcal{T}(\mathcal{P})$, it is compact. $\square$

Recall that we denote the feasible set of probability measures, i.e., ambiguity set, for primal K-DRO (2) by $\mathcal{K}_{\mathcal{C}} = \{P\colon \int \phi \, dP = \mu, \mu \in \mathcal{C}, P \in \mathcal{P}\}$. It is convex by straightforward verification. Let us derive the following compactness property of the ambiguity set.

Lemma C.4. *If $\mathcal{X}$ is compact, under Assumption 3.1, $\mathcal{K}_{\mathcal{C}}$ is compact.*

Proof. We first note $\mathcal{K}_{\mathcal{C}} = \mathcal{T}^{-1}(C_p)$ and $\mathcal{T}$ is an isometric isomorphism (i.e., bijective isometry) between $\mathcal{K}_{\mathcal{C}}$ and C_p. Then $\mathcal{K}_{\mathcal{C}}$ is compact since C_p is compact. $\square$