Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation

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
Currently at Google Research, NYC, USA
wittawatj@gmail.com

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

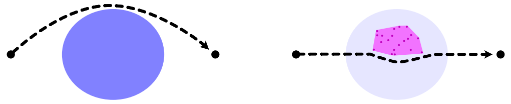
Bernhard Schölkopf
Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany
bernhard.schoelkopf@tuebingen.mpg.de

Abstract

We propose *kernel distributionally robust optimization* (Kernel DRO) using insights from the robust optimization theory and functional analysis. Our method uses reproducing kernel Hilbert spaces (RKHS) to construct a wide range of convex ambiguity sets, which can be generalized to sets based on integral probability metrics and finite-order moment bounds. This perspective unifies multiple existing robust and stochastic optimization methods. We prove a theorem that generalizes the classical duality in the mathematical problem of moments. Enabled by this theorem, we reformulate the maximization with respect to measures in DRO into the dual program that searches for RKHS functions. Using universal RKHSs, the theorem applies to a broad class of loss functions, lifting common limitations such as polynomial losses and knowledge of the Lipschitz constant. We then establish a connection between DRO and stochastic optimization with expectation constraints. Finally, we propose practical algorithms based on both batch convex solvers and stochastic functional gradient, which apply to general optimization and machine learning tasks.

1 INTRODUCTION

Imagine a hypothetical scenario in the illustrative figure where we want to arrive at a destination while avoiding unknown obstacles. A *worst-case robust optimization* (RO) (Ben-Tal et al., 2009) 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

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

distributionally robust optimization (DRO) (Delage and Ye, 2010; Scarf, 1958).

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 robustness properties, e.g., (Christmann and Steinwart, 2007; Xu et al., 2009). However, this robustness only applies to kernelized models. This paper extends the robustness of kernel methods using the robust counterpart formulation techniques (Ben-Tal et al., 2009) as well as the principled conic duality theory (Shapiro, 2001). We term our approach *kernel distributionally robust optimization* (Kernel DRO), which can robustify general optimization solutions not limited to kernelized models.

The *main contributions* of this paper are:

1. We rigorously prove the generalized duality theorem (Theorem 3.1) that reformulates general DRO into a convex dual problem searching for RKHS functions, lifting common limitations of DRO on the loss functions, such as the knowledge of Lipschitz constant. The theorem also constitutes a generalization of the duality results from the literature of mathematical problem of moments.
2. We use RKHSs to construct a wide range of convex ambiguity sets (in Table 1, 3), including sets based on integral probability metrics (IPM) and finite-order moment bounds. This perspective unifies existing RO and DRO methods.
3. We propose computational algorithms based on both convex solvers and stochastic approximation, which can be applied to robustify general optimization and machine learning models not limited to kernelized or known-Lipschitz-constant ones.
4. Finally, we establish an explicit connection between DRO and stochastic optimization with expectation constraints. This leads to a novel stochastic functional gradient DRO (SFG-DRO) algorithm which can scale up to modern machine learning tasks.

In addition, we give complete self-contained proofs in the appendix that shed light on the connection between RKHSs, conic duality, and DRO. We also show that universal RKHSs are large enough for DRO from the perspective of functional analysis through concrete examples.

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. $\operatorname{ri}(\cdot)$ denotes the relative interior of a set. A function f is upper semicontinuous on $\mathcal{X}$ if $\limsup_{x\to x_0} f(x) \le f(x_0), \forall x_0 \in \mathcal{X}$; it is proper if it is not identically $-\infty$. When there is no ambiguity, we simplify the loss function notation $l(\theta, \cdot)$ by using l to indicate that results hold for θ point-wise.

2.1 Robust and distributionally robust optimization

Robust optimization (RO) (Ben-Tal et al., 2009) 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*. Recently, RO has been applied to the setting of adversarially robust learning, e.g., in (Madry et al., 2019; Wong and Kolter, 2018), which we will visit in this paper. In the optimization literature, a typical approach to solving RO is via reformulating the min-max program using duality to obtain a single minimization problem. 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. Compared with RO, DRO only robustifies the solution against a subset $\mathcal{K}$ of distributions on $\mathcal{X}$ and is, therefore, less conservative (since $\sup_{P\in\mathcal{K}}\{\int l(\theta, \xi)\ dP(\xi)\} \le \sup_{\xi\in\mathcal{X}} l(\theta, \xi)$).

The inner problem of (1), historically known as the *problem of moments* traced back at least to Thomas Joannes Stieltjes, estimates the worst-case risk under uncertainty in distributions. The modern approaches, pioneered by the work of (Isii, 1962) (see also (Lasserre, 2002; Shapiro, 2001; Bertsimas and Popescu, 2005; Popescu, 2005; Vandenberghe et al., 2007; Van Parys et al., 2016; Zhu et al., 2020)), typically seek a sharp upper bound via duality. This duality, rigorously justified in (Shapiro, 2001), 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 (finite-order) moment constraints has been studied in (Delage and Ye, 2010; Scarf, 1958; Zymler et al., 2013). The authors of (Ben-Tal et al., 2013; Iyengar, 2005; Nilim and El Ghaoui, 2005; Wang et al., 2016; Duchi et al., 2016) studied DRO using likelihood bounds as well as $\phi-$divergence. Wasserstein-distance-based DRO has been studied by the authors of (Mohajerin Esfahani and Kuhn, 2018; Zhao and Guan, 2018; Gao and Kleywegt, 2016; Blanchet et al., 2019), and applied in a large body of literature. Many existing approaches require either the assumptions such as quadratic loss functions or the knowledge of Lipschitz constant or RKHS norm of the loss l, which are often hard to obtain in practice; see (Virmaux and Scaman, 2018; Bietti et al., 2019).

2.2 Reproducing kernel Hilbert spaces

A symmetric function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 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 Hilbert space $\mathcal{H}$ and 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 denote the canonical feature map as $\phi(x):=k(x,\cdot)$. 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 $C(\mathcal{X})$ (Steinwart and Christmann, 2008, 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 (SVM) for classification problems (Cortes and Vapnik, 1995; Boser et al., 1992; Schölkopf et al., 2000). More recently, the use of RKHSs has been extended to manipulating and comparing probability distributions via kernel mean embedding (Smola et al., 2007). 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 $\mu_P\in\mathcal{H}$ (Smola et al., 2007, 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 $\mathcal{P}$ (Gretton et al., 2012). 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) (Gretton et al., 2012). 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 the optimum f^* is a witness function (Gretton et al., 2012; Sriperumbudur et al., 2012).

3 THEORY

We make the following assumption for the proof.

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

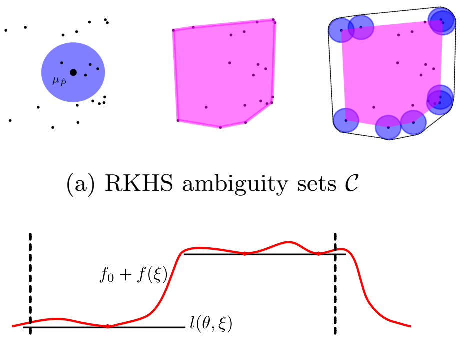
This assumption is general in that it does not require the knowledge of the Lipschitz constant or the RKHS $l(\theta,\cdot)$ lives in. Generally speaking, the DRO problem (1) requires two essential elements: an appropriate ambiguity set that contains meaningful distributions and a sharp reformulation of the min-max problem. We first present the former in Section 3.1, and then the latter in Section 3.2. Complete proofs of our theory are deferred to the appendix.

3.1 Generalized primal formulation

We now present the primal formulation of kernel distributionally robust optimization (Kernel DRO) as a generalization of existing DRO frameworks.

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

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) ambiguity 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 usual ambiguity set. Intuitively, the set $\mathcal{C}$ restricts the RKHS embeddings of distributions in the ambiguity



(a) RKHS ambiguity sets $\mathcal{C}$

(b) Interpretation of Kernel DRO

Figure 1: **(a)**: Geometric intuition for choosing ambiguity 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 Kernel DRO (4). 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 $\mathcal{X}$.

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.

To better understand our unifying formulation, let us examine the celebrated SVM through the lens of our generalized formulation.

Example 3.1 (SVM as generalized DRO)**.** Let us consider SVM for regression, without using slack variables or regularization for simplicity. This can be formulated as optimizing the loss $\min_{f\in\mathcal{H}} \max_i [|y_i - f(x_i)| - \eta]_+$ where $\eta > 0$ is the parameter for the hinge loss. This can be seen as the generalized DRO

$$\min_{f\in\mathcal{H}} \sup_{P\in\mathcal{K}} \int [|y - f(x)| - \eta]_+ dP(x, y),$$

where the ambiguity set is given by the polytope $\mathcal{K} = \mathrm{clconv}\{\delta_{\xi_1}, \ldots, \delta_{\xi_N}\}, \xi_i = (x_i, y_i)$.

Let us now consider a small RKHS to understand the effect of different RKHSs on Kernel DRO.

Example 3.2 (DRO with 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. $\mathrm{MMD}_{k_1}(\hat{P}, Q_v) = 0, \forall v \neq 0$ since they share the first moment. Therefore, any $\mathcal{C}$ that contains $\mu_{\hat{P}}$ also contains all the distributions in $\{\mu_{Q_v}, v \neq 0\}$.

This example shows that small RKHSs force Kernel DRO to robustify against a large set of distributions, resulting in conservativeness. In the extreme, if we choose the smallest possible RKHS $\mathcal{H} = \{0\}$, then the space does not contain functions to separate any distinct distributions. This renders Kernel DRO (4) overly conservative since we can only choose $f = 0$ in (4) — it is precisely reduced to worst-case RO. On the other extreme, the next example shows the downside of function spaces that are too large.

Example 3.3 (DRO with large function space)**.** Suppose $\mathcal{H}$ is the space of all bounded measurable functions, the metric induced by $\mathcal{H}$ becomes the *total variation* distance (Sriperumbudur et al., 2011). While the induced topology is strong, (4) has a trivial solution $f = l, f_0 = 0$. By plugging this solution into (4), we recover (2). Hence the reformulation becomes meaningless.

We distinguish between DRO without metrics, e.g., moment constraints and SVMs, and DRO with probability metric or divergence, e.g., Wasserstein metric. Let us first examine an instance of the former using Kernel DRO. We return to the latter at the end of this section.

Example 3.4 (Reduction to DRO with moment constraints)**.** Kernel DRO with the second-order polynomial kernel $k_2(x, y) := (1 + x^\top y)^2$ and a singleton ambiguity set $\mathcal{C} = \{\mu_{\hat{P}}\}$ robustifies against all distributions sharing the first two moments with $\hat{P}$. This is equivalent to DRO with known first two moments, such as in (Delage and Ye, 2010; Scarf, 1958). More generally, the choice of the pth-order polynomial kernel $k_p(x, y) := (1 + x^\top y)^p$ corresponds to DRO with known first p moments.

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. (Sriperumbudur et al., 2011)). Then the induced topology (MMD) is strong enough to separate distinct probability measures. Meanwhile, RKHS allows for efficient computation using tools from kernel methods, as shown in Section 4. Therefore, our insight is that *universal RKHSs are large enough* for DRO applications.

Remark. For the RKHS associated with the Gaussian kernel, the diameter of the space can be computed: $\forall p, q, \|\mu_p - \mu_q\|_\mathcal{H} \leq \|\mu_p\|_\mathcal{H} + \|\mu_q\|_\mathcal{H} \leq 2\sup_{x,y} \sqrt{k(x, y)} = 2$. Hence, if $\epsilon \geq 2$, $\mathcal{C}$ contains all probability distributions. Then, Kernel DRO is again reduced to worst-case RO on domain $\mathcal{X}$.

We now turn to DRO with a generalized class of integral probability metrics (IPM).

Example 3.5 (Generalization to IPM-DRO)**.** Suppose $d_\mathcal{F}(P, \hat{P}) := \sup_{f\in\mathcal{F}} \int f d(P - \hat{P})$ is the IPM defined by some function class $\mathcal{F}$. The IPM-DRO primal formulation is given by

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

RKHS ambiguity set $\mathcal{C}$	Support function $\delta^*_{\mathcal{C}}(f)$
RKHS norm-ball $\mathcal{C} = \{\mu\colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \leq \epsilon\}$	$\frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$
Polytope $\mathcal{C} = \text{conv}\{\phi(\xi_1), \ldots, \phi(\xi_N)\}$	$\max_i f(\xi_i)$ (scenario optimization; SVMs with no slack)
Minkowski sum $\mathcal{C} = \sum_{i=1}^N \mathcal{C}_i$	$\sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f)$
Whole space $\mathcal{C} = \mathcal{H}$	0 if $f = 0$, ∞ otherwise (worst-case RO (Ben-Tal et al., 2009))

$$\min_{\theta} \sup_{d_{\mathcal{F}}(P,\hat{P})\leq\epsilon} \int l(\theta,\xi)\ dP(\xi). \tag{3}$$

If we choose the class $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, we recover Kernel DRO with the RKHS norm-ball set in Table 1. Similarly, $\mathcal{F} = \{f : \text{lip}(f) \leq 1\}$ recovers the (type-1) Wasserstein-DRO. This puts Wasserstein-DRO and Kernel DRO into a unified perspective.

3.2 Generalized duality theorem

We now present the main theorem of this paper, the generalized duality theorem of Kernel DRO (2).

Theorem 3.1 (**Generalized Duality**). *Under Assumption 3.1,* (2) *is equivalent to*

$$\begin{aligned}(D) := &\min_{\theta, f_0\in\mathbb{R}, f\in\mathcal{H}} && f_0 + \delta^*_{\mathcal{C}}(f)\\ &\text{subject to} && l(\theta,\xi) \leq f_0 + f(\xi),\ \forall\xi\in\mathcal{X}\end{aligned} \tag{4}$$

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.,* $(P) = (D)$, 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. If l is convex in θ, then (4) is a *convex program.* Formulation (4) 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. Note the term duality here refers to the inner moment problem. The statement can be further simplified by replacing $f_0 + f$ with f. However, we choose the current notation for the sake of its explicit connection to RO.

Proof sketch. Our weak duality proof follows standard paradigms of Lagrangian relaxation by introduing dual variables. Notably, we associate the functional constraint $\int \phi\ dP = \mu$ with a dual function $f \in \mathcal{H}$, which is the decision variable in the dual problem (4). Using the reproducing property of RKHSs and conic duality, we arrive at (4) with weak duality. Our strong duality proof is an extension of the conic strong duality in Eulidean spaces. We rely on the existance of separating hyperplnes between convex sets in locally convex function spaces, e.g., $\mathcal{H}$. See the illustration in Figure 2. In our generalized duality theorem, this separating hyperplane is determined by the witness function f^*, which is the optimal dual variable in (4). See the appendix for the full proof.

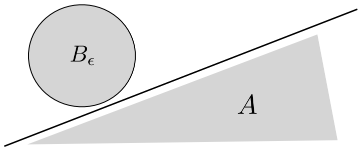


Figure 2: Illustration of a separating hyperplane in $\mathcal{H}$

Theorem 3.1 generalizes the classical bounds in generalized moment problems (Isii, 1962; Lasserre, 2002; Shapiro, 2001; Bertsimas and Popescu, 2005; Popescu, 2005; Vandenberghe et al., 2007; Van Parys et al., 2016) to infinitely many moments using RKHSs. A distinction between Theorem 3.1 and other DRO approaches is that it uses the density of universal RKHSs to find a surrogate which can sharply bound the worst-case risk. This means that we do not require the loss $l(\theta,\cdot)$ to be affine, quadratic, or living in a known RKHS, nor do we require the knowledge of Lipschitz constant or RKHS norm of $l(\theta,\cdot)$. To our knowledge, existing works typically require one of such assumptions.

Moreover, Theorem 3.1 generalizes existing RO and DRO in the sense that it gives us a unifying tool to work with various ambiguity and ambiguity sets, which may be customized for specific applications. We outline a few closed-form expressions of the support function $\delta^*_{\mathcal{C}}(f)$ in Table 1, and more in Table 3. We now return IPM-DRO with a duality result.

Corollary 3.1.1 (IPM-DRO duality)**.** Given the integral probability metric $d_{\mathcal{F}}(P,\hat{P}) := \sup_{f\in\mathcal{F}}\int f d(P - \hat{P})$, a dual program to (3) is given by

$$\begin{aligned}&\min_{\theta,\lambda\geq 0, f_0\in\mathbb{R}, f\in\mathcal{F}} && f_0 + \frac{1}{N}\sum_{i=1}^N \lambda f(\xi_i) + \lambda\epsilon\\ &\text{subject to} && l(\theta,\xi) \leq f_0 + \lambda f(\xi),\ \forall\xi\in\mathcal{X}.\end{aligned} \tag{5}$$

The reduction to (4) as a special case can be seen by replacing λf with f and choosing $\mathcal{F} = \mathcal{H}$.

We now establish an explicit connection between DRO and stochastic optimization with expectation constraint, whose solution methods using stochastic approximation are an topic of active research (Lan and Zhou, 2020; Xu, 2020).

Corollary 3.1.2. (Stochastic optimization with expectation constraint) Under the Assumption 3.1, the optimal value of (2) coincides with that of

$$\begin{aligned} \min_{\theta, f_0\in\mathbb{R}, f\in\mathcal{H}} \quad & f_0 + \delta^*_{\mathcal{C}}(f) \\ \text{subject to} \quad & \mathbb{E}h\left(l\left(\theta, \zeta\right) - f_0 - \lambda f\left(\zeta\right)\right) \leq 0 \end{aligned} \tag{6}$$

for some function h that satisfies $h(t) = 0$ if $t \leq 0, h(t) > 0$ if $t > 0$, and random variable $\zeta \sim \mu$ whose probability measure places positive mass on any non-empty open subset of $\mathcal{X}$, i.e., $\mu(B) > 0,\ \forall B \subseteq \mathcal{X}, B \neq \emptyset, B$ is open.

A choice for h is $h(\cdot) = [\cdot]_+$, which is used in the conditional value-at-risk (Rockafellar and Uryasev, 2000). We will see the computational implication of Corollary 3.1.2 in Section 4.

We now establish further theoretical results as a consequence of the generalized duality theorem to help us understand the geometric intuition of how Kernel DRO works. By the weak duality $(P) \leq (D)$ of (11) and (12), we have $\int l\ dP \leq f_0 + \delta^*_{\mathcal{C}}(f)$. Specifically, if $\mathcal{C}$ is the RKHS norm-ball in Table 1 , this inequality becomes $\int l\ dP \leq f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$. Its right-hand-side can be seen as a computable bound for the worst-case risk when generalizing to P. This may be useful when the Lipschitz constant of l is not known or hard to obtain, as is often the case in practice. The following insight is a consequence of a generalization of the classical *complementarity condition* of convex optimization; see the appendix.

Corollary 3.1.3 (Interpolation property)**.** Given θ, let P^*, f^*, f_0^* be a set of optimal primal-dual solutions associated with (P) and (D), 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. On the other hand, the density of universal RKHS allows the interpolation of general loss functions.

It is tempting to approximately solve (4) by relaxing the constraint to hold for only the empirical samples, i.e., $l(\theta, \xi_i) \leq f_0 + f(\xi_i),\ i = 1 \dots N$. The following observation cautions us against this.

Example 3.6 (Counterexample: relaxation of the semi-infinite constraint)**.** Let $\mathcal{H}$ be a Gaussian RKHS with the bandwidth $\sigma = \sqrt{2}$. Suppose our data set is $\{0\}$ and the ambiguity set is $\mathcal{C} := \{\mu\colon \|\mu - \phi(0)\|_{\mathcal{H}} \leq \epsilon\}$. Let $\epsilon = \sqrt{2 - 2/e}$. We consider the loss function $l(\xi) = [|\theta + \xi| - 1]_+$ and relaxing the constraint of (4) to only hold at the empirical sample, i.e.,

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

which admits an optimal solution $\theta^* = 0, f^* = 0, f_0^* = 0$ and the worst-case risk $(d) = 0$. However, let $\mu_{P'} = \frac{1}{2}\phi(0) + \frac{1}{2}\phi(2)$. 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'.

4 COMPUTATION

Given a certain parametrization of the RKHS function f, (4) is a semi-infinite program (SIP) (Guerra Vázquez et al., 2008). In the following, we propose two computational methods that do not require a polynomial loss l or the knowledge of its Lipschitz constant. For simplicity, we only derive the formulations for the RKHS-norm-ball ambiguity set, while other formulations are given in Table 1, 3.

A batch approach by discretization of SIP. We first consider an approach based on the discretization method of SIP (Guerra Vázquez et al., 2008). Let us consider an ambiguity set smaller than the RKHS-norm ball of distributions supported on some $\{\zeta_j\}_{j=1}^M \subseteq \mathcal{X}$. Then it suffices to consider the following program, which relaxes the constraint of (4) to finite support.

$$\begin{aligned} \min_{\theta, f\in\mathcal{H}, f_0\in\mathbb{R}} \quad & f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \\ \text{subject to} \quad & l(\theta, \zeta_i) \leq f(\zeta_j) + f_0,\ j = 1 \dots M. \end{aligned} \tag{7}$$

Note (7) is a *convex program* if $l(\theta, \xi)$ in convex in θ.

We can parametrize the RKHS function f by a wealth of tools from kernel methods, such as the random features $\hat{f}(\xi) = w^\top \hat{\phi}(\xi)$ for large scale problems (Rahimi and Recht, 2008). Alternatively, for small problems, we can parametrize f by a kernel expansion on the the points ζ_i. We provide concrete plug-in forms in the appendix.

As an interesting by-product of (7), let us derive an unconstrained version of (7), which gives rise to a generalized risk measure that we term *kernel conditional value-at-risk* (Kernel CVaR).

Example 4.1 (Kernel CVaR)**.**

$$\begin{aligned}\text{K-CVaR}_\alpha(X) := \inf_{f\in\mathcal{H}, f_0\in\mathbb{R}} \frac{1}{\alpha M}\sum_{j=1}^{M}[X - f(\zeta_j) - f_0]_+ \\ + f_0 + \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}. \quad (8)\end{aligned}$$

If $f = 0$ and $\{\zeta_j\}_{j=1}^M = \{\xi_i\}_{i=1}^N$, then (8) is reduced to the classical CVaR (Rockafellar and Uryasev, 2000).

Program (7) can be readily solved using off-the-shelf convex solvers. However, to scale up to large data sets, we next develop a stochastic approximation (SA) method for Kernel DRO.

Stochastic functional gradient DRO. We now present our SA approach enabled by Theorem 3.1 by employing two key tools: 1) scalable approximate RKHS features, such as random Fourier features (Rahimi and Recht, 2008; Dai et al., 2014; Carratino et al., 2018), and 2) stochastic approximation with semi-infinite and expectation constraints (Tadić et al., 2006; Lan and Zhou, 2020; Baes et al., 2011; Xu, 2020).

Let us summon Corollary 3.1.2 to formulate a stochastic program with expectation constraint.

$$\begin{aligned}\min_{\theta, f_0\in\mathbb{R}, f\in\mathcal{H}} \quad & f_0 + \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \\ \text{subject to} \quad & \mathbb{E}[l\,(\theta, \zeta) - f_0 - \lambda f\,(\zeta)]_+ \le 0\end{aligned} \qquad (9)$$

where ζ follows a certain proposal distribution on $\mathcal{X}$, e.g., uniform or by adaptive sampling. An alternative is to directly solve (4) using SA techniques with SI constraints, such as (Tadić et al., 2006; Wei et al., 2020). (4), (6), and (9) are all convex in function f. We can compute the functional gradient by

$$\nabla_f f = \phi, \quad \nabla_f \|f\|_{\mathcal{H}} = \frac{f}{\|f\|_{\mathcal{H}}}. \qquad (10)$$

When used with approximate features of the form $\hat{f}(\xi) = w^\top \hat{\phi}(\xi)$, we further have $\nabla_w \hat{f}(\xi) = \hat{\phi}(\xi), \nabla_w \|\hat{f}\|_{\mathcal{H}} = w/\|w\|_2$. We outline our stochastic functional gradient DRO (SFG-DRO) in Algorithm 1.

Compared with many batch-setting DRO approaches, SFG-DRO can be used with general model classes, such as neural networks, and is applicable to a broad class of optimization and modern learning tasks. The convergence guarantee follows that of the specific SA routine used in Step 5 of the algorithm. It is worth noting that, when used with a primal SA approach such as (Lan and Zhou, 2020), SFG-DRO completely operates in the dual space (an RKHS) since Kernel DRO (4) is based on the generalized duality Theorem 3.1. This interplay between the primal (measures) and dual (functions) is the essence of our theory.

Algorithm 1 Stochastic Functional Gradient DRO (SFG-DRO)

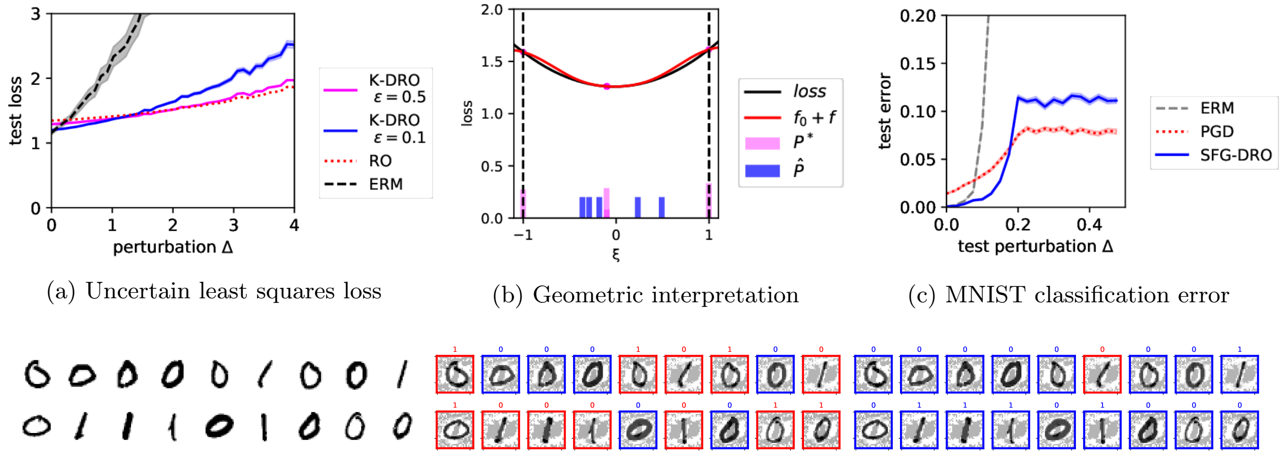
```
1: for $k = 1, 2, \dots$ do
2:     Sample mini-batch data $\{\xi_i^k\} := \{x_i, y_i\}$. Sample $\{\zeta_i\}$ from some proposing distribution.
3:     Approximate $f$, e.g., by random Fourier feature $\hat{f}(\xi_i^k) = w^\top \hat{\phi}(\xi_i^k)$.
4:     Estimate the stochastic functional gradient of the objective and constraint in (9) using (10).
5:     Update $\theta, f_0, f$ using the functional gradient with any SA routine with expectation or semi-infinite constraints, e.g., (Lan and Zhou, 2020; Xu, 2020; Tadić et al., 2006; Wei et al., 2020) .
```

5 NUMERICAL STUDIES

This section showcases the applicability of Kernel DRO (and hence SFG-DRO) and discusses the robustness-optimality trade-off. Our purpose is not to benchmark state-of-art performances or to demonstrate the superiority of a specific algorithm. Indeed, we believe both RO and DRO are elegant theoretical frameworks that have their specific use cases. We note that our theory can be applied to a broader scope of applications than the examples here, such as stochastic optimal control. See the appendix for more experimental results. The code is available at `https://github.com/jj-zhu/kdro`.

Distributionally robust solution to uncertain least squares. We first consider a robust least squares problem adapted from (El Ghaoui and Lebret, 1997), which demonstrated an important application of RO to statistical learning historically. (See also (Boyd et al., 2004, 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 Kernel 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 (El Ghaoui and Lebret, 1997). We consider a data-driven setting with given samples $\{\xi_i\}_{i=1}^N$ with the Kernel DRO formulation $\min_\theta \max_{P\in\mathcal{P}, \mu\in\mathcal{C}} \mathbb{E}_{\xi\sim P}\ \|A(\xi)\ \theta - b\|_2^2$ subject to $\int \phi\, dP = \mu$, where we choose the ambiguity set to be the ϵ-norm-ball in the RKHS (Table 1).

Empirical samples $\{\xi_i\}_{i=1}^N (N = 10)$ are generated uniformly from $[-0.5, 0.5]$. We then apply Kernel



(a) Uncertain least squares loss

(b) Geometric interpretation

(c) MNIST classification error

(d) (Left) unperturbed data (Center) ERM classification result (red indicates errors) (Right) SFG-DRO (Kernel DRO)

Figure 3: **Uncertain least squares.** **(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 Kernel 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. **Classification under perturbation** **(c)** We plot the classification error rate during test time. The $x-$axis is the perturbation magnitude allowed on the test data. For ERM, PGD, and SFG-DRO (Kernel DRO), we train 5 independent models. Each model is tested on 500 randomly sampled images. **(d)** We visualize the predictions of ERM and SFG-DRO on the perturbed images with perturbation magnitude $\Delta = 0.2$. Blue frames indicate correct predictions while the red ones indicate errors.

DRO formulation (7). 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, Kernel DRO achieves some level of optimality while retaining robustness.

We then ran Kernel 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 the inner moment problem (see appendix) 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. From Figure 3b, we can see that the *intuition of Kernel DRO is to flatten the loss curve using a smooth function.*

Distributionally robust learning under adversarial perturbation. We now demonstrate the framework of SFG-DRO in Algorithm 1 in a non-convex setting. For simplicity, we consider a MNIST binary classification task with a two-layer neural network. We emphasize that the deliberate choice of this simple architecture ablates factors known to implicitly influence robustness, such as regularization and dropout. The data set contains images of zero and one (i.e., two classes). Each image x is represented by $x \in [0, 1]^{28\times 28}$. The test data is perturbed by an *unknown* disturbance, i.e., $\tilde{x}_{\text{test}} := x + \delta$ where $x \sim P_{\text{test}}$ is the unperturbed test data and δ is the perturbation. In the plots, δ is generated by the PGD algorithm (Madry et al., 2019) using projected gradient descent to find the worst-case perturbation within a box $\{\delta : \|\delta\|_\infty \leq \Delta\}\}$. We compared SFG-DRO (Kernel DRO) with ERM and PGD (cf. (Madry et al., 2019; Madry)). Note the overall loss of PGD is an average loss instead of a worst-case one. Hence it is already less conservative than RO. We train a classification model $g_\theta : x \mapsto y$ using SFG-DRO in Algorithm 1, with the SA subroutine of (Lan and Zhou, 2020). During training, we set the ambiguity size of SFG-DRO as $\epsilon = 0.5$ and domain $\mathcal{X}$ to be norm-balls around the training data $\mathcal{X} = \{\zeta = X + \delta : \|\delta\|_\infty \leq 0.5\}$ where X is the training data.

Figure 3d (left) plots unperturbed test samples. Figure 3c shows the classification error rate as we increase the magnitude of the perturbation Δ. We observe that ERM attains good performance when there is no test-time perturbation but quickly underperforms as the noise level increases. PGD is the most robust under large perturbation, but has the worst nominal performance. SFG-DRO possesses improved robustness while its performance under no perturbation does not become much worse. This is consistent with our theoretical insights into RO and DRO.

6 OTHER RELATED WORK AND DISCUSSION

This paper uses similar techniques of reformulating min-max programs as in (Ben-Tal et al., 2015; Bertsimas et al., 2017), but our ambiguity set is constructed in an RKHS. Duchi et al. (2020) proposed variational approximations to marginal DRO to treat covariate shift in supervised learning. The authors of (Zhu et al., 2020) used kernel mean embedding for the inner moment problem (but not DRO) and proved the statistical consistency of the solution. The work of Staib and Jegelka (2019) used insights from DRO to motivate a regularizer for kernel ridge regression. DRO has been also applied to Bayesian optimization in (Rontsis et al., 2020; Kirschner et al., 2020), where the latter work used MMD ambiguity sets of distributions over discrete spaces. In terms of scalability, recent works such as (Sinha et al., 2020; Li et al., 2019; Namkoong and Duchi, 2016) also explored DRO for modern machine learning tasks. To the best of our knowledge, no existing work contains the results such as generalized ambiguity set constructions in Table 1, 3, generalized duality theory underpinned by Theorem 3.1, or the stochastic functional gradient algorithm SFG-DRO.

In summary, this paper proves Theorem 3.1 that generalizes the classical duality theory in the literature of mathematical problem of moments and DRO. Using the density of universal RKHSs, the dual bound in Theorem 3.1 is sharp while lifting restrictions on the loss function class. The generalized primal formulations shed light on the connection between Kernel DRO and existing robust and stochastic optimization approaches. Finally, the proposed stochastic approximation algorithm SFG-DRO enables the applications of Kernel DRO to modern learning tasks.

The compactness assumption on $\mathcal{X}$ can be further extended, as universality can be extended to non-compact domains (Sriperumbudur et al., 2011). In the special case of RKHS-norm-ball ambiguity sets, choosing the size ϵ can be motivated using kernel statistical testing (Gretton et al., 2012). However, when DRO is used in the setting where test distributions are perturbed as in our examples, existing statistical guarantees in the literature for unperturbed settings cannot be directly applied. This is a topic of future work.

Acknowledgements

We 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 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

Michel Baes, Michael Bürgisser, and Arkadi Nemirovski. A randomized Mirror-Prox method for solving structured large-scale matrix saddle-point problems. *arXiv:1112.1274 [math]*, December 2011.

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

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

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. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1120.1641.

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. ISSN 1436-4646. doi: 10.1007/s10107-014-0750-8.

Dimitris Bertsimas and Ioana Popescu. Optimal Inequalities in Probability Theory: A Convex Optimization Approach. *SIAM Journal on Optimization*, 15(3):780–804, January 2005. ISSN 1052-6234, 1095-7189. doi: 10.1137/S1052623401399903.

Dimitris Bertsimas, Nathan Kallus, and Vishal Gupta. *Data-Driven Robust Optimization.* Springer Berlin Heidelberg, 2017. ISBN 1010701711258. doi: 10. 1007/s10107-017-1125-8.

Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A Kernel Perspective for Regularizing Deep Neural Networks. *arXiv:1810.00363 [cs, stat]*, May 2019.

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. ISSN 0021-9002, 1475-6072. doi: 10.1017/jpr.2019.49.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

Stephen Boyd, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 978-0-521-83378-3.

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. ISSN 2334-3303. doi: 10.1109/TAC.2006.875041.

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and Random Features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10192–10203. Curran Associates, Inc., 2018.

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, August 2007. ISSN 1350-7265. doi: 10.3150/07-BEJ5102.

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

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

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. Scalable Kernel Methods via Doubly Stochastic Gradients. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3041–3049. Curran Associates, Inc., 2014.

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. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1090.0741.

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

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses for Latent Covariate Mixtures. *arXiv:2007.13982 [cs, stat]*, July 2020.

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. ISSN 0895-4798. doi: 10.1137/S0895479896298130.

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

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.

F. Guerra Vázquez, J. J. Rückmann, O. Stein, and G. Still. Generalized semi-infinite programming: A tutorial. *Journal of Computational and Applied Mathematics*, 217(2):394–419, August 2008. ISSN 0377-0427. doi: 10.1016/j.cam.2007.02.012.

Keiiti Isii. On sharpness of tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14 (1):185–197, December 1962. ISSN 1572-9052. doi: 10.1007/BF02868641.

Garud N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. ISSN 0364-765X.

Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally Robust Bayesian Optimization. *arXiv:2002.09038 [cs, stat]*, March 2020.

Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, February 2020. ISSN 0926-6003, 1573-2894. doi: 10.1007/s10589-020-00179-x.

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

Jiajin Li, Sen Huang, and Anthony Man-Cho So. A First-Order Algorithmic Framework for Wasserstein Distributionally Robust Logistic Regression. *arXiv:1910.12778 [cs, math, stat]*, October 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, September 2019.

Zico Kolter and Aleksander Madry. Adversarial Robustness - Theory and Practice. http://adversarial-ml-tutorial.org/.

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. ISSN 1436-4646. doi: 10.1007/s10107-017-1172-1.

Hongseok Namkoong and John C Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In D. D. Lee,

M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2208–2216. Curran Associates, Inc., 2016.

Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, October 2005. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1050.0216.

Imre Pólik and Tamás Terlaky. A Survey of the S-Lemma. *SIAM Review*, 49(3):371–418, January 2007. ISSN 0036-1445, 1095-7200. doi: 10.1137/S003614450444614X.

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. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.1040.0137.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

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

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

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

Nikitas Rontsis, Michael A. Osborne, and Paul J. Goulart. Distributionally Ambiguous Optimization for Batch Bayesian Optimization. *Journal of Machine Learning Research*, 21(149):1–26, 2020. ISSN 1533-7928.

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

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.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

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. ISBN 978-1-4419-5204-2 978-1-4757-3403-4. doi: 10.1007/978-1-4757-3403-4_7.

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

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571 [cs, stat]*, May 2020.

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.

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. ISSN ISSN 1533-7928.

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. ISSN 1935-7524. doi: 10.1214/12-EJS722.

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

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

Vladislav B. Tadić, Sean P. Meyn, and Roberto Tempo. Randomized Algorithms for Semi-Infinite Programming Problems. In Giuseppe Calafiore and Fabrizio Dabbene, editors, *Probabilistic and Randomized Methods for Design under Uncertainty*, pages 243–261. Springer, London, 2006. ISBN 978-1-84628-095-5. doi: 10.1007/1-84628-095-8_9.

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. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-015-0878-1.

Lieven. Vandenberghe, Stephen. Boyd, and Katherine. Comanor. Generalized Chebyshev Bounds via Semidefinite Programming. *SIAM Review*, 49 (1):52–64, January 2007. ISSN 0036-1445. doi: 10.1137/S0036144504440543.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle,

K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3835–3844. Curran Associates, Inc., 2018.

Zizhuo Wang, Peter W. Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2): 241–261, April 2016. ISSN 1619-697X, 1619-6988. doi: 10.1007/s10287-015-0240-3.

Bo Wei, William B. Haskell, and Sixiang Zhao. The CoMirror algorithm with random constraint sampling for convex semi-infinite programming. *Annals of Operations Research*, September 2020. ISSN 1572-9338. doi: 10.1007/s10479-020-03766-7.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Yangyang Xu. Primal-Dual Stochastic Gradient Method for Convex Programs with Many Functional Constraints. *SIAM Journal on Optimization*, 30(2): 1664–1692, January 2020. ISSN 1052-6234, 1095-7189. doi: 10.1137/18M1229869.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, March 2018. ISSN 01676377. doi: 10.1016/j.orl.2018.01.011.

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.

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. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-011-0494-7.