Kernel Distributionally Robust Optimization: A Generalization Theorem

Jia-Jie Zhu JIA-JIE.ZHU@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany
Wittawat Jitkrittum WITTAWATJ@GMAIL.COM
Google Research, Max Planck Institute for Intelligent Systems, Tübingen, Germany
Moritz Diehl MORITZ.DIEHL@IMTEK.UNI-FREIBURG.DE
University of Freiburg, Freiburg, Germany
Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

This paper is an in-depth investigation using kernel methods to robustify 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 a wide range of convex ambiguity sets, including sets based on integral probability metrics and finite-order moment bounds. This perspective unifies multiple existing robust optimization methods. We then prove a theorem that reformulates the maximization with respect to measures into the dual problem that searches for smooth functions. Using universal RKHSs, the theorem applies to a broad class of loss functions, lifting common limitations such as quadratic loss and knowledge of Lipschitz constant.

1. Introduction

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) [8, 25].

Today's learning tasks suffer from various manifestations of distributional ambiguity — e.g., covariate shift, adversarial attacks — phenomena that are caused by the discrepancy between training and test distributions. Kernel methods are known to possess robustness properties, e.g., [6, 35]. However, this robustness only applies to kernelized models. This paper extends the robustness of kernel methods using the robust counterpart formulation techniques [1] as well as the conic duality theory [27]. We term our approach *kernel distributionally robust optimization* (K-DRO), which can robustify general optimization solutions not limited to kernelized models.

The *main contributions* of this paper are the following.

1. We rigorously prove Theorem 1 for reformulating general DRO into a convex dual problem searching for RKHS functions, lifting common limitations of DRO on the loss functions, such as quadratic loss classes and knowledge of Lipschitz constant.
2. We use RKHSs to construct a wide range of convex ambiguity sets (in Table 1) including sets based on integral probability metrics (IPM) and finite-order moment bounds. This perspective unifies existing RO and DRO methods.

In addition, we give complete self-contained proofs in the full version of this paper[1] that shed light on the connection between RKHSs, conic duality, and DRO.

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 $\mathcal{X}$. 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 $-\infty$.

Distributionally robust optimization. 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. 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 [8, 25, 38]. The authors of [2, 9, 16, 21, 34] studied DRO using likelihood bounds as well as $\phi-$divergence. Wasserstein-distance-based DRO has been studied by the authors of [4, 12, 20, 36], and applied in a large body of literature. Notably, the authors of [26] applied Wasserstein-DRO to a kernel-based learning task, which should not be confused with K-DRO in this paper.

Reproducing kernel Hilbert spaces. A symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel if $\sum_{i=1}^{n} \sum_{i=1}^{n} a_i a_j k(x_i, x_j) \ge 0$ 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 $\phi(x) := k(x, \cdot)$. A continuous kernel k on a compact metric space $\mathcal{X}$ is said to be *universal* if $\mathcal{H}$ is dense in $C(\mathcal{X})$ [31, Section 4.5].

RKHSs first gained widespread attention following the advent of the kernelized support vector machine for classification problems [7]. More recently, the use of RKHSs has been extended to manipulating and comparing probability distributions via kernel mean embedding [28]. Given a distribution P, and a (positive definite) kernel k, the *kernel mean embedding* of P is defined

[1] A full version of this paper is available at https://arxiv.org/pdf/2006.06981.pdf

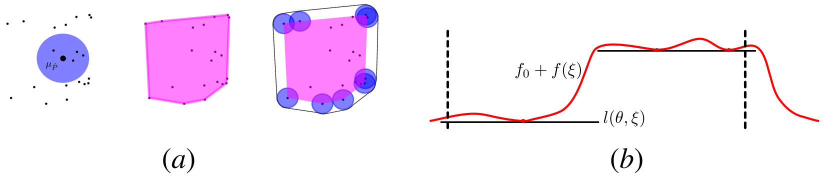


Figure 1: **(a)**: Geometric intuition for choosing uncertainty set $\mathcal{C}$ in $\mathcal{H}$ such as norm-ball, polytope, and Minkowski sum of sets. The scattered points are the embeddings of empirical samples. **(b)**: Geometric interpretation of K-DRO (3). The (red) curve depicts $f_0 + f$, which *majorizes* $l(\theta, \cdot)$ (black). The horizontal axis is ξ. The dashed lines denote the boundary of the domain $\mathcal{X}$.

as $\mu_P := \int k(x, \cdot)\, dP$. If $\mathbb{E}_{x \sim P}[k(x, x)] < \infty$, then $\mu_P \in \mathcal{H}$ [28, Section 1.2]. 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 $\mathcal{P}$ [13]. 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) [13]. 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.

3. Main result

To solve the DRO problem (1), we need two essential elements: an appropriate ambiguity set that contains meaningful distributions and a sharp reformulation of the min-max problem. We now present the kernel distributionally robust optimization (K-DRO), 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 (see the full version for more cases), and illustrate our intuition in Figure 1 (a). We make the following regularity assumptions used in the proofs.

Assumption 1 *$l(\theta, \xi)$ is proper, upper semicontinuous in ξ. $\mathcal{C}$ is closed convex.* $\mathrm{ri}(\mathcal{K}_{\mathcal{C}}) \neq \emptyset$.

We first give the main result for solving K-DRO (2). All proofs are deferred to the full version of this paper.

Table 1: Examples of support functions for K-DRO. See the full version 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 $\mathcal{C} = \mathrm{conv}\{\phi(\xi_1), \ldots, \phi(\xi_N)\}$	$\max_i f(\xi_i)$ (equivalent to scenario approach [5], and 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 (equivalent to RO)

Theorem 1 (K-DRO reformulation) *Under Assumption 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. If l is convex in θ, then (3) is a *convex program*. 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 1 (b).

A distinction between Theorem 1 and other DRO approaches is that it does not use the functional (semi-)norm of the loss itself. Rather, it uses the universality of RKHS to find a surrogate which can sharply bound the worst-case risk. This means we do not require the loss $l(\theta, \cdot)$ to be affine, quadratic, or living in a known RKHS. Nor does it require the knowledge of Lipschitz constant or RKHS norm of the loss. To our knowledge, existing works, such as Wasserstein DRO [20], typically require one of those assumptions.

Theorem 1 generalizes existing RO and DRO in the sense that it gives us a flexible tool to work with various ambiguity and uncertainty sets, which may be customized for specific applications. It reveals the relationship between existing methods such as SVMs, worst-case RO and K-DRO. We outline a few closed-form expressions of the support function $\delta^*_{\mathcal{C}}(f)$ in Table 1, while more are given in Table **??**. In the following, while we pay special attention to the RKHS norm-balls since they provide the elementary topological sets, our proof holds for general choices of $\mathcal{C}$.

Example 1 (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 [8, 25]. 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.*

The example above reveals the relationship between Theorem 1 and the classical bound in generalized moment problems [3, 15, 19, 22, 27, 32, 33]: Theorem 1 generalizes the moment bound to infinite orders.

Our K-DRO reformulation (3) can be further generalized to the 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 $\min_\theta \sup_{d_{\mathcal{F}}(P,\hat{P})\le\epsilon} \int l(\theta,\xi)\, dP(\xi)$ as the following.

$$\min_{\theta,\lambda\ge 0, f_0\in\mathbb{R}, f\in\mathcal{F}} f_0 + \frac{1}{N}\sum_{i=1}^{N} \lambda f(\xi_i) + \lambda\epsilon \quad \text{subject to } l(\theta,\xi) \le f_0 + \lambda f(\xi),\ \forall \xi\in\mathcal{X}. \tag{4}$$

If we choose the class $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$, we recover K-DRO (by noting $\|\lambda f\|_{\mathcal{H}} = \lambda, \forall f\in\mathcal{F}$). 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.

4. Computation and numerical example

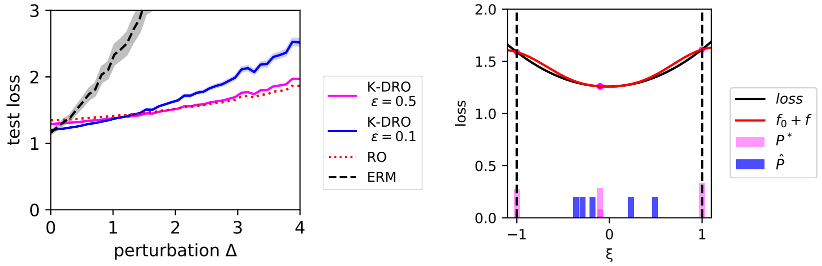
In the following, we propose a strightforward approximation to K-DRO solutions based on the discretization method for solving SIP [14]. Let usconsider K-DRO restricted to a smaller ambiguity set of distributions supported on some $\{\zeta_j\}_{j=1}^M \subseteq \mathcal{X}$. Then it suffices to consider the following program, which relaxes the constraint of (3).

$$\min_{\theta, f\in\mathcal{H}, f_0\in\mathbb{R}} f_0 + \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \quad \text{subject to}\, l(\theta,\xi_i) \le f(\zeta_j) + f_0,\ j = 1\ldots M. \tag{5}$$

This idea of using a smaller optimistic ambiguity set has also been used in other DRO approaches, e.g., in [18, Thoerem 6]. We can parametrize the RKHS function f by a wealth of kernel methods, such as the random Fourier features approximation $f(x) \approx= \sum_{i=1}^N \alpha^\top \hat{\phi}_i(x)$ for large scale learning [23]. Note (5) is *convex* in the decision variable θ and f if $l(\theta,\xi)$ in convex in θ.

Distributionally robust solution to uncertain least squares We consider a robust least squares problem adapted from [11], which demonstrated an historically important application of RO to statistical learning. 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 [11]. 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 (5). 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 2 (a) 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 2. Recall it is an over-estimator of the loss $l(\theta,\cdot)$. We then solve the inner moment problem 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.



(a) (b)

Figure 2: **(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.

5. Other related work

The authors of [10] proposed variational approximations to marginal DRO to treat covariate shift in supervised learning. The authors of [37] used kernel mean embedding for the inner moment problem. The work of [30] used insights from DRO to motivate a regularizer for kernel ridge regression. DRO has been also applied to Bayesian optimization in [17, 24], where the latter work used MMD ambiguity sets of distributions over discrete spaces. To the best of our knowledge, no existing work has explored the results, such as generalized ambiguity set constructions in Table 1, as well as a bound in the form of Theorem 1 for general loss functions.

6. Discussion

The compactness assumption on $\mathcal{X}$ may be further extended, just like universality can be extended to non-compact domains [29]. Choosing ϵ in this paper can be further motivated using kernel statistical testing [13] 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 for large scale learning.

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. In Serguei Netessine, Douglas Shier, and Harvey J. Greenberg, editors, *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, October 2019. ISBN 978-0-9906153-3-0. doi: 10.1287/educ.2019.0198.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and Regularization of Support Vector Machines. page 26.

[36] 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.

[37] 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.

[38] 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.