Appendix: Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation

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

Theorems	Generalized Duality Theorem 3.1 Strong duality of the inner moment problem Proposition A.1 Interpolation property Proposition 3.1.3 Complementarity condition Lemma A.2 Robust representer theorem Proposition B.1 IPM-DRO duality Corollary 3.1.1 Kernel DRO as stochastic optimization with expectation constraint Corollary 3.1.2
Formulations	Kernel DRO primal (P) (2), dual (D) (4) IPM-DRO primal (3), dual (5) Formulations for various RKHS ambiguity sets Table 1, 3 Stochastic program with expectation constraint formulation of Kernel DRO (6),(9) Program to compute worst-case distributions (23),(24),(25) Kernel DRO convex program by the discretization of SIP (7) Kernel conditional value-at-risk (8)

In general, we refer to standard texts in optimization (Boyd et al., 2004; Shapiro et al., 2014; Ben-Tal et al., 2009), convex analysis (Rockafellar, 1970; Barvinok, 2002), and functional analysis (Conway, 2019) 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 $\mathcal{K}^* := \{h\colon \int h \,\mathrm{d}m \geq 0, \forall m \in \mathcal{K}, h \;\text{ measurable}\}$. 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 the Generalized Duality Theorem 3.1

We now derive our key result for Kernel DRO — the Generalized Duality Theorem, in Theorem 3.1. Let us first consider the inner moment problem of (2)

$$\sup_{P\in\mathcal{P},\mu\in\mathcal{C}} \int l\; dP \quad \text{subject to } \int \phi\; dP = \mu, \tag{11}$$

where we suppress θ in $l(\theta, \cdot)$ as we fix it for the moment. (11) generalizes the *problem of moments* 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 A.1 (Strong dual to (11))**.** *Under Assumption 3.1,* (11) *is equivalent to solving*

$$\begin{aligned} \min_{f_0\in\mathbb{R},f\in\mathcal{H}} \quad & \delta^*_{\mathcal{C}}(f) + f_0 \\ \text{subject to} \quad & l(\xi) \leq f_0 + f(\xi),\ \forall \xi \in \mathcal{X} \end{aligned} \tag{12}$$

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

Using Proposition A.1, we can reformulate the inner moment problem in (2) to obtain Theorem 3.1. We now prove this generalized duality result for the inner moment problem in Proposition A.1. 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 $P \in \operatorname{co}(\mathcal{P})$. To constrain P to still be a probability measure, we impose $\int 1\ dP(x) = 1$, which results in the primal problem equivalent to (11)

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

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.

$$\begin{aligned}\mathcal{L}(P, \mu;\ f, f_0) &= \int l\ dP - \delta_{\mathcal{C}}(\mu) + \langle \mu - \int \phi\ dP, f\rangle_{\mathcal{H}} + f_0(1 - \int 1\ dP) \\ &= \int l\ dP - \delta_{\mathcal{C}}(\mu) + \langle \mu, f\rangle_{\mathcal{H}} - \int f\ dP + f_0 - \int f_0\ dP \\ &= \int l - f - f_0\ dP + (\langle \mu, f\rangle_{\mathcal{H}} - \delta_{\mathcal{C}}(\mu)) + f_0. \end{aligned} \tag{13}$$

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 D.1, this conic constraint is equivalent to the constraint of (12), $l(\xi) \le f_0 + f(\xi),\ \forall \xi \in \mathcal{X}$.

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. □

Strong duality can potentially be adapted from the strong duality result of moment problem, e.g., (Shapiro, 2001). 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 ((Ben-Tal et al., 2009) Theorem A.2.1) to infinite dimensions. Figure 4 illustrates the idea of the proof.

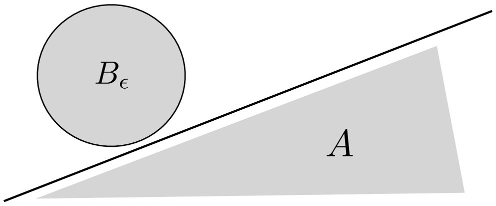


Figure 4: Illustration of the strong duality proof that uses a separating hyperplane. See the proof for detailed descriptions.

Proof. We assume the dual optimal value of (12) 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 $\mathbb{R} \times \mathcal{H} \times \mathbb{R}$

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

where co again denotes conic hull.

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

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

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 $x_3 = \int l \, dP' \le (P) < t + \epsilon = x_3$, 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

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

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

By the separation theorem, (see, e.g., (Barvinok, 2002) 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

$$\begin{gathered} f_0 + \langle f, \mu \rangle_\mathcal{H} + \tau(t + \epsilon) < s, \\ f_0 \int 1 \, dP + \langle f, \int \phi dP \rangle_\mathcal{H} + \tau \int l \, dP > s. \end{gathered}$$

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

$$\begin{gathered} f_0 + \langle f, \mu \rangle_\mathcal{H} + \tau(t + \epsilon) < 0, \forall \mu \in \mathcal{C}, \\ f_0 \int 1 \, dP + \langle f, \int \phi dP \rangle_\mathcal{H} + \tau \int l \, dP \ge 0, \forall P \in \mathrm{co}(\mathcal{P}). \end{gathered} \tag{14}$$

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 D) , the primal optimum is attained by the extreme value theorem. Suppose P^* is a primal optimal solution, from the second inequality of (14),

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

Using this and the first inequality of (14), we obtain $\tau < 0$. Without loss of generality, we let $\tau = -1$.

From the second inequality of (14), we have

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

This tells us that f_0, f is a feasible dual solution because it satisfies the semi-infinite constraint in (12).

By the first inequality of (14),

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

where the left-hand side is precisely the dual objective in (12). This implies $(D) \le (P) + \epsilon$. By weak duality, $(D) \ge (P)$. Therefore, strong duality holds. □

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.

Table 3: Robust counterpart formulations of Kernel DRO.

RKHS ambiguity set $\mathcal{C}$	Robust counterpart formulation
norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon\}$	$f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$
convex hull $\mathcal{C} = \operatorname{conv}\{\mathcal{C}_1, \dots, \mathcal{C}_N\}$ (same under closure 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 SVMs/scenario opt. (Calafiore and Campi, 2006))
Minkowski sum $\sum_{i=1}^N C_i$	$f_0 + \sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f)$
example: $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ $\mathcal{C}_1 = \{\mu \colon \|\mu\|_{\mathcal{H}} \le \epsilon\}$ $\mathcal{C}_2 = \operatorname{conv}\{\phi(\xi_1), \dots, \phi(\xi_N)\}$	$f_0 + \max_i f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$
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 $\mathcal{C} = \{\alpha\mu_{\hat{P}} + (1-\alpha)\mu_Q : \mu_Q \in \mathcal{C}_Q\}$	$f_0 + \frac{\alpha}{N}\sum_{i=1}^N f(\xi_i) + (1-\alpha)\delta^*_{\mathcal{C}_Q}(f)$
Intersection $\mathcal{C} = \cap_{i=1}^N \mathcal{C}_i$	$f_0 + \sum_{i=1}^N \delta^*_{\mathcal{C}_i}(f_i),\ \sum_{i=1}^N f_i = f$
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: $\mathcal{C}_i = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon_i\}$	$f_0 + \frac{1}{N}\sum_{i=1}^N \sum_{i=j}^N f_i(\xi_j) + \epsilon \sum_{i=1}^N \|f_i\|_{\mathcal{H}_i}$
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 (Ben-Tal et al., 2009))

Finally, we summarize the results above to prove the Kernel DRO Generalized Duality Theorem 3.1

Proof. Theorem (3.1) is obtained by reformulating the inner moment problem in (2) using the strong duality result in Proposition A.1, i.e.,

$$\begin{aligned}\min_\theta \sup_{P,\mu} & \left\{ \int l(\theta, \xi)\ dP(\xi) \colon \int \phi\ dP = \mu, P \in \mathcal{P}, \mu \in \mathcal{C} \right\} \\ &= \min_\theta \min_{f_0 \in \mathbb{R}, f \in \mathcal{H}} \left\{ f_0 + \delta^*_{\mathcal{C}}(f) : l(\theta, \xi) \le f_0 + f(\xi),\ \forall \xi \in \mathcal{X} \right\}, \quad (15)\end{aligned}$$

which results in formulation (4). □

A.2 Table 3 deriving formulations for various choices of RKHS ambiguity set $\mathcal{C}$

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

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

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

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 ambiguity set) The result for convex hull follows from standard support function calculus. If the ambiguity set $\mathcal{C}$ is described by the polytope $\operatorname{conv}\{\phi(\xi_1), \ldots, \phi(\xi_N)\}$, 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 [1] . The equivalence to the scenario approach in (Calafiore and Campi, 2006) 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, intersection) Those cases follow directly from the support function calculus; cf. (Ben-Tal et al., 2015).

(Kernel DRO with multiple kernels) Let us consider multiple ambiguity sets from different RKHSs. Suppose $\mathcal{H}_1, \ldots, \mathcal{H}_{N_h}$ are RKHSs associated with feature maps $\phi_1, \ldots, \phi_{N_h}$. Let $\mathcal{C}_1, \ldots, \mathcal{C}_{N_h}$ be the ambiguity sets in the respective RKHSs. Kernel DRO formulation with multiple kernels is given By

$$\min_{\theta} \sup_{P,\mu} \left\{ \int l(\theta, \xi)\ dP(\xi) \colon \int \phi_i\ dP = \mu_i, P \in \mathcal{P}, \mu_i \in \mathcal{C}_i, i = 1 \ldots N_h \right\}, \tag{16}$$

Using the same proof as Proposition A.1, we have the Kernel DRO reformulation

$$\begin{aligned} \min_{\theta, f_0\in\mathbb{R}, f_i\in\mathcal{H}_i} \quad & f_0 + \sum_{i=1}^{N} \delta^*_{\mathcal{C}_i}(f_i) \\ \text{subject to} \quad & l(\theta, \xi) \leq f_0 + \sum_{i=1}^{N} f_i(\xi),\ \forall \xi \in \mathcal{X} \end{aligned} . \tag{17}$$

Hence we obtain the formulation in Table 3.

(Singleton ambiguity 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 ambiguity 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 (4) is reduced to

$$\begin{aligned} \min_{\theta, f_0\in\mathbb{R}} \quad & f_0 \\ \text{subject to} \quad & l(\theta, \xi) \leq f_0,\ \forall \xi \in \mathcal{X} \end{aligned} \tag{18}$$

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

A.3 Complementarity condition and proof

Lemma A.2 (Complementarity condition)**.** *Let P^*, f^*, f_0^* be a set of optimal primal-dual solutions of (P) and (D), then*

$$\int l - f^* - f_0^*\ dP^* = 0, \quad \delta^*_{\mathcal{C}}(f^*) = \int f^*\ dP^* \tag{19}$$

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

$$\int \frac{f^*}{\|f^*\|_{\mathcal{H}}}\ d(P^* - \hat{P}) = \operatorname{MMD}(P^*, \hat{P}), \tag{20}$$

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. (Shapiro, 2001). The following is a straightforward proof.

[1] 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.

[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.

Proof. Plug P^*, f^*, f_0^* into Lagrangian (13),

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

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.4 Proof of Proposition 3.1.3 (Interpolation property)

Proof. Since f_0^*, f^* is a solution to the inner moment problem of Kernel DRO (12), we have $l(\theta, \xi) \leq f_0^* + f^*(\xi),\ \forall \xi \in \mathcal{X}$ for any given θ. By the first equation in the complementarity condition (19), we have $\int l - f^* - f_0^*\ dP^* = 0$. Hence the integrand must be zero P^*-a.e. □

A.5 Corollary 3.1.1 IPM-DRO duality

We provide a derivation using a technique alternative to the proof of Proposition A.1.

Proof. We consider the Lagrangian

$$\begin{aligned}\mathcal{L}(P;\lambda) &= \int l\ dP - \lambda(d_{\mathcal{F}}(P,\hat{P}) - \epsilon) \\ &= \int l\ dP - \lambda \sup_{f\in\mathcal{F}} \int f d(P - \hat{P}) + \lambda\epsilon \\ &= \inf_{f\in\mathcal{F}} \int l - \lambda f\ dP + \lambda \int f d\hat{P} + \lambda\epsilon \\ &\leq \inf_{f\in\mathcal{F}} \sup_{\xi\in\mathcal{X}} [l(\xi) - \lambda f(\xi)] + \frac{\lambda}{N}\sum_{i=1}^{N} f(\xi_i) + \lambda\epsilon. \qquad (21)\end{aligned}$$

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 (5). □

A.6 Corollary 3.1.2 Kernel DRO as stochastic optimization with expectation constraint

Using the known relationship between semi-infinite constraint and expectation constraint (see, e.g., (Tadić et al., 2006, Theorem 1)), the SI constraint in (4) is equivalent to the expectation constraint in (6).

B COMPUTATIONAL FORMULATIONS

We now provide practical plug-in formulations for computation. Specifically, we can parametrize the RKHS function f by, e.g., the following methods. We note that the random feature method is well-suited for large scale problems, such as in SFG-DRO applications.

B.1 Random features

Common ways to parametrize an RKHS function include the representer theorem as well as approximations such as the random Fourier features (Rahimi and Recht, 2008). Recall that an RKHS function can be approximated by the finite feature expansion

$$f(\xi) \approx \hat{f}(\xi) = w^\top \hat{\phi}(\xi), \quad k(x,x') \approx \sum_{i=1}^{N} \hat{\phi}_i(x)\hat{\phi}_i(x')$$

where $\{\hat{\phi}_i(x)\}_{i=1}^N$ are the random features, e.g., random Fourier features $\hat{\phi}_i(x) = \cos(w_i x + b_i), w_i \sim \mathrm{N}(0, \sigma^2), b_i \sim \mathrm{Uniform}[0, 2\pi]$. If x is a vector, then $w_i \sim \mathcal{N}(0, I\sigma^2)$, and $w_i x$ is the dot product. See, e.g., (Rahimi and Recht, 2008), for more properties.

One strength of the Generalized Duality Theorem 3.1 is that it does not require the knowledge of the RKHS that the loss l lives in, which is typically not available in non-kernelized models. This enables us to use approximate features for commonly used RKHSs, e.g., random Fourier feature. This is a strength of our Kernel DRO theory.

Note program (7) is a convex optimization problem with the random feature parametrization.

B.2 Distributionally robust version of representer theorem

In program (7), we may parametrize the RKHS function by $f(\cdot) = \sum_{i=1}^N \beta_i k(\xi_i, \cdot) + \sum_{j=1}^M \gamma_j k(\zeta_j, \cdot)$, $\|f\|_{\mathcal{H}} = \sqrt{\alpha^\top K \alpha}$, where $\alpha = (\beta_1, \ldots, \beta_N, \gamma_1, \ldots, \gamma_M)^\top, K = [k(\eta_i, \eta_j)], \eta = (\xi_1, \ldots, \xi_N, \zeta_1, \ldots, \zeta_M)^\top$. We justify this parametrization by the following DRO version of the RKHS representer theorem (Schölkopf et al., 2001).

The intuition of the following result is to restrict Kernel DRO to a smaller ambiguity set of distributions supported on $\{\zeta_i\}_{i=1}^M$ (i.e., replace $P \in \mathcal{P}$ by $P \in \mathcal{P}_M$, an inner approximation depending on M). In this setting, the ambiguity set only contains only distributions supported on (a subset of) ζ_i. Then it suffices to parametrize f in (7) by $f(\cdot) = \sum_{i=1}^N \beta_i k(\xi_i, \cdot) + \sum_{j=1}^M \gamma_j k(\zeta_j, \cdot)$.

Lemma B.1 (Robust representer)**.** *Given data $\{\xi_i\}_{i=1}^N$ and the ambiguity set chosen to be a set of embeddings with the form $\sum_{j=1}^M \alpha_j \phi(\zeta_j)$, for some $0 \le \alpha_j \le 1, \sum_{j=1}^M \alpha_j = 1$, and within the RKHS norm-ball $\mathcal{C} = \{\mu \colon \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \le \epsilon\}$. Then, it suffices to consider the RKHS function of the form $f(\cdot) = \sum_{i=1}^N \beta_i k(\xi_i, \cdot) + \sum_{j=1}^M \gamma_j k(\zeta_j, \cdot)$ for some $\beta_i, \gamma_j \in \mathbb{R}, i = 1...N, j = 1...M$.*

Lemma B.1 states that the expansion points of the RKHS representer in (7) are exactly the support of the empirical and worst-case distributions. It extends the classical RKHS representer theorem (Schölkopf et al., 2001), which uses only the empirical samples as expansion points. The implication is that, to be distributionally robust, we should choose the representers as in Lemma B.1 instead of only using empirical samples. Below is a proof that is similar to the original representer theorem.

Proof. In (7) we consider $f = f_s + f_\perp$, where $f_s = \sum_{i=1}^N \beta_i k(\xi_i, \cdot) + \sum_{j=1}^M \gamma_j k(\zeta_j, \cdot)$ belongs to a subspace of the $\mathcal{H}$ and $f_\perp$ its complement. Plug in $f = f_s + f_\perp$ to (7) and note the orthogonality, we obtained

$$\begin{aligned} \min_{\theta, f_s, f_\perp, f_0} \quad & f_0 + \frac{1}{N} \sum_{i=1}^N f_s(\xi_i) + \epsilon(\|f_s\|_{\mathcal{H}} + \|f_\perp\|_{\mathcal{H}}) \\ \text{subject to} \quad & l(\theta, \zeta_i) \le f_s(\zeta_j) + f_0, \ j = 1 \ldots M. \end{aligned} \tag{22}$$

It suffices to choose $f_\perp = 0$ in this optimization problem. Hence the conclusion follows. □

Remark. Note the existence of a worst case distribution in more general settings is not yet proven. The discussion here is restricted to the setting of (7).

C FURTHER NUMERICAL EXPERIMENT RESULTS

We carry out additional numerical experiments to study Kernel DRO.

C.1 Testing other variants of Kernel DRO

We empirically test the following proposed variants of Kernel DRO.

- Relaxed Kernel DRO formulation (Kernel DRO-relaxed) with constraint hold for only the empirical samples, i.e., $l(\theta, \xi_i) \le f_0 + f(\xi_i), \ i = 1 \ldots N$.

- Unconstrained Kernel DRO using Kernel CVaR in Example 4.1

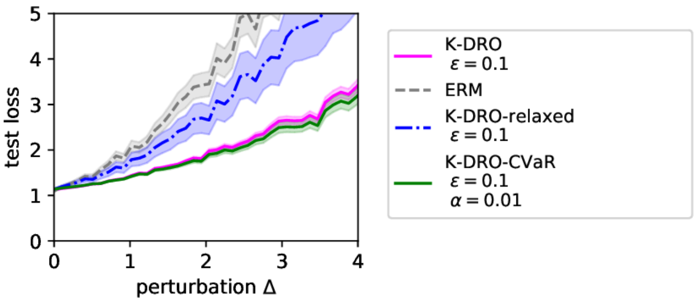


Figure 5: Comparing Kernel DRO-relaxed, Kernel DRO-KCVaR, ERM, and regular Kernel DRO. y-axis limit is adjusted to show the plot. All error bars are in standard error.

We compare Kernel DRO-relaxed with the ERM as well as the regular Kernel DRO. Compared with ERM, Kernel 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 Kernel DRO.

We then test the Kernel DRO using the unconstrained objective given by Kernel CVaR. We observe no significant difference in performance between Kernel DRO-KCVaR (with small chance constraint level α.) and regular Kernel DRO (7).

C.2 Analyzing the generalization behavior

An insight can be obtained by observing the plot of the MMD estimator between the training and test data in Figure 6 (left). As Kernel DRO with $\epsilon = 0.5$ robustified against perturbation less than the level 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 Kernel 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 6 (right). This empirically validates our theoretical results for robustification.

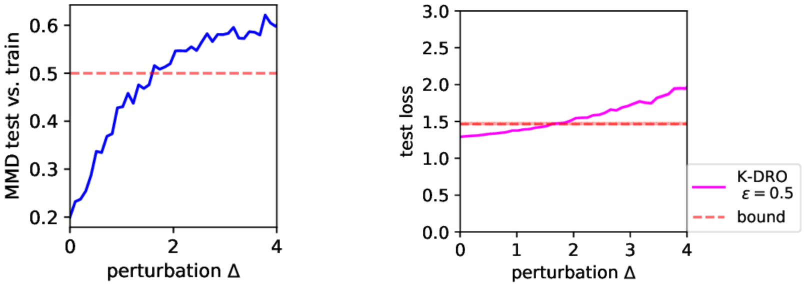
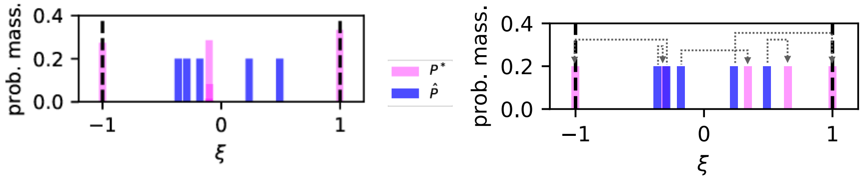


Figure 6: (Left) MMD estimator between the empirical samples and test samples. The level MMD $= 0.5$ is marked in red. (Right) Loss compared to the generalization bound. As the test data falls outside the robustification level ϵ, the loss starts to exceed the generalization bound (red) $f_0 + \frac{1}{N}\sum_{i=1}^N f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$.

C.3 Miscellaneous details for experimental set-up

Robust least squares example. Our experiments are implemented in Python. The convex optimization problems are solved using ECOS or MOSEK interfaced with CVXPY. In the experiments, we chose the bandwidth for the Gaussian kernel using the medium heuristic (Gretton et al., 2012). ϵ in this paper are fixed to constants below 2 for Gaussian kernels. Choosing ϵ can be further motivated by kernel statistical tests (Gretton et al., 2012) and is left for future work.

Sampled ζ_j In applying Kernel DRO using (7), we may obtain 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



(a) Support sampling

(b) Perturbation

Figure 7: Computing the worst-case distribution P^* using **(a)**: (23) samples possible support ζ_j then optimizes w.r.t. weights α. **(b)**: (24) moves the empirical samples directly.

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. For example, in the robust least squares experiment, we sampled the support ζ_i uniformly random from $[-1, 1]$.

Robust learning under adversarial perturbation example. For the MNIST robust classification example, we used a neural network with two hidden layers with 64 units each. For the training of ERM and PGD, we used the ADAM optimization routine implemented in the PyTorch library. In Step 3 of SFG-DRO, we used random Fourier features (Rahimi and Recht, 2008) with 500 features. In Step 5 of SFG-DRO, we used the SA routine from CSA algorithm (Lan and Zhou, 2020). While other SA routines can be used, we prefer the simplicity of CSA in that it does not use a dual variable. We set the threshold and step-size of the CSA algorithm (Lan and Zhou, 2020) to decay at the rate of $\frac{1}{\sqrt{k}}$ as suggested in that paper. We did not attempt further adaptive tuning of the step-sizes or the proposing distribution for ζ (we generate 3000 samples uniformly in Step 2 of SFG-DRO), which may further improve the performance. Parameter (weights of the neural nets) averaging is used for training all models. In the visualization of the predictions in Figure 3d, we perturbed the images by the PGD method (Madry et al., 2019; Madry) based on the ERM loss and linear model. SFG-DRO does not have the knowledge of the perturbation method.

C.4 Computing worst-case distributions

We have proposed Kernel DRO for making the decision θ via reformulation (4). 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 7.

Support sampling. We consider the moment problem (11) 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,

$$\max_{\alpha \in S_M} \sum_{j=1}^M \alpha_i l(\theta, \zeta_j) \quad \text{subject to} \quad \left\| \sum_{j=1}^M \alpha_i \phi(\zeta_j) - \frac{1}{N} \sum_{i=1}^N \phi(\xi_i) \right\|_{\mathcal{H}} \leq \epsilon. \tag{23}$$

(23) 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 (Pólik and Terlaky, 2007) (cf. appendix). Alternatively, (23) can be directly handled by convex solvers for a given θ. Note this approach was previously used in solving the problem of moments in (Zhu et al., 2020).

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,

$$\max_{\substack{d_i, i=1\ldots N,\\ \xi_i + d_i \in \mathcal{X}}} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i + d_i) \quad \text{subject to} \quad \left\| \frac{1}{N} \sum_{i=1}^N (\phi(\xi_i + d_i) - \phi(\xi_i)) \right\|_{\mathcal{H}} \leq \epsilon. \tag{24}$$

[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.

Compared with (23), (24) 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, (24) may become a nonlinear program. However, its feasibility is guaranteed since it can always be initialized with a feasible solution $d_i = 0$.

We now empirically examine the *support sampling* method (23) and *perturbation* method (24) to recover the worst-case distribution. Since both programs (23) and (24) search for the worst-case distribution within a subset of all distributions, their optimal values lower-bound the true worst-case risk (P) in (11), i.e., with finite samples, they are optimistic bound.

Under the experimental setting as in Figure 3b, we ran Kernel DRO with fewer empirical samples ($N = 5$). After we obtain the Kernel DRO solution θ^*, we plug it into (23) and (24), respectively, to compute the worst-case distribution P^*. Figure 7 plots the results. Note (23) is a convex optimization problem, while (24) results in a nonlinear program (with Gaussian kernel). Nonetheless, we solve it with an always-feasible initialization $d_i = 0$.

C.5 SDP dual via S-lemma

We consider a discretized version of the primal moment problem in (23) where the distribution is constrained to be a discrete distribution. We rewrite (23) as a quadratically constrained program using the plug-in estimator of MMD,

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

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 (Pólik and Terlaky, 2007), this program can be reformulated as the following SDP,

$$\begin{aligned} \min_{\lambda\ge 0, x, y\ge 0, t} \quad & t \\ \text{subject to} \quad & \begin{bmatrix} \lambda P & -\lambda q - \frac{1}{2}(l + x\cdot \mathbb{1} + y) \\ (-\lambda q - \frac{1}{2}(l + x\cdot \mathbb{1} + y))^\top & t - \lambda\epsilon^2 + x + \lambda r \end{bmatrix} \ge 0, \end{aligned} \tag{25}$$

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

D SUPPORTING LEMMAS

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

D.1 Reducing conic constraint to infinite constraint

To derive the semi-infinite constraint in (12), we need a standard result from the literature of the moment problem. We give a self-contained proof below.

Lemma D.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*

$$l(\theta, \xi) \le f_0 + f(\xi),\ \forall \xi \in \mathcal{X}. \tag{26}$$

Proof. “ $\Longrightarrow$ ”: Let us consider the set of all Dirac measures on $\mathcal{X}$, $\mathcal{D} := \{\delta_\xi \colon \xi \in \mathcal{X}\}$. 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 $l - f_0 - f \in -K^*$. □

Note an extension of this result to generating classes other than all Dirac measures $\mathcal{D}$ can be proved using Choquet theory, cf. (Shapiro et al., 2014, Proposition 6.66) (Popescu, 2005, Lemma 3.1), as well as in (Shapiro, 2001; Rogosinski).

D.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 D.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. □

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

Lemma D.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. □

Recall that we denote the feasible set of probability measures, i.e., ambiguity set, for primal Kernel 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 D.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. □