Distributionally Robust Optimization and Generalization in Kernel Methods

Matthew Staib  
MIT CSAIL  
mstaib@mit.edu

Stefanie Jegelka  
MIT CSAIL  
stefje@csail.mit.edu

Abstract

Distributionally robust optimization (DRO) has attracted attention in machine learning due to its connections to regularization, generalization, and robustness. Existing work has considered uncertainty sets based on $\phi$-divergences and Wasserstein distances, each of which have drawbacks. In this paper, we study DRO with uncertainty sets measured via maximum mean discrepancy (MMD). We show that MMD DRO is roughly equivalent to regularization by the Hilbert norm and, as a byproduct, reveal deep connections to classic results in statistical learning. In particular, we obtain an alternative proof of a generalization bound for Gaussian kernel ridge regression via a DRO lens. The proof also suggests a new regularizer. Our results apply beyond kernel methods: we derive a generically applicable approximation of MMD DRO, and show that it generalizes recent work on variance-based regularization.

1 Introduction

Distributionally robust optimization (DRO) is an attractive tool for improving machine learning models. Instead of choosing a model $f$ to minimize empirical risk $\mathbb{E}_{x \sim \hat{P}_n}[\ell_f(x)] = \frac{1}{n} \sum_i \ell_f(x_i)$, an adversary is allowed to perturb the sample distribution within a set $\mathcal{U}$ centered around the empirical distribution $\hat{P}_n$. DRO seeks a model that performs well regardless of the perturbation: $\inf_f \sup_{Q \in \mathcal{U}} \mathbb{E}_{x \sim Q}[\ell_f(x)]$. The induced robustness can directly imply generalization: if the data that forms $\hat{P}_n$ is drawn from a population distribution $P$, and $\mathcal{U}$ is large enough to contain $P$, then we implicitly optimize for $P$ too and the DRO objective value upper bounds out of sample performance. More broadly, robustness has gained attention due to adversarial examples [17, 39, 26]; indeed, DRO generalizes robustness to adversarial examples [33, 36].

In machine learning, the DRO uncertainty set $\mathcal{U}$ has so far always been defined as a $\phi$-divergence ball or Wasserstein ball around the empirical distribution $\hat{P}_n$. These choices are convenient, due to a number of structural results. For example, DRO with $\chi^2$-divergence is roughly equivalent to regularizing by variance [18, 22, 51], and the worst case distribution $Q \in \mathcal{U}$ can be computed exactly in $O(n \log n)$ [37]. Moreover, DRO with Wasserstein distance is asymptotically equivalent to certain common norm penalties [15], and the worst case $Q \in \mathcal{U}$ can be computed approximately in several cases [28, 14]. These structural results are key, because the most challenging part of DRO is solving (or bounding) the DRO objective.

However, there are substantial drawbacks to these two types of uncertainty sets. Any $\phi$-divergence uncertainty set $\mathcal{U}$ around $\hat{P}_n$ contains only distributions with the same (finite) support as $\hat{P}_n$. Hence, the population $P$ is typically not in $\mathcal{U}$, and so the DRO objective value cannot directly certify out of sample performance. Wasserstein uncertainty sets do not suffer from this problem. But, they are more computationally expensive, and the above key results on equivalences and computation need nontrivial assumptions on the loss function and the specific ground distance metric used.

In this paper, we introduce and develop a new class of DRO problems, where the uncertainty set \( \mathcal{U} \) is defined with respect to the maximum mean discrepancy (MMD) \([19]\), a kernel-based distance between distributions. MMD DRO complements existing approaches and avoids some of their drawbacks, e.g., unlike \( \phi \)-divergences, the uncertainty set \( \mathcal{U} \) will contain \( \mathbb{P} \) if the radius is large enough.

First, we show that MMD DRO is roughly equivalent to regularizing by the Hilbert norm \( \| \ell_f \|_H \) of the loss \( \ell_f \) (not the model \( f \)). While, in general, \( \| \ell_f \|_H \) may be difficult to compute, we show settings in which it is tractable. Specifically, for kernel ridge regression with a Gaussian kernel, we prove a bound on \( \| \ell_f \|_H \) that, as a byproduct, yields generalization bounds that match (up to a small constant) the standard ones. Second, beyond kernel methods, we show how MMD DRO generalizes variance-based regularization. Finally, we show how MMD DRO can be efficiently approximated empirically, and in fact generalizes variance-based regularization.

Overall, our results offer deeper insights into the landscape of regularization and robustness approaches, and a more complete picture of the effects of different divergences for defining robustness. In short, our contributions are:

1. We prove fundamental structural results for MMD DRO, and its rough equivalence to penalizing by the Hilbert norm of the loss.
2. We give a new generalization proof for Gaussian kernel ridge regression by way of DRO. Along the way, we prove bounds on the Hilbert norm of products of functions that may be of independent interest.
3. Our generalization proof suggests a new regularizer for Gaussian kernel ridge regression.
4. We derive a computationally tractable approximation of MMD DRO, with application to general learning problems, and we show how the aforementioned approximation generalizes variance regularization.

2 Background and Related Work

Distributionally robust optimization (DRO) \([16, 2]\), introduced by Scarf \([32]\), asks to not only perform well on a fixed problem instance (parameterized by a distribution), but simultaneously for a range of problems, each determined by a distribution in an uncertainty set \( \mathcal{U} \). This results in more robust solutions. The uncertainty set plays a key role: it implicitly defines the induced notion of robustness.

The DRO problem we address asks to learn a model \( f \) that solves

\[
\begin{align*}
\text{(DRO)} \quad & \inf_f \sup_{Q \in \mathcal{U}} \mathbb{E}_{x \sim Q}[\ell_f(x)],
\end{align*}
\]

where \( \ell_f(x) \) is the loss incurred under prediction \( f(x) \).

In this work, we focus on data-driven DRO, where \( \mathcal{U} \) is centered around an empirical sample \( \hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \), and its size is determined in a data-dependent way. Data-driven DRO yields a natural approach for certifying out-of-sample performance.

**Principle 2.1** (DRO Generalization Principle). Fix any model \( f \). Let \( \mathcal{U} \) be a set of distributions containing \( \mathbb{P}_n \). Suppose \( \mathcal{U} \) is large enough so that, with probability \( 1 - \delta \), \( \mathcal{U} \) contains the population \( \mathbb{P} \). Then with probability \( 1 - \delta \), the population loss \( \mathbb{E}_{x \sim \mathbb{P}}[\ell_f(x)] \) is bounded by

\[
\mathbb{E}_{x \sim \mathbb{P}}[\ell_f(x)] \leq \sup_{Q \in \mathcal{U}} \mathbb{E}_{x \sim Q}[\ell_f(x)].
\]

Essentially, if the uncertainty set \( \mathcal{U} \) is chosen appropriately, the corresponding DRO problem gives a high probability bound on population performance. The two key steps in using Principle 2.1 are 1. arguing that \( \mathcal{U} \) actually contains \( \mathbb{P} \) with high probability (e.g. via concentration); 2. solving the DRO problem on the right hand side, or an upper bound thereof.

In practice, \( \mathcal{U} \) is typically chosen as a ball of radius \( \epsilon \) around the empirical sample \( \hat{\mathbb{P}}_n \): \( \mathcal{U} = \{ Q : d(Q, \hat{\mathbb{P}}_n) \leq \epsilon \} \). Here, \( d \) is a discrepancy between distributions, and is of utmost significance: the choice of \( d \) determines how large \( \epsilon \) must be, and how tractable the DRO problem is.

In machine learning, two choices of the divergence \( d \) are prevalent, \( \phi \)-divergences \([1, 11, 22]\), and Wasserstein distance \([28, 53, 6]\). The first option, \( \phi \)-divergences, have the form \( d_\phi(\mathbb{P}, Q) = \int \log \frac{d \mathbb{P}}{d Q} \ d\mathbb{P} \).
well as fast convergence properties, which are of chief importance in our work.

The main focus of this paper is Distributionally Robust Optimization where the uncertainty set is

defined via the MMD distance

\[ \text{MMD}(P, Q) := \sup_{f \in \mathcal{H}} \mathbb{E}_{x \sim P} [f(x)] - \mathbb{E}_{x \sim Q} [f(x)]. \]  

Fact 2.1. Define the mean embedding \( \mu_P \) of the distribution \( P \) by \( \mu_P = \mathbb{E}_{x \sim P} [k(x, \cdot)] \). Then the MMD between distributions \( P \) and \( Q \) can be equivalently written

\[ \text{MMD}(P, Q) = \| \mu_P - \mu_Q \|_H. \]  

MMD and (more generally) kernel mean embeddings have been used in many applications, particularly in two- and one-sample tests [19, 21, 25, 9] and in generative modeling [12, 24, 38, 5]. We refer the interested reader to the monograph by Muandet et al. [30]. MMD admits efficient estimation, as well as fast convergence properties, which are of chief importance in our work.

Further related work. Beyond \( \phi \)-divergences and Wasserstein distances, work in operations research has considered DRO problems that capture uncertainty in moments of the distribution, e.g. [10]. These approaches typically focus on first- and second-order moments; in contrast, an MMD uncertainty set allows high order moments to vary, depending on the choice of kernel.

Robust and adversarial machine learning have strong connections to our work and DRO more generally. Robustness to adversarial examples [39, 17], where individual inputs to the model are perturbed in a small ball, can be cast as a robust optimization problem [26]. When the ball is a norm ball, this robust formulation is a special case of Wasserstein DRO [35, 36]. Xu et al. [42] study the connection between robustness and regularization in SVMs, and perturbations within a (possibly Hilbert) norm ball. Unlike our work, their results are limited to SVMs instead of general loss minimization. Moreover, they consider only perturbation of individual data points instead of shifts in the entire distribution. Bietti et al. [4] show that many regularizers used for neural networks can also be interpreted in light of an appropriately chosen Hilbert norm [3].

3 Generalization bounds via MMD DRO

The main focus of this paper is Distributionally Robust Optimization where the uncertainty set is defined via the MMD distance \( d_{\text{MMD}} \):

\[ \inf \sup_{Q : d_{\text{MMD}}(Q, \hat{P}_n) \leq \epsilon} \mathbb{E}_{x \sim Q} [\ell_f(x)]. \]
One motivation for considering MMD in this setting are its possible implications for Generalization. Recall that for the DRO Generalization Principle [2.1] to apply, the uncertainty set $\mathcal{U}$ must contain the population distribution with high probability. To ensure this, the radius of $\mathcal{U}$ must be large enough. But, the larger the radius, the more pessimistic is the DRO minimax problem, which may lead to over-regularization. This radius depends on how quickly $d_{\text{MMD}}(P, P_n)$ shrinks to zero, i.e., on the empirical accuracy of the divergence.

In contrast to Wasserstein distance, which converges at a rate of $O(n^{-1/d})$ [13], MMD between the empirical sample $P_n$ and population $P$ shrinks as $O(n^{-1/2})$:

**Lemma 3.1 (Modified from [30], Theorem 3.4).** Suppose that $k(x, x) \leq M$ for all $x$. Let $\hat{P}_n$ be an $n$ sample empirical approximation to $P$. Then with probability $1 - \delta$,

$$d_{\text{MMD}}(P, \hat{P}_n) \leq 2\sqrt{\frac{M}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}. \quad (6)$$

The constant $M$ is dimension-independent for many common universal kernels, e.g. Gaussian, Laplace, and Matern kernels. With Lemma 3.1 in hand, we conclude a simple high probability bound on out-of-sample performance:

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

$$\mathbb{E}_{x \sim P}[\ell_f(x)] \leq \sup_{Q: d_{\text{MMD}}(Q, \hat{P}_n) \leq \epsilon} \mathbb{E}_{x \sim Q}[\ell_f(x)]. \quad (7)$$

We refer to the right hand side as the DRO adversary’s problem. In the next section we develop results that enable us to bound its value, and consequently bound the DRO problem [9].

### 3.1 Bounding the DRO adversary’s problem

The DRO adversary’s problem seeks the distribution $Q$ in the MMD ball so that $\mathbb{E}_{x \sim Q}[\ell_f(x)]$ is as high as possible. Reasoning about the optimal worst-case $Q$ is the main difficulty in DRO. With MMD, we take two steps for simplification. First, instead of directly optimizing over distributions, we optimize over their mean embeddings in the Hilbert space (described in Fact 2.1). Second, while the adversary’s problem (7) makes sense for general $Q$, often $k$ is a universal kernel, meaning under mild conditions $\ell_f$ can be approximated arbitrarily well by a member of $\mathcal{H}$ [30, Definition 3.3].

With the additional assumption that $\ell_f \in \mathcal{H}$, the risk $\mathbb{E}_{x \sim P}[\ell_f(x)]$ can also be written as $\langle \ell_f, \mu_P \rangle_{\mathcal{H}}$. Then we obtain

$$\sup_{Q: d_{\text{MMD}}(Q, P) \leq \epsilon} \mathbb{E}_{x \sim Q}[\ell_f(x)] \leq \sup_{\mu_Q \in \mathcal{H}: \|\mu_Q - \mu_P\|_{\mathcal{H}} \leq \epsilon} \langle \ell_f, \mu_Q \rangle_{\mathcal{H}}, \quad (8)$$

where we have an inequality because not every function in $\mathcal{H}$ is the mean embedding of some probability distribution. If $k$ is a characteristic kernel [30, Definition 3.2], the mapping $P \mapsto \mu_P$ is injective. In this case, the only looseness in the bound is due to discarding the constraints that $Q$ integrates to one and is nonnegative. However it is difficult to constrain the mean embedding $\mu_Q$ in this way as it is a function.

The mean embedding form of the problem is simpler to work with, and leads to further interpretations.

**Theorem 3.1.** Let $\ell_f, \mu_P \in \mathcal{H}$. We have the following equality:

$$\sup_{\mu_Q \in \mathcal{H}: \|\mu_Q - \mu_P\|_{\mathcal{H}} \leq \epsilon} \langle \ell_f, \mu_Q \rangle_{\mathcal{H}} = \langle \ell_f, \mu_P \rangle_{\mathcal{H}} + \epsilon \|\ell_f\|_{\mathcal{H}} = \mathbb{E}_{x \sim P}[\ell_f(x)] + \epsilon \|\ell_f\|_{\mathcal{H}}. \quad (9)$$

In particular, the adversary’s optimal solution is $\mu_Q^* = \mu_P + \frac{\epsilon}{\|\ell_f\|_{\mathcal{H}}} \ell_f$.

Combining Theorem 3.1 with equation (8) yields our main result for this section:

**Corollary 3.2.** Let $\ell_f \in \mathcal{H}$, let $P$ be a probability distribution, and fix $\epsilon > 0$. Then,

$$\sup_{Q: d_{\text{MMD}}(P, Q) \leq \epsilon} \mathbb{E}_{x \sim Q}[\ell_f(x)] \leq \mathbb{E}_{x \sim P}[\ell_f(x)] + \epsilon \|\ell_f\|_{\mathcal{H}} \quad \text{and therefore} \quad (10)$$

$$\inf_f \sup_{Q: d_{\text{MMD}}(P, Q) \leq \epsilon} \mathbb{E}_{x \sim Q}[\ell_f(x)] \leq \inf_f \mathbb{E}_{x \sim P}[\ell_f(x)] + \epsilon \|\ell_f\|_{\mathcal{H}}. \quad (11)$$
Combining Corollary 3.2 with Corollary 3.1 shows that minimizing the empirical risk plus a norm on \( \ell_f \) leads to a high probability bound on out-of-sample performance. This result is similar to results that equate Wasserstein DRO to norm regularization. For example, Gao et al. [15] show that under appropriate assumptions on \( \ell_f \), DRO with a \( p \)-Wasserstein ball is asymptotically equivalent to
\[
\mathbb{E}_{x \sim \mathcal{P}_n}[\ell_f(x)] + \epsilon \|\nabla x \ell_f\|_{\mathcal{P}_n,q},
\]
where \( \|\nabla_x \ell_f\|_{\mathcal{P}_n,q} = \left( \frac{1}{n} \sum_{i=1}^n \|\nabla_x \ell_f(x_i)\|^q \right)^{1/q} \) measures a kind of \( q \)-norm average of \( \|\nabla_x \ell_f(x_i)\| \) at each data point \( x_i \) (here \( q \) is such that \( 1/p + 1/q = 1 \), and \( \| \cdot \|_\ast \) is the dual norm of the metric defining the Wasserstein distance).

There are a few key differences between our result and that of Gao et al. [15]. First, the norms are different. Second, their result penalizes only the gradient of \( \ell_f \), while ours penalizes \( \ell_f \) directly. Third, except for certain special cases, the Wasserstein results cannot serve as a true upper bound; there are higher order terms that only shrink to zero as \( \epsilon \to 0 \). These higher order terms may not be so small: in high dimension \( d \), the radius \( \epsilon \) of the uncertainty set needed so that \( \mathbb{P} \in \mathcal{U} \) shrinks very slowly, as \( O(n^{-1/d}) \) [13].

**Remark 3.1.** Theorem 3.1 and Corollary 3.2 require that \( \ell_f \) is in the RKHS \( \mathcal{H} \). Though this may seem restrictive, if the kernel \( k \) is universal, as is the case for many kernels used in practice such as Gaussian and Laplace kernels, we can readily extend our results to all bounded continuous functions. Suppose \( \ell_f \) is a bounded continuous function on a compact metric space \( \mathcal{X} \). By definition (e.g., [30]), Definition 3.3, if \( k \) is a universal kernel on \( \mathcal{X} \), then for any \( \epsilon > 0 \), there is some \( \ell' \in \mathcal{H} \) with
\[
\sup_{x \in \mathcal{X}} |\ell_f(x) - \ell'(x)| < \epsilon.
\]
It follows that for any measure \( \mathbb{P} \), we can bound the expectation of \( \ell_f(x) \) by that of \( \ell' \).

**4 Connections to kernel ridge regression**

After applying Corollary 3.2, we are interested in solving:
\[
\inf_{f} \mathbb{E}_{x \sim \mathcal{P}_n}[\ell_f(x)] + \epsilon \|\ell_f\|_\mathcal{H}.
\] (12)

Here, we penalize our model \( f \) by \( \|\ell_f\|_\mathcal{H} \). This looks similar to but is very different from the usual penalty \( \|f\|_\mathcal{H} \) in kernel methods. In fact, Hilbert norms of function compositions such as \( \ell_f \) pose several challenges. For example, \( f \) and \( \ell_f \) may not belong to the same RKHS \( \mathcal{H} \) – it is not hard to construct counterexamples, even when \( \ell \) is merely quadratic. So, the objective (12) is not yet computable.

Despite these challenges, we next develop tools that will allow us to bound \( \|\ell_f\|_\mathcal{H} \) and use it as a regularizer. These tools may be of independent interest to bound RKHS norms of composite functions (e.g., for settings as in [41]). Due to the difficulty of this task, we specialize to Gaussian kernels \( k_\sigma(x,y) = \exp(-\|x-y\|^2/(2\sigma^2)) \). Since we will need to take care regarding the bandwidth \( \sigma \), we explicitly write it out for the inner product \( \langle \cdot, \cdot \rangle_\sigma \) and norm \( \| \cdot \|_\sigma \), of the corresponding RKHS \( \mathcal{H}_\sigma \).

To make the setting concrete, consider kernel ridge regression, with Gaussian kernel \( k_\sigma \). As usual, we assume there is a simple target function \( h \) that fits our data: \( h(x_i) = y_i \). Then the loss \( \ell_f \) of \( f \) is
\[
\ell_f(x) = (f(x) - h(x))^2,
\]
so we wish to solve
\[
\inf_{f} \mathbb{E}_{x \sim \mathcal{P}_n}[(f(x) - h(x))^2] + \epsilon \|f - h\|^2_\sigma.
\] (13)

**4.1 Bounding norms of products**

To bound \( \|(f-h)^2\|_\sigma \), it will suffice to bound RKHS norms of products. The key result for this subsection is the following deceptively simple-looking bound:

**Theorem 4.1.** Let \( f, g \in \mathcal{H}_\sigma \), that is, the RKHS corresponding to the Gaussian kernel \( k_\sigma \) of bandwidth \( \sigma \). Then,
\[
\|fg\|_{\sigma/\sqrt{2}} \leq \|f\|_\sigma \|g\|_\sigma.
\]

Indeed, there are already subtleties: if \( f, g \in \mathcal{H}_\sigma \), then, to discuss the norm of the product \( fg \), we need to decrease the bandwidth from \( \sigma \) to \( \sigma/\sqrt{2} \).

We prove Theorem 4.1 via two steps. First, we represent the functions \( f, g \), and \( fg \) exactly in terms of traces of certain matrices. This step is highly dependent on the specific structure of the Gaussian kernel. Then, we can apply standard trace inequalities. Proofs of both results are given in Appendix B.

5
With the help of Theorem 4.1, we can develop DRO-based bounds for actual learning problems. In these intermediate results in hand, we can prove the main bound of interest:

\[ \|fg\|_\sigma^2 = \text{tr}(A^2B^2), \quad \|f\|_\sigma^2 = \text{tr}(A^2), \quad \|g\|_\sigma^2 = \text{tr}(B^2), \]

where \( A = \sum_i a_i z_i z_i^T \) and \( B = \sum_j b_j z_j z_j^T \).

**Lemma 4.1.** Let \( X, Y \) be symmetric and positive semidefinite. Then \( \text{tr}(XY) \leq \text{tr}(X) \text{tr}(Y) \).

With these intermediate results in hand, we can prove the main bound of interest:

**Proof of Theorem 4.2** By Proposition 4.1 we may write

\[ \|fg\|_\sigma^2 = \text{tr}(A^2B^2), \quad \|f\|_\sigma^2 = \text{tr}(A^2), \quad \|g\|_\sigma^2 = \text{tr}(B^2), \]

where \( A = \sum_i a_i z_i z_i^T \) and \( B = \sum_j b_j z_j z_j^T \) are chosen as described in Proposition 4.1. Since \( A \) and \( B \) are each symmetric, it follows that \( A^2 \) and \( B^2 \) are each symmetric and positive semidefinite. Then we can apply Lemma 4.1 to conclude that

\[ \|fg\|_\sigma^2 = \text{tr}(A^2B^2) \leq \text{tr}(A^2) \text{tr}(B^2) = \|f\|_\sigma^2 \|g\|_\sigma^2. \]

\[ \Box \]

### 4.2 Implications: kernel ridge regression

With the help of Theorem 4.1 we can develop DRO-based bounds for actual learning problems. In this section we develop such bounds for Gaussian kernel ridge regression, i.e. problem (13).

For shorthand, we write \( R_Q(f) = E_{x \sim q}[f(x)] = E_{x \sim q}(f(x) - h(x))^2 \) for the risk of \( f \) on a distribution \( Q \). Generalization amounts to proving that the population risk \( R_p(f) \) is not too different from the empirical risk \( R_{\hat{n}}(f) \).

**Theorem 4.2.** Assume the target function \( h \) satisfies \( \|h\|_{\sigma/\sqrt{2}} \leq \Lambda_h \) and \( \|h\|_{\sigma} \leq \Lambda_h \). Then, for any \( \delta > 0 \), with probability \( 1 - \delta \), the following holds for all functions \( f \) satisfying \( \|f\|_{\sigma/\sqrt{2}} \leq \Lambda_f \) and \( \|f\|_{\sigma} \leq \Lambda_f \):

\[ R_p(f) \leq R_{\hat{n}}(f) + \frac{2}{\sqrt{n}} \left( 1 + \sqrt{\frac{\log(1/\delta)}{2}} \right) (\Lambda_f + \Lambda_h + 2\Lambda_f \Lambda_h). \]

\[ \text{(14)} \]

**Proof.** We utilize the DRO Generalization Principle 2.1. By Lemma 3.1 we know that with probability \( 1 - \delta \), \( d_{\text{MMD}}(P_{\hat{n}}, P) \leq \epsilon \) for \( \epsilon = (2 + \sqrt{2\log(1/\delta)})/\sqrt{n} \), since \( k_{\sigma}(x, x) \leq M = 1 \). Note the bandwidth \( \sigma \) does not affect the convergence result. As a result of Lemma 3.1 with probability \( 1 - \delta \):

\[ R_p(f) = E_{x \sim \hat{P}}[(f(x) - h(x))^2] \]

\[ \leq E_{x \sim \hat{P}}[(f(x) - h(x))^2] + \epsilon \|f - h\|_{\sigma/\sqrt{2}}^2 \]

\[ \leq R_{\hat{n}}(f) + \epsilon \left( \|f\|_{\sigma/\sqrt{2}}^2 + \|h\|_{\sigma/\sqrt{2}}^2 + 2\|fh\|_{\sigma/\sqrt{2}}^2 \right) \]

\[ \leq R_{\hat{n}}(f) + \epsilon \left( \Lambda_f^2 + \Lambda_h^2 + 2\Lambda_f \Lambda_h \right), \]

\[ \text{(18)} \]

where (a) is by Corollary 3.2, (b) is by the triangle inequality, and (c) follows from Theorem 4.1 and our assumptions on \( f \) and \( h \). Plugging in the bound on \( \epsilon \) yields the result. \[ \Box \]

We placed different bounds on each of \( f, h, f^2, h^2 \) to emphasize the dependence on each. Since each is bounded separately, the DRO based bound in Theorem 4.2 allows finer control of the complexity of the function class than is typical. Since, by Theorem 4.1, the norms of \( f^2, h^2 \) and \( fh \) are bounded by those of \( f \) and \( h \), we may also state Theorem 4.2 just with \( \|f\|_{\sigma} \) and \( \|h\|_{\sigma} \).

**Corollary 4.1.** Assume the target function \( h \) satisfies \( \|h\|_{\sigma} \leq \Lambda \). Then, for any \( \delta > 0 \), with probability \( 1 - \delta \), the following holds for all functions \( f \) satisfying \( \|f\|_{\sigma} \leq \Lambda \):

\[ R_p(f) \leq R_{\hat{n}}(f) + \frac{8 \Lambda^2}{\sqrt{n}} \left( 1 + \sqrt{\frac{\log(1/\delta)}{2}} \right). \]

\[ \text{(19)} \]
Proof. We reduce to Theorem 4.2. By Theorem 4.1, we know that \( \| f \|_{\sigma/\sqrt{2}} \leq \| f \|_{T}^{2} \), which may be bounded above by \( \Lambda^{2} \) (and similarly for \( \Lambda \)). Therefore we can take \( \Lambda f_{x} = \Lambda f \) and \( \Lambda h_{z} = \Lambda h = \Lambda \) in Theorem 4.2. The result follows by bounding
\[
\Lambda f_{x} + \Lambda h_{z} + 2 \Lambda y \Lambda h \leq \Lambda^{2} + \Lambda^{2} + 2 \Lambda \cdot \Lambda = 4 \Lambda^{2}.
\]

Generalization bounds for kernel ridge regression are of course not new; we emphasize that the DRO viewpoint provides an intuitive approach that also grants finer control over the function complexity. Moreover, our results take essentially the same form as the typical generalization bounds for kernel ridge regression, reproduced below:

**Theorem 4.3** (Specialized from [29], Theorem 10.7). Assume the target function \( h \) satisfies \( \| h \|_{\sigma} \leq \Lambda \). Then, for any \( \delta > 0 \), with probability \( 1 - \delta \), it holds for all functions \( f \) satisfying \( \| f \|_{\sigma} \leq \Lambda \) that
\[
R_{\delta}(f) \leq R_{\delta,n}(f) \leq 8 \frac{\Lambda^{2}}{\sqrt{n}} \left( 1 + \frac{1}{2} \sqrt{\frac{\log(1/\delta)}{2}} \right).
\]

Hence, our DRO-based Theorem 4.2 evidently recovers standard results up to a universal constant.

### 4.3 Algorithmic implications

The generalization result in Theorem 4.3 is often used to justify penalizing by the norm \( \| f \|_{\sigma} \), since it is the only part of the bound (other than the risk \( R_{\delta,n}(f) \)) that depends on \( f \). In contrast, our DRO-based generalization bound in Theorem 5.2 is of the form
\[
R_{\delta}(f) - R_{\delta,n}(f) \leq \epsilon \left( \| f \|_{\sigma/\sqrt{2}}^{2} + h^{2} \| f \|_{\sigma/\sqrt{2}}^{2} + 2 \| f \|_{\sigma} \right) \| h \|_{\sigma},
\]
which depends on \( f \) through both norms \( \| f \|_{\sigma} \) and \( \| f \|_{\sigma/\sqrt{2}} \). This bound motivates the use of both norms as regularizers in kernel regression, i.e. we would instead solve
\[
\inf_{f \in \mathcal{H}_{\sigma}} \mathbb{E}_{(x,y) \sim \mathbb{P}_{n}}[(f(x) - y)^{2} + \lambda_{1} \| f \|_{\sigma} + \lambda_{2} \| f \|_{\sigma/\sqrt{2}}^{2}].
\]

Given data \((x_{i}, y_{i})\)\(_{i=1}^{n}\), for kernel ridge regression, the Representer Theorem implies that it is sufficient to consider only \( f \) of the form \( f = \sum_{i=1}^{n} a_{i} k_{\sigma}(x_{i}, \cdot) \). Here this is not in general possible due to the norm of \( f^{2} \). However, it is possible to evaluate and compute gradients of \( \| f \|_{\sigma/\sqrt{2}}^{2} \); let \( K \) be the matrix with \( K_{ij} = k_{\sigma}(x_{i}, x_{j}) \), and let \( D = \text{diag}(a) \). Using Proposition 4.1, we can prove \( \| f \|_{\sigma/\sqrt{2}}^{2} = \text{tr}(DK^{4}) \). A complete proof is given in the appendix.

### 5 Approximation and connections to variance regularization

In the previous section we studied bounding the MMD DRO problem (5) via Hilbert norm penalization. Going beyond kernel methods where we search over \( f \in \mathcal{H} \), it is even less clear how to evaluate the Hilbert norm \( \| f \|_{\mathcal{H}} \). To circumvent this issue, next we approach the DRO problem from a different angle: we directly search for the adversarial distribution \( \mathbb{Q} \). Along the way, we will build connections to variance regularization [27][18][22][31], where the empirical risk is regularized by the empirical variance of \( \ell_{f} \): \( \text{Var}_{\mathbb{P}_{n}}(\ell_{f}) = \mathbb{E} \mathbb{E}_{x \sim \mathbb{P}_{n}}[\ell_{f}(x)]^{2} - \mathbb{E} \mathbb{E}_{x \sim \mathbb{P}_{n}}[\ell_{f}(x)]^{2} \). In particular, we show in Theorem 5.1 that MMD DRO yields stronger regularization than variance.

Searching over all distributions \( \mathbb{Q} \) in the MMD ball is intractable, so we restrict our attention to those with the same support \((x_{i})\)\(_{i=1}^{n}\) as the empirical sample \( \mathbb{P}_{n} \). All such distributions \( \mathbb{Q} \) can be written as \( \mathbb{Q} = \sum_{i=1}^{n} w_{i} \delta_{x_{i}} \), where \( w \) is in the \( n \)-dimensional simplex. By restricting the set of candidate distributions \( \mathbb{Q} \), we make the adversary weaker:

\[
\sup_{\mathbb{Q}} \mathbb{E}_{x \sim \mathbb{Q}}[\ell_{f}(x)] \quad \text{s.t.} \quad d_{\text{MMD}}(\mathbb{Q}, \mathbb{P}_{n}) \leq \epsilon \geq \sup_{w} \sum_{i=1}^{n} w_{i} \ell_{f}(x_{i}) \quad \text{s.t.} \quad d_{\text{MMD}}(\sum_{i=1}^{n} w_{i} \delta_{x_{i}}, \mathbb{P}_{n}) \leq \epsilon \sum_{i=1}^{n} w_{i} = 1 \quad w_{i} \geq 0 \quad \forall i = 1, \ldots, n.
\]

By restricting the support of \( \mathbb{Q} \), it is no longer possible to guarantee out of sample performance, since it typically will have different support. Yet, as we will see, problem (23) has nice connections.
As a consequence, we conclude that with the right choice of kernel \( k \), MMD DRO is a stronger regularizer than variance:

**Theorem 5.1.** There exists a kernel \( k \) so that MMD DRO bounds the variance regularized problem:

\[
\mathbb{E}_{x \sim \hat{p}_n}[\ell_f(x)] \leq \mathbb{E}_{x \sim \hat{p}_n}[\ell_f(x)] + \epsilon \sqrt{n \Var_{\hat{p}_n}(\ell_f)} \leq \sup_{Q: d_{\text{MMD}}(Q, \hat{p}_n) \leq \epsilon} \mathbb{E}_Q[\ell_f(x)].
\]  

### 6 Experiments

In subsection 4.3 we proposed an alternate regularizer for kernel ridge regression, specifically, penalizing \( \|f\|_{\sigma/\sqrt{2}}^2 \) instead of \( \|f\|_2^2 \). Here we probe the new regularizer on a synthetic problem where we can precisely compute the population risk \( R_0(f) \). Consider the Gaussian kernel \( k_\sigma \) with \( \sigma = 1 \). Fix the ground truth \( h = k_\sigma(1, \cdot) - k_\sigma(-1, \cdot) \in \mathcal{H}_\sigma \). Sample \( 10^4 \) points from a standard one-dimensional Gaussian, and set this as the population \( \mathbb{P} \). Then subsample \( n \) points \( x_i = h(x_i) + \epsilon_i \),
where $\epsilon_i$ are Gaussian. We consider both an easy regime, where $n = 10^3$ and $\text{Var}(\epsilon_i) = 10^{-2}$, and a hard regime where $n = 10^2$ and $\text{Var}(\epsilon_i) = 1$. On the empirical data, we fit $f \in \mathcal{H}_\sigma$ by minimizing square loss plus either $\lambda \Vert f \Vert_\sigma^2$ (as is typical) or $\lambda \Vert f^2 \Vert_\sigma/\sqrt{2}$ (our proposal). We average over $10^2$ resampling trials for the easy case and $10^3$ for the hard case, and report 95% confidence intervals. Figure 1 shows the result in each case for a parameter sweep over $\lambda$. If $\lambda$ is tuned properly, the tighter regularizer $\Vert f^2 \Vert_\sigma/\sqrt{2}$ yields better performance in both cases. It also appears the regularizer $\Vert f^2 \Vert_\sigma/\sqrt{2}$ is less sensitive to the choice of $\lambda$: performance decays slowly when $\lambda$ is too low.

7 Conclusion

We introduce MMD DRO, distributionally robust optimization with maximum mean discrepancy uncertainty sets. We prove fundamental structural results and upper bounds for MMD DRO, and unearth deep connections, in particular to Gaussian kernel ridge regression and variance regularization.

Several open questions remain. In terms of theory, our MMD DRO approach to generalization bounds leaves much new ground to explore. In particular, we conjecture that our approach might also work for ridge regression with non-Gaussian kernels. Practically, there is also much left to do to make MMD DRO a general purpose tool. We have presented two approximations of MMD DRO, each with strengths and drawbacks: the upper bound in Corollary 3.2 enables our kernel ridge regression generalization bound, but is potentially loose, and is difficult to use more generally because the Hilbert norm is tricky to compute; the discrete approximation in Section 5 is more practical but is not an upper bound on the MMD DRO problem. Future work could address these drawbacks, or potentially develop a tractable exact reformulation of the DRO problem.

Acknowledgements

This work was supported by The Defense Advanced Research Projects Agency (grant number YFA17 N66001-17-1-4039). The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. We thank Cameron Musco and Joshua Robinson for helpful conversations, and Marwa El Halabi and Sebastian Claici for comments on the draft.

References


