Generalization Bounds with Minimal Dependency on Hypothesis Class via Distributionally Robust Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Established approaches to obtain generalization bounds in data-driven optimization and machine learning mostly build on solutions from empirical risk minimization (ERM), which depend crucially on the functional complexity of the hypothesis class. In this paper, we present an alternate route to obtain these bounds on the solution from distributionally robust optimization (DRO), a recent data-driven optimization framework based on worst-case analysis and the notion of ambiguity set to capture statistical uncertainty. In contrast to the hypothesis class complexity in ERM, our DRO bounds depend on the ambiguity set geometry and its compatibility with the true loss function. Notably, when using maximum mean discrepancy as a DRO distance metric, our analysis implies generalization bounds whose dependence on the hypothesis class appears the minimal possible: The bound depends solely on the true loss function, independent of any other candidates in the hypothesis class. To our best knowledge, it is the first generalization bound of this type in the literature, and we hope our findings can open the door for a better understanding of DRO, especially its benefits on loss minimization and other machine learning applications.

1 Introduction

We study generalization error in the following form. Let $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ be a loss function over the function class $\mathcal{F}$ and sample space $\mathcal{Z} \subset \mathbb{R}^d$. Let $P\ell(f, \cdot) := \mathbb{E}_{z\sim P}[\ell(f, z)]$ be the expected loss under the true distribution P for z. This $P\ell(f, \cdot)$ can be an objective function ranging from various operations research applications to the risk of a machine learning model. Given iid samples $\{z_i\}_{i=1}^n \sim P$, we solve a data-driven optimization problem or fit model to give solution f_{data}. The *excess risk*, or optimality gap of f_{data} with respect to the oracle best solution $f^* \in \arg\min_{f\in\mathcal{F}} P\ell(f, \cdot)$, is given by

$$P\ell(f_{\text{data}}, \cdot) - P\ell(f^*, \cdot) \tag{1}$$

This gap measures the relative performance of a data-driven solution in future test data, giving rise to a direct measurement on the generalization error. Established approaches to obtain high probability bounds for (1) have predominantly focused on empirical risk minimization (ERM)

$$\hat{P}\ell(f, \cdot) := \mathbb{E}_{z\sim\hat{P}}[\ell(f, z)], \quad \hat{f}^* \in \arg\min_{f\in\mathcal{F}} \hat{P}\ell(f, \cdot)$$

for empirical distribution $\hat{P}$, or its regularized versions. A core determination of the quality of these bounds is the functional complexity of the model, or hypothesis class, which dictates the richness of the function class $\mathcal{F}$ or $\mathcal{L} := \{l(f, \cdot) \mid f \in \mathcal{F}\}$ and leads to well-known measures such as the Vapnik-Chervonenkis (VC) dimension [1]. On a high level, these complexity measures arise from the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

need to uniformly control the empirical error, which in turn arises from the a priori uncertainty on the decision variable f in the optimization.

In this paper, we present an alternate route to obtain concentration bounds for (1) on solutions obtained from distributionally robust optimization (DRO). The latter started as a decision-making framework for optimization under stochastic uncertainty [2–4], and has recently surged in popularity in machine learning, thanks to its abundant connections to regularization and variability penalty [5–12] and risk-averse interpretations [13, 14]. Instead of replacing the unknown true expectation $\mathbb{E}_{z\sim P}[\cdot]$ by an empirical expectation $\mathbb{E}_{z\sim \hat{P}}[\cdot]$, DRO hinges on the creation of an *uncertainty set* or *ambiguity set* $\mathcal{K}$. This set lies in the space of probability distribution $\mathcal{P}$ on z and is calibrated from data. It obtains a solution f^*_{dro} by minimizing the worst-case expected loss among $\mathcal{K}$

$$P_{\text{dro}}\ell(f,\cdot) := \max_{Q\in\mathcal{K}} \mathbb{E}_{z\sim Q}[\ell(f,z)], \quad f^*_{\text{dro}} \in \arg\min_{f\in\mathcal{F}} P_{\text{dro}}\ell(f,\cdot). \tag{2}$$

The risk-averse nature of DRO is evident from the presence of an adversary that controls Q in (2). Moreover, if $\mathcal{K}$ is suitably chosen so that it contains the true probability with confidence (in some suitable sense) and shrinks to singleton as data size grows, then one would expect f^*_{dro} to eventually approach the true solution, which also justifies the approach as a consistent training method. The latter can often be achieved by choosing $\mathcal{K}$ as a neighborhood ball surrounding a baseline distribution that estimates the ground truth (notably the empirical distribution), and the ball size is measured via a statistical distance.

Our main goal is to present a line of analysis to bound (1) for DRO solutions, i.e.,

$$P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot) \tag{3}$$

which, instead of using functional complexity measures as in ERM, relies on the ambiguity set geometry and its compatibility with *only* the true loss function. More precisely, this bound depends on two ingredients: (i) the probability that the true distribution lies in $\mathcal{K}$, i.e., $\mathbb{P}[P\in\mathcal{K}]$ and, (ii) given that this occurs, the difference between the robust and true objective functions evaluated at the true solution f^*, i.e., $P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot)$. These ingredients allow us to attain generalization bounds that depend on the hypothesis class in a distinct, less sensitive fashion from ERM. Specifically, when using the maximum mean discrepancy (MMD) [15] as a statistical distance in DRO, our analysis implies generalization bounds whose dependence on the hypothesis class $\mathcal{L}$ appears the minimal possible: The bound depends solely on the true loss function $\ell(f^*,\cdot)$, independent of any other $\ell(f,\cdot)\in\mathcal{L}$.

Note that although there exist generalization bounds that do not utilize conventional complexity measures like the VC dimension [1], they are all still dependent on other candidates in the hypothesis class and therefore, are different from our results. Examples include generalization bounds based on hypothesis stability [16], algorithmic robustness [17], and the RKHS norm [18]. These approaches all rely on the data-driven algorithmic output f_{data}, which varies randomly among the hypothesis class due to its dependence on the random training data. Therefore, to translate into generalization bounds, they subsequently require taking expectation over f_{data} ([16, Definition 3 and Theorem 11]) or taking a uniform bound over $\mathcal{F}$ ([17, Definition 2 and Theorem 1]; [18, Theorem 4.2]). In contrast, our bound depends on the hypothesis class through the deterministic $\ell(f^*,\cdot)$, in a way that links to the choice of ambiguity set distance. To our best knowledge, generalization bounds of our type have never appeared in the literature (not only for DRO but also other ML approaches). We hope such a unique property will be useful in developing better machine learning algorithms in the future, especially in harnessing DRO on loss minimization and broader statistical problems.

Finally, our another goal is to conduct a comprehensive review on how DRO intersects with machine learning, which serves to position our new bounds over an array of motivations and interpretations of DRO. This contribution is in Section 2 before we present our main results (Section 3 - 5) and numerical experiments (Section 6).

2 Related Work and Comparisons

DRO can be viewed as a generalization of (deterministic) robust optimization (RO) [19, 20]. The latter advocates the handling of unknown or uncertain parameters in optimization problems via a worst-case perspective, which often leads to minimax formulations. [21, 22] show the equivalence of

ERM regularization with RO in some statistical models, and [17] further concretizes the relation of generalization with robustness.

DRO, which first appeared in [23] in the context of inventory management, applies the worst-case idea to stochastic problems where the underlying distribution is uncertain. Like in RO, it advocates a minimax approach to decision-making, but with the inner maximization resulting in the worst-case distribution over an ambiguity set $\mathcal{K}$ of plausible distributions. This idea has appeared across various disciplines like stochastic control [24] and economics [25, 26]. Data-driven DRO constructs and calibrates $\mathcal{K}$ based on data when available. The construction can be roughly categorized into two approaches: (i) neighborhood ball using statistical distance, which include most commonly ϕ-divergence [27–31] and Wasserstein distance [32–34, 6]; (ii) partial distributional information including moment [35, 2–4, 36], distributional shape [37–40] and marginal [41–43] constraints. The former approach has the advantage that the ambiguity set or the attained robust objective value consistently approaches the truth [27, 44]. The second approach, on the other hand, provides flexibility to decision-maker when limited data is available which proves useful on a range of operational or risk-related settings [45–47].

The first approach above, namely statistical-distance-based DRO, has gained momentum especially in statistics and machine learning in recent years. We categorize its connection with statistical performance into three lines, and position our results in this paper within each of them.

2.1 Absolute Bounds on Expected Loss

The classical approach to obtain guarantees for data-driven DRO is to interpret the ambiguity set $\mathcal{K}$ as a nonparametric confidence region, namely that $\mathbb{P}[P \in \mathcal{K}] \geq 1 - \delta$ for small $\delta \in (0, 1)$. In this case, the confidence guarantee on the set can be translated into a confidence bound on the true expected loss function evaluated at the DRO solution f^*_{dro} in the form

$$\mathbb{P}[P\ell(f^*_{\text{dro}}, \cdot) \leq P_{\text{dro}}\ell(f^*_{\text{dro}}, \cdot)] \geq 1 - \delta \tag{4}$$

via a direct use of the worst-case definition of P_{dro}. This implication is very general, with $\mathcal{K}$ taking possibly any geometry (e.g., [2, 27, 44, 32]). A main concern on results in the form (4) is that the bound could be loose (i.e., a large $P_{\text{dro}}\ell(f^*_{\text{dro}}, \cdot)$). This, in some sense, is unsurprising as the analysis only requires a confidence guarantee on the set $\mathcal{K}$, with no usage of other more specific properties, which is also the reason why the bound (4) is general. When $\mathcal{K}$ is suitably chosen, a series of work has shown that the bound (4) can achieve tightness in some well-defined sense. [48, 49] show that when $\mathcal{K}$ is a Kullback-Leibler divergence ball, $P_{\text{dro}}\ell(f^*_{\text{dro}}, \cdot)$ in (4) is the minimal among all possible data-driven formulations that satisfy a given exponential decay rate on the confidence. [8, 50, 51] show that for divergence-based $\mathcal{K}$, $P_{\text{dro}}\ell(f, \cdot)$ matches the confidence bound obtained from the standard central limit theorem (CLT) by deducing that it is approximately P_{dro} plus a standard deviation term (see more related discussion momentarily).

Our result leverages part of the above "confidence translation" argument, but carefully twisted to obtain excess risk bounds for (1). We caution that (4) is a result on the validity of the estimated objective value $P_{\text{dro}}\ell(f^*_{\text{dro}}, \cdot)$ in bounding the true objective value $P\ell(f^*_{\text{dro}}, \cdot)$. The excess risk (1), on the other hand, measures the generalization performance of a solution in comparison with the oracle best. The latter is arguably more intricate as it involves the unknown true optimal solution f^* and, as we will see, (4) provides an intermediate building block in our analysis of (1).

2.2 Variability Regularization

In a series of work [30, 31, 7, 9, 51], it is shown that DRO using divergence-based ball, i.e., $\mathcal{K} = \{Q \in \mathcal{P} : D(Q, \hat{P}) \leq \eta\}$ for some threshold $\eta > 0$ and D a ϕ-divergence (e.g., Kullback-Leibler, χ^2-distance), satisfies a Taylor-type expansion

$$P_{\text{dro}}\ell(f, \cdot) = \hat{P}\ell(f, \cdot) + C_1(f)\sqrt{\eta} + C_2(f)\eta + \cdots \tag{5}$$

where $C_1(f)$ is the standard deviation of the loss function, $\sqrt{\text{Var}[\ell(f, z)]}$, multiplied by a constant that depends on ϕ. Similarly, if D is the Wasserstein distance and η is of order $1/n$, (5) holds with $C_1(f)$ being the gradient norm or the Lipschitz norm [12, 10, 5, 52]. Furthermore, [18], which is perhaps closest to our work, studies MMD as the DRO statistical distance and derives a high-probability bound for $P_{\text{dro}}\ell(f, \cdot)$ similar to the RHS of (5), with $C_1(f)$ being the reproducing kernel

Hilbert space (RKHS) norm of $\ell(f,\cdot)$. Results of the form (5) can be used to show that $P_{\text{dro}}\ell(f,\cdot)$, upon choosing η properly (of order $1/n$), gives a confidence bound on $P\ell(f,\cdot)$ [8, 51]. Moreover, this result can be viewed as a duality of the empirical likelihood theory [53, 12, 51, 54].

In connecting (5) to the solution performance, there are three implications. First, the robust objective function $P_{\text{dro}}\ell(f,\cdot)$ can be interpreted as approximately a mean-variance optimization, and [7, 55] prove that, thanks to this approximation, the DRO solution can lower the variance of the attained loss which compensates for its under-performance in the expected loss, thus overall leading to a desirable risk profile. [7, 55] have taken a viewpoint that the variance of the attained loss is important in the generalization. On the other hand, when the expected loss, i.e., the true objective function $P\ell(f,\cdot)$, is the sole consideration, the approximation (5) is used in two ways. One way is to obtain bounds in the form

$$P\ell(f^*_{\text{dro}},\cdot) \leq \min_{f\in\mathcal{F}}\{P\ell(f,\cdot) + C_1(f)/\sqrt{n}\} + O(1/n) \tag{6}$$

thus showing that DRO performs optimally, up to $O(1/n)$ error, on the variance-regularized objective function [9]. From this, [9] deduces that under special situations where there exists f with both small risk $P\ell(f,\cdot)$ and variance $\text{Var}[\ell(f,z)]$, (6) can be translated into a small excess risk bound of order $1/n$. Moreover, such an order also desirably appears for DRO in some non-smooth problems where ERM could bear $1/\sqrt{n}$. The second way is to use DRO as a mathematical route to obtain uniform bounds in the form

$$P\ell(f,\cdot) \leq \hat{P}\ell(f,\cdot) + C_1(f)/\sqrt{n} + O(1/n),\ \forall f\in\mathcal{F}, \tag{7}$$

which is useful for proving the generalization of ERM. In particular, we can translate (7) into a high probability bound for $P\ell(\hat{f}^*,\cdot) - P\ell(f^*,\cdot)$ of order $1/\sqrt{n}$. This use is studied in, e.g., [56] in the case of Wasserstein and [18] in the case of MMD.

Despite these rich results, both (6) and (7), and their derived bounds on the excess risk (1), still depend on other candidates in the hypothesis class through the choice of the ball size η or the coefficient $C_1(f)$. Our main message in this paper is that this dependence can be completely removed, when using the solution of a suitably constructed DRO. Note that this is different from localized results for ERM, e.g., local Rademacher complexities [57]. Although the latter establishes better generalization bounds by mitigating some dependence on the hypothesis class, local dependence on other candidates still exists and cannot be removed.

2.3 Risk Aversion

It is known that any coherent risk measure (e.g., conditional value-at-risk) [13, 14] of a random variable (in our case $\ell(f,z)$) is equivalent to the robust objective value $P_{\text{dro}}\ell(f,\cdot)$ with a particular $\mathcal{K}$. Thus, DRO is equivalent to a risk measure minimization, which in turn explains its benefit in controlling tail performances. In machine learning, this rationale has been adopted to enhance performance on minority subpopulations and in safety or fairness-critical systems [58]. A related application is adversarial training in deep learning, in which RO or DRO is used to improve test performances on perturbed input data and adversarial examples [59, 60]. DRO has also been used to tackle distributional shifts in transfer learning [61, 62]. In these applications, numerical results suggest that RO and DRO can come at a degradation to the average-case performance [63, 60, 64] (though not universally, e.g., [65] observes that robust training can help reduce generalization error with very few training data). To connect, our result in this paper serves to justify a generalization performance of DRO even in the average case that can potentially outperform ERM.

3 General Results

We first present a general DRO bound.

Theorem 3.1. (A General DRO Bound) *Let $\mathcal{Z}\subset\mathbb{R}^d$ be a sample space, P be a distribution on $\mathcal{Z}$, $\{z_i\}_{i=1}^n$ be iid samples from P, and $\mathcal{F}$ be the function class. For loss function $\ell:\mathcal{F}\times\mathcal{Z}\to\mathbb{R}$ and DRO $\min_{f\in\mathcal{F}}\max_{Q\in\mathcal{K}}\mathbb{E}_{z\sim Q}[\ell(f,z)]$, DRO solution f^*_{dro} satisfies that for any $\varepsilon>0$,*

$$\mathbb{P}[P\ell(f^*_{dro},\cdot) - P\ell(f^*,\cdot) > \varepsilon] \leq \mathbb{P}[P\notin\mathcal{K}] + \mathbb{P}[P_{dro}\ell(f^*,\cdot) - P\ell(f^*,\cdot) > \varepsilon | P\in\mathcal{K}]. \tag{8}$$

Proof of Theorem 3.1. We first rewrite $P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot)$ as

$$[P\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot)] + [P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*,\cdot)] + [P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot)]. \quad (9)$$

Note that the second term $P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*,\cdot) \leq 0$ almost surely by the DRO optimality of f^*_{dro}. Therefore,

$$\mathbb{P}[P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot) > \varepsilon] \leq \mathbb{P}[[P\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot)] + [P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot)] > \varepsilon].$$

Now, by definition of $P_{\text{dro}}l(\cdot,\cdot)$ as the worst-case objective function, we have $P_{\text{dro}}\ell(f,\cdot) \geq P\ell(f,\cdot)$ for all $f \in \mathcal{F}$ as long as $P \in \mathcal{K}$. Thus,

$$\begin{aligned} & \mathbb{P}[P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot) > \varepsilon] \leq \mathbb{P}[P \notin \mathcal{K}] + \mathbb{P}[P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) > \varepsilon, P \in \mathcal{K}] \\ \leq \quad & \mathbb{P}[P \notin \mathcal{K}] + \mathbb{P}[P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) > \varepsilon | P \in \mathcal{K}]. \end{aligned} \quad (10)$$

This completes our proof. □

A key feature of the bound in Theorem 3.1 is that it requires only the true loss function $\ell(f^*,\cdot)$. This contrasts sharply with other bounds that rely on other candidates in the hypothesis class. To see where this distinction arises, note that in (9) we divide $P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot)$ into three parts where the second term is trivially ≤ 0 and the third term depends only on the true loss. The key is that the first term satisfies $P\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot) \leq 0$ as long as $P \in \mathcal{K}$, thanks to the worst-case definition of the robust objective function P_{dro}. In contrast, in the ERM case for instance, the same line of analysis gives $[P\ell(\hat{f}^*,\cdot) - \hat{P}\ell(\hat{f}^*,\cdot)] + [\hat{P}\ell(\hat{f}^*,\cdot) - \hat{P}\ell(f^*,\cdot)] + [\hat{P}\ell(f^*,\cdot) - P\ell(f^*,\cdot)]$, and while the second and third terms are handled analogously, the first term depends on $\hat{f}^*$ that varies randomly among the hypothesis class and is typically handled by the uniform bound $\sup_{f\in\mathcal{F}} |P\ell(f,\cdot) - \hat{P}\ell(f,\cdot)|$ that requires empirical process analysis [66] and the complexity of the hypothesis class $\mathcal{F}$ or $\mathcal{L}$. Thus, in a sense, the worst-case nature of DRO "transfers" the uniformity requirement over the hypothesis class into alternate geometric requirements on only the true loss function.

4 Specialization to MMD DRO

4.1 Generalization Bounds

Our next step is to use Theorem 3.1 to derive generalization bounds for a concrete DRO formulation. In the following, we use ambiguity set $\mathcal{K}$ as a neighborhood ball of the empirical distribution measured by statistical distance, namely

$$\mathcal{K} = \{Q \in \mathcal{P} : D_{\text{MMD}}(Q, \hat{P}) \leq \eta\} \quad (11)$$

for a threshold $\eta > 0$. Moreover, we specialize in max mean discrepancy (MMD) [15] as the choice of distance $D_{\text{MMD}}(\cdot,\cdot)$. MMD is a distance derived from the RKHS norm, by using test functions constrained by this norm in an Integral Probability Metric (IPM) [67]. Specifically, let $k : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ be a positive definite kernel function on $\mathcal{Z}$ and $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be the corresponding RKHS. For any distributions Q_1 and Q_2, the MMD distance is defined by the maximum difference of the integrals over the unit ball $\{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq 1\}$. That is, $D_{\text{MMD}}(Q_1, Q_2) := \sup_{h\in\mathcal{H}:\|h\|_{\mathcal{H}}\leq 1} \int h \mathrm{d}Q_1 - \int h \mathrm{d}Q_2$. In this way, MMD-DRO can be formulated as

$$\min_{f\in\mathcal{F}} \max_{Q: D_{\text{MMD}}(Q,\hat{P})\leq\eta} \mathbb{E}_{z\sim Q}[\ell(f,z)] \quad (12)$$

MMD is known to be less prone to the curse of dimensionality [68–70], a property that we leverage in this paper, and such a property has allowed successful applications in statistical inference [15], generative models [71, 72], reinforcement learning [73], and DRO [18, 74]. In particular, [75] studies the duality and optimization procedure for MMD DRO. For a recent comprehensive review of MMD, we refer readers to [76].

In the sequel, we adopt bounded kernels, i.e., $\sup_{z,z'\in\mathcal{Z}} \sqrt{k(z,z')} < +\infty$, to conduct our analysis. This assumption applies to many popular choices of kernel functions and more importantly, guarantees the so-called kernel mean embedding (KME) [76] is well-defined in RKHS. That is, $\mu_Q := \int k(z,\cdot)\mathrm{d}Q \in \mathcal{H}$, $\forall\, Q \in \mathcal{P}$. This well-definedness of KME gives two important implications. One is that MMD is equivalent to the norm distance in KME: $D_{\text{MMD}}(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}$ [77, 15]. Second, it bridges the expectation and the inner product in the RKHS space [69]: $\forall\, f \in \mathcal{H}$, $\mathbb{E}_{x\sim P}[f(x)] = \int_x \langle f, k(x,\cdot)\rangle \mathrm{d}P(x) = \langle f, \mu_P\rangle$. Both properties above will facilitate our analysis. We refer the reader to Appendix A for further details on KME.

Theorem 4.1. (Generalization Bounds for MMD DRO) *Adopt the notation and assumptions in Theorem 3.1. Let $\mathcal{Z}$ be a compact subspace of $\mathbb{R}^d$, $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded continuous positive definite kernel, and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the corresponding RKHS. Suppose also that (i) $\sup_{z \in \mathcal{Z}} \sqrt{k(z,z)} \leq K$; (ii) $\ell(f^*, \cdot) \in \mathcal{H}$ with $\|\ell(f^*, \cdot)\|_{\mathcal{H}} \leq M$. Then, for all $\delta \geq 0$, if we choose ball size $\eta = \frac{K}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)})$ for MMD DRO* (12)*, then MMD DRO solution f^*_{dro} satisfies*

$$P\ell(f^*_{dro}, \cdot) - P\ell(f^*, \cdot) \leq \frac{2KM}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)})$$

with probability at least $1 - \delta$.

We briefly overview our proof of Theorem 4.1, leaving the details to Appendix B. In view of Theorem 3.1, the key lies in choosing the ball size η to make a good trade-off between $\mathbb{P}[D_{\text{MMD}}(P, \hat{P}) \geq \eta]$ and an upper bound of $P_{\text{dro}}\ell(f^*, \cdot) - P\ell(f^*, \cdot)$ when $D_{\text{MMD}}(P, \hat{P}) \leq \eta$. The former can be characterized by the established concentration rate of KME (See Proposition B.1). To get the latter bound, we use the properties of KME to translate expectations into inner products:

$$\begin{aligned} P_{\text{dro}}\ell(f^*, \cdot) - P\ell(f^*, \cdot) &= \max_{Q \in \mathcal{K}} \langle \ell(f^*, \cdot), \mu_Q - \mu_P \rangle \leq \max_{Q \in \mathcal{K}} \|\ell(f^*, \cdot)\|_{\mathcal{H}} D_{\text{MMD}}(Q, P) \\ &\leq \|\ell(f^*, \cdot)\|_{\mathcal{H}} \max_{Q \in \mathcal{K}} \{D_{\text{MMD}}(Q, \hat{P}) + D_{\text{MMD}}(\hat{P}, P)\} \leq \|\ell(f^*, \cdot)\|_{\mathcal{H}} \cdot 2\eta. \end{aligned} \tag{13}$$

From these, we can choose η as indicated in Theorem 4.1 to conclude the result. In a nutshell, here we harness the compatibility between $\mathcal{K}$ and the true loss function $\ell(f^*, \cdot)$, more specifically the RKHS norm on $\ell(f^*, \cdot)$, and the dimension-free concentration of KME. Our general bound in Theorem 3.1 allows us to stitch these two ingredients together to obtain generalization bounds for MMD DRO.

4.2 Strengths and Limitations Compared with ERM

Our generalization bound in Theorem 4.1 in fact still utilizes a uniformity argument, hidden in the distance construction. Note that our bound comprises essentially the product of $\|\ell(f^*, \cdot)\|_{\mathcal{H}}$, which depends on the true loss function, and η, the size of the ambiguity set. The latter is controlled by $D_{\text{MMD}}(P, \hat{P})$, which is equivalent to the supremum of the empirical process indexed by $\mathcal{H}_1 := \{h \mid h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq 1\}$. Note that this supremum is expressed through $\sup_{z \in \mathcal{Z}} \sqrt{k(z,z)}$, which is independent of the hypothesis class.

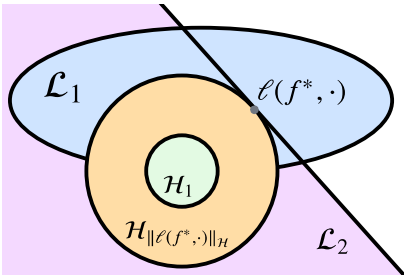


Figure 1: Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two hypothesis classes with the same true loss function $\ell(f^*, \cdot)$. $\mathcal{L}_1$ corresponds to functions in the oval and $\mathcal{L}_2$ corresponds to functions below the straight line. In ERM, the uniform bound is taken directly over $\mathcal{L}_1$ (blue area) or $\mathcal{L}_2$ (pink area), while DRO gives the same bound that corresponds to uniformity over $\mathcal{H}_{\|\ell(f^*,\cdot)\|_{\mathcal{H}}}$ (orange area).

To understand this further, Figure 1 provides geometric representations on the classes of functions over which ERM and our DRO apply uniform bounds. From this we also argue how DRO relies less heavily on the hypothesis class. In ERM, the uniform bound is taken directly over the hypothesis class, say $\mathcal{L}$. In DRO, from our discussion above, the uniform bound is taken over a dilation of the $\mathcal{H}_1$-ball by a factor $\|\ell(f^*, \cdot)\|_{\mathcal{H}}$, equivalently $\mathcal{H}_{\|\ell(f^*,\cdot)\|_{\mathcal{H}}} := \{h \mid h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq \|\ell(f^*, \cdot)\|_{\mathcal{H}}\}$. Thus, if two hypothesis classes contain the same true loss function $\ell(f^*, \cdot)$, their DRO generalization bounds would be the same, while ERM could differ. In this sense, DRO generalization exhibits less reliance on the hypothesis class.

With the above understanding, we can see several situations where DRO could outperform ERM: (i) *When $\|\ell(f^*, \cdot)\|_{\mathcal{H}}$ is small compared to the size of $\mathcal{L}$.* Though one may argue that $\ell(f^*, \cdot)$ is typically unknown and hence we ultimately resort to using uniformity over $\mathcal{L}$ to obtain any computable bound, our bound in terms of $\ell(f^*, \cdot)$ explains why DRO can sometimes substantially outperform ERM. In Section 6, we provide an example where $\|l(\theta^*, \cdot)\|_{\mathcal{H}} = 0$ and the excess risk of DRO achieves zero with high probability. (ii) *When uniformly bounding over $\mathcal{H}_1$ is "cheap".* Many commonly used kernel spaces are bounded and therefore uniform bounds over $\mathcal{H}_1$ are trivial to establish (by Proposition B.1), which in turn leads to generalization bounds that depend only on a

dilation (by $\|\ell(f^*,\cdot)\|_{\mathcal{H}}$) of these uniform bounds over $\mathcal{H}_1$. This is regardless of how complex the hypothesis class $\mathcal{L}$ is, including when $\|\mathcal{L}\|_{\mathcal{H}} = \infty$ or $\mathcal{L} = \mathcal{H}$. ERM, in contrast, uses uniformity over $\mathcal{L}$ and can result in an exploding generalization bound when $\|\mathcal{L}\|_{\mathcal{H}} = \infty$. (iii) *When $\mathcal{L}$ resembles the geometry of $\mathcal{H}$.* For instance, when $\mathcal{L} = \mathcal{H}_B := \{h \mid h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq B\}$, the ERM bound yields $2B \sup_{h\in\mathcal{H}_1}\{Ph - \hat{P}h\}$ since $B = \sup_{\ell\in\mathcal{L}} \|\ell(\cdot)\|_{\mathcal{H}}$, while the DRO bound gives $2\|\ell(f^*,\cdot)\|_{\mathcal{H}} \sup_{h\in\mathcal{H}_1}\{Ph - \hat{P}h\}$ that depends only on the true loss function and is smaller.

Despite the strengths, our bound in Theorem 4.1 also reveals situations where DRO could underperform ERM: When $\mathcal{L}$ has a shape deviating from the RKHS ball such that $\mathcal{L} \subset \mathcal{H}_{\|\ell(f^*,\cdot)\|_{\mathcal{H}}}$, then taking uniform bounds over $\mathcal{H}_{\|\ell(f^*,\cdot)\|_{\mathcal{H}}}$ is costly compared to over $\mathcal{L}$, in which case DRO generalization bound is worse than ERM. Moreover, note that $\ell(f^*,\cdot)$, the optimal loss within $\mathcal{L}$, can shift when $\mathcal{L}$ becomes richer. In this sense Theorem 4.1 cannot completely remove the dependency on the hypothesis class, but rather diminishes it to the minimum possible: The resulting bound does not rely on any other candidates in the hypothesis class except $\ell(f^*,\cdot)$.

Remark 4.2. There exists another line of work that establishes generalization bounds with a faster rate $O(n^{-1})$ for ERM [78, 57, 79, 80] and ϕ-divergence DRO [9]. However, they are all based on additional curvature conditions on $\mathcal{F}$ or $\mathcal{L}$ to relate the loss/risk and variance. Improving our bounds towards $O(n^{-1})$ rate is potentially achievable and constitutes future work.

4.3 Comparison with [18]

For bounds that do not utilize conventional complexity measures, to the best of our knowledge, they are still dependent on other candidates in the hypothesis class; recall the end of Section 1. As a concrete example, we compare our results with [18], which also studies generalization related to MMD DRO and appears the most relevant to our paper. First, different from our research goal, [18] aims to use MMD DRO as an analysis tool to derive uniform bounds for $P\ell(f,\cdot) - \hat{P}\ell(f,\cdot),\ \forall f \in \mathcal{F}$. More specifically, the main argument for their Theorem 4.2 is that, with high probability, one has $P\ell(f,\cdot) - \hat{P}\ell(f,\cdot) = O(\|\ell(f,\cdot)\|_{\mathcal{H}}\sqrt{\frac{\log(1/\delta)}{n}})$ for any given f. Although this bound seems to depend on a single loss function $\ell(f,\cdot)$, any algorithmic output, say f_{data}, needs to be learnt from data and is thus random. This means that to construct a generalization bound, one has to either average over f_{data} or take a supremem over $f \in \mathcal{F}$. [18] adopts the latter approach and establishes a bound $P\ell(f,\cdot) - \hat{P}\ell(f,\cdot) = O(\sup_{f\in\mathcal{F}} \|\ell(f,\cdot)\|_{\mathcal{H}}\sqrt{\frac{\log(1/\delta)}{n}}),\ \forall\ f \in \mathcal{F}$. in their Theorem 4.2. Our Theorem 4.1, on the other hand, only depends on $\|\ell(f^*,\cdot)\|_{\mathcal{H}}$ which is potentially much tighter and provides a distinct perspective.

Second, our Theorem 4.1 also distinguishes from [18] in that our bound applies specifically to the MMD DRO solution f^*_{dro} and reveals a unique property of this solution. This should be contrasted to [18] whose result is a uniform bound for any $f \in \mathcal{F}$, without any specialization to the DRO solution. This also relates to our next discussion.

4.4 Comparison with RKHS Regularized ERM

We also compare our results to RKHS regularized ERM:

$$\min_{f\in\mathcal{F}} \ell(f,z) + \eta\|f\|_{\mathcal{H}}. \tag{14}$$

When we additionally assume $\mathcal{L} \subset \mathcal{H}$, it is known that MMD DRO is equivalent to

$$\min_{f\in\mathcal{F}} \ell(f,z) + \eta\|\ell(f,\cdot)\|_{\mathcal{H}} \tag{15}$$

according to [18, Theorem 3.1]. Although (14) and (15) appear similar, their rationales are different. When regularizing $\|f\|_{\mathcal{H}}$, the goal is to maintain the simplicity of f. In comparison, through (15) we are in disguise conducting DRO and utilizing its variability regularization property (Recall Section 2.2). This observation motivates [18] to consider regularizing $\ell(f,\cdot)$ instead of f itself. Our Theorem 4.1 reveals the benefit of regularizing $\ell(f,\cdot)$ at a new level: Regularizing $\|\ell(f,\cdot)\|_{\mathcal{H}}$ increases the likelihood of $P\ell(f^*_{\text{dro}},\cdot) - P_{\text{dro}}\ell(f^*_{\text{dro}},\cdot) \leq 0$, which is argued by looking at the *distribution* space, instead of the loss function space, more specifically from the likelihood of a DRO ambiguity set covering P. Because of this, we can establish generalization bounds that depend on the hypothesis

class only through the true loss $\ell(f^*, \cdot)$ for RKHS-regularized ERM when the regularization is on $\ell(f, \cdot)$. Furthermore, thanks to this we can relax the assumption $\mathcal{L} \subset \mathcal{H}$ adopted in (15) to merely $\ell(f^*, \cdot) \in \mathcal{H}$ which is all we need in MMD DRO. Lastly, we point out that our analysis framework is not confined only to MMD, but can be applied to other distances in defining DRO (See Appendix D).

5 Extension to $\ell(f^*, \cdot) \notin \mathcal{H}$

In Theorem 4.1, we assume (i) $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} \leq K$; (ii) $\ell(f^*, \cdot) \in \mathcal{H}$ with $\|\ell(f^*, \cdot)\|_{\mathcal{H}} \leq M$. Although (i) is natural for many kernels, (ii) can be restrictive since many popular loss functions do not satisfy $\ell(f^*, \cdot) \in \mathcal{H}$. For instance, if k is a Gaussian kernel defined on $\mathcal{Z} \subset \mathbb{R}^d$ that has nonempty interior, then any nonzero polynomial on $\mathcal{Z}$ does not belong to the induced RKHS $\mathcal{H}$ [81]. This is in fact a common and well-known problem for almost all kernel methods [82], though many theoretical kernel method papers just assume the concerned function $g \in \mathcal{H}$ for simplicity, e.g., [83, 84, 18]. This theory-practice gap motivates us to extend our result to $\ell(f^*, \cdot) \notin \mathcal{H}$.

To this end, note that under universal kernels [85], any bounded continuous function can be approximated arbitrarily well (in L^∞ norm) by functions in $\mathcal{H}$. However, the RKHS norm of the approximating function, which appears in the final generalization bound, may grow arbitrarily large as the approximation becomes finer [86, 82]. This therefore requires analyzing a trade-off between the function approximation error and RKHS norm magnitude. More specifically, we define the approximation rate $I(\ell(f^*, \cdot), R) := \inf_{\|g\|_{\mathcal{H}} \leq R} \|\ell(f^*, \cdot) - g\|_\infty$, and $g_R := \arg\inf_{\|g\|_{\mathcal{H}} \leq R} \|\ell(f^*, \cdot) - g\|_\infty$.[1] As a concrete example, we adopt the Gaussian kernel and the Sobolev space to quantify such a rate and establish Theorem 5.1 below.

Theorem 5.1. (Generalization Bounds for Sobolev Loss Functions) *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be the RKHS induced by $k(z, z') = \exp(-\|z - z'\|_2^2/\sigma^2)$ on $\mathcal{Z} = [0,1]^d$. Suppose that $\ell(f^*, \cdot)$ is in the Sobolev space $H^d(\mathbb{R}^d)$. Then, there exists a constant C independent of R but dependent on $\ell(f^*, \cdot)$, d, and σ such that for all $\delta \geq 0$, if we choose ball size $\eta = \frac{1}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)})$ for MMD DRO* (12)*, then MMD DRO solution f^*_{dro} satisfies*

$$P\ell(f^*_{dro}, \cdot) - P\ell(f^*, \cdot) \leq 2 \inf_{R \geq 1} \left\{ C(\log R)^{-\frac{d}{16}} + \frac{R}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)}) \right\} \tag{16}$$

with probability at least $1 - \delta$.

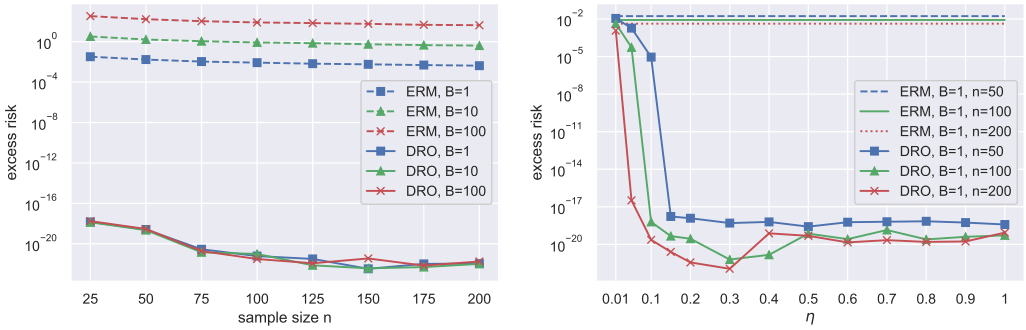
We briefly outline our proof, leaving the details to Appendix C. When $\ell(f^*, \cdot) \notin \mathcal{H}$, we cannot use KME to rewrite expectation as inner products as shown in (13). Rather, we will approximate $\ell(f^*, \cdot)$ by some $g_R \in \mathcal{H}$ and obtain $P\ell(f^*, \cdot) \leq Pg_R + I(\ell(f^*, \cdot), R)$. Since $g_R \in \mathcal{H}$, similar to (13) we can derive a generalization bound of g_R, which appears in the second term of (16). Then, it suffices to take infinum over $R \geq 1$ to establish the aforementioned trade-off between the approximation error $I(\ell(f^*, \cdot), R)$ (the first term of (16)) and the RKHS norm $\|g_R\|_{\mathcal{H}} \leq R$ (the second term of (16)).

Compared to Theorem 4.1, here we have to pay an extra price on the solution quality when the loss function is less compatible with the chosen RKHS, i.e., $\ell(f^*, \cdot) \notin \mathcal{H}$. Concretely, in the setting of Theorem 5.1, $I(\ell(f^*, \cdot), R) \leq C(\log R)^{-d/16}$ [86] and this results in a worse convergence rate in n. For example, inserting $R = \sqrt{n}(\log n)^{-d/16}$ into (16) yields $\mathcal{O}(\max\{C, \sqrt{\log(1/\delta)}\}(\log n)^{-16/d})$. This convergence rate is rather slow. However, it mainly results from the slow approximation rate $I(\ell(f^*, \cdot), R)$ established by non-parametric analysis. If we further utilize the structure of $\ell(f^*, \cdot)$, then a faster convergence rate or even close to $O(n^{-1/2})$ rate can be potentially achieved. Nonetheless, our main message in Theorem 5.1 is that even if $\ell(f^*, \cdot) \notin \mathcal{H}$, our bound (16) still depends minimally on the hypothesis class $\mathcal{L}$: The bound depends solely on the true loss function $\ell(f^*, \cdot)$, independent of any other $\ell(f, \cdot) \in \mathcal{L}$.

6 Numerical Experiments

We conduct simple experiments to study the numerical behaviors of our MMD DRO and compare with ERM on simulated data, which serves to illustrate the potential of MMD DRO in improving

[1] g_R is well-defined since for fixed R, $h(g) = \|\ell(f^*, \cdot) - g\|_\infty$ is a continuous function over a compact set $\{g : \|g\|_{\mathcal{H}} \leq R\}$.



(a) Performance of ERM and DRO under varying n (b) Performance of ERM and DRO under varying η

Figure 2: Excess risks of ERM and DRO

generalization and validate our developed theory. We adapt the experiment setups from [9, Section 5.2] and consider a quadratic loss with linear perturbation: $l(\theta, z) = \frac{1}{2}\|\theta - v\|_2^2 + z^\top(\theta - v)$, where $z \sim \text{Unif}[-B, B]^d$ with constant B varying from $\{1, 10, 100\}$ in the experiment. Here, $\mathcal{F} = \{\theta \mid \theta \in \Theta\}$ and we use f and θ interchangeably. v is chosen uniformly from $[1/2, 1]^d$ and is known to the learner in advance. In applying MMD DRO, we use Gaussian kernel $k(z, z') = \exp(-\|z - z'\|_2^2/\sigma^2)$ with σ set to the median of $\{\|z_i - z_j\|_2 \mid \forall\, i, j\}$ according to the well-known median heuristic [15]. To solve MMD DRO, we adopt the semi-infinite dual program and the constraint sampling approach from [75], where we uniformly sample a constraint set of size equal to the data size n. The sampled program is then solved by CVX [87] and MOSEK [88]. Each experiment takes the average performance of 500 independent trials. Our computational environment is a Mac mini with Apple M1 chip, 8 GB RAM and all algorithms are implemented in Python 3.8.3.

In Figure 2(a), we present the excess risks of ERM and DRO, i.e., $P\ell(f_{\text{data}}, \cdot) - P\ell(f^*, \cdot)$, where f_{data} denotes the ERM or DRO solution. We set $d = 5$ and tune the ball size via the best among $\eta \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.3, \ldots, 1.0\}$ (more details to be discussed on Figure 2(b) momentarily). Figure 2(a) shows that, as n increases, both DRO and ERM incur a decreasing loss. More importantly, DRO appears to perform much better than ERM. For $n = 200$ and $B = 100$ for instance, the expected loss of DRO is less than 10^{-20}, yet that of ERM remains at around 10^1.

We attribute the outperformance of DRO to its dependency on the hypothesis class only through the true loss function $\ell(f^*, \cdot)$. This can be supported by an illustration on the effect of ball size η. In Figure 2(b), we present excess risk $P\ell(f_{\text{data}}, \cdot) - P\ell(f^*, \cdot)$ for $B = 1, d = 5, n = 50$, and varying η. We see that the excess risk of the DRO solution drops sharply when η is small, and then for sufficiently big η, i.e., $\eta \in [0.5, 1]$, it remains at a fixed level of 10^{-19}, which is close to the machine accuracy of zero (such an "L"-shape excess loss also occurs under other choices of sample size n, and we leave out those results to avoid redundancy). Such a phenomenon coincides with our theorems. First, note that $f^* = v$ and the optimal loss function satisfies $\|\ell(f^*, \cdot)\|_{\mathcal{H}} = \|0\|_{\mathcal{H}} = 0$, so that our assumptions in Theorem 4.1 are satisfied. Recall that by Theorem 3.1,

$$\mathbb{P}[P\ell(f^*_{\text{dro}}, \cdot) - P\ell(f^*, \cdot) > \varepsilon] \leq \mathbb{P}[P \notin \mathcal{K}] + \mathbb{P}[P_{\text{dro}}\ell(f^*, \cdot) - P\ell(f^*, \cdot) > \varepsilon | P \in \mathcal{K}]. \tag{17}$$

Conditioning on $P \in \mathcal{K}$, we have $f^*_{\text{dro}} = f^* = v$ and $P_{\text{dro}}\ell(f^*_{\text{dro}}, \cdot) = P\ell(f^*, \cdot) = 0$. Thus, (17) yields $\mathbb{P}[P\ell(f^*_{\text{dro}}, \cdot) - P\ell(f^*, \cdot) > \varepsilon] \leq \mathbb{P}[P \notin \mathcal{K}]$. The proof of Theorem 4.1 reveals that $\forall\, \delta > 0$, if the ball size is big enough, namely $\eta \geq \frac{K}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)})$, then excess risk will vanish, i.e., $P\ell(f^*_{\text{dro}}, \cdot) - P\ell(f^*, \cdot) = 0$ with probability $\geq 1 - \delta$. This behavior shows up precisely in Figure 2(b).

Our discussion above is not to suggest the choice of η in practice. Rather, it serves as a conceptual proof for the correctness of our theorem. In fact, to the best of our knowledge, none of the previous results, including the multiple approaches to explain the generalization of DRO reviewed in Section 2, can explain the strong outperformance of DRO with the "L"-shape phenomenon in this example. These previous results focus on either absolute bounds (Section 2.1) or using variability regularization (Section 2.2) whose bounds, when converted to excess risks, typically increase in η when η is sufficiently large and, moreover, involve complexity measures on the hypothesis class (e.g., the ϕ-divergence case in [9, Theorems 3, 6] and the Wasserstein case in [56, Corollaries 2, 4]). Our developments in Section 3 thus appear to provide a unique explanation for the significant outperformance of DRO in this example.

References

[1] Vladimir Vapnik and Alexey Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3): 283–305, 1991.

[2] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[3] Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-Part-1):902–917, 2010.

[4] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[5] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2017.

[6] Ruidi Chen and Ioannis C Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018.

[7] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters*, 46(4):448–452, 2018.

[8] Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[9] John C Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20:68–1, 2019.

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

[11] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[12] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[13] Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006.

[14] R Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. Informs, 2007.

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[16] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[17] Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.

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

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

[20] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[21] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7), 2009.

[22] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and Lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

[23] Herbert Scarf. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 1958.

[24] Ian R Petersen, Matthew R James, and Paul Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000.

[25] Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton University Press, 2008.

[26] Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

[27] 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, 2013.

[28] Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *Tutorials in Operations Research*, pages 1–19. INFORMS, 2015.

[29] Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1):291–327, 2016.

[30] Henry Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[31] Henry Lam. Sensitivity to serial dependency of input processes: A robust approach. *Management Science*, 64(3):1311–1327, 2018.

[32] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[33] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[34] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[35] Laurent El Ghaoui, Maksim Oks, and Francois Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

[36] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62, 2015.

[37] Ioana Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

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

[39] Bowen Li, Ruiwei Jiang, and Johanna L Mathieu. Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming*, pages 1–42, 2017.

[40] Xi Chen, Simai He, Bo Jiang, Christopher Thomas Ryan, and Teng Zhang. The discrete moment problem with nonconvex shape constraints. *Operations Research*, 2020.

[41] Louis Chen, Will Ma, Karthik Natarajan, David Simchi-Levi, and Zhenzhen Yan. Distributionally robust linear and discrete optimization with marginals. *Available at SSRN 3159473*, 2018.

[42] Xuan Vinh Doan, Xiaobo Li, and Karthik Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research*, 63(6):1468–1488, 2015.

[43] Anulekha Dhara, Bikramjit Das, and Karthik Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. *INFORMS Journal on Computing*, 33(1):370–389, 2021.

[44] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282, 2018.

[45] Yiling Zhang, Siqian Shen, and Johanna L Mathieu. Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads. *IEEE Transactions on Power Systems*, 32(2):1378–1388, 2016.

[46] Henry Lam and Clementine Mottet. Tail analysis without parametric models: A worst-case perspective. *Operations Research*, 65(6):1696–1711, 2017.

[47] Chaoyue Zhao and Ruiwei Jiang. Distributionally robust contingency-constrained unit commitment. *IEEE Transactions on Power Systems*, 33(1):94–102, 2017.

[48] Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 2020.

[49] Tobias Sutter, Bart PG Van Parys, and Daniel Kuhn. A general framework for optimal data-driven optimization. *arXiv preprint arXiv:2010.06606*, 2020.

[50] Vishal Gupta. Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science*, 65(9):4242–4260, 2019.

[51] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

[52] Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, 2019.

[53] Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301 – 307, 2017. ISSN 0167-6377.

[54] Jose Blanchet and Yang Kang. Sample out-of-sample inference based on Wasserstein distance. *Operations Research*, 2021.

[55] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 2021.

[56] Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

[57] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[58] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[59] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[60] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[61] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[62] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8662–8670, 2021.

[63] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[64] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[65] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[66] Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[67] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

[68] John Shawe-Taylor, Nello Cristianini, et al. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[69] 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.

[70] Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, 2012.

[71] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

[72] Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

[73] Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *arXiv preprint arXiv:2007.12354*, 2020.

[74] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust Bayesian optimization. In *Artificial Intelligence and Statistics*, pages 2174–2184. PMLR, 2020.

[75] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *Artificial Intelligence and Statistics*, volume 130, pages 280–288. PMLR, 2021.

[76] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. ISSN 1935-8237.

[77] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[78] Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[79] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[80] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[81] Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32 (2):307–338, 2010.

[82] Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *International Conference on Machine Learning*, 2012.

[83] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

[84] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

[85] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

[86] Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

[87] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

[88] MOSEK ApS. *The MOSEK optimization toolbox for Python manual. Version 9.2.46.*, 2021. URL `https://docs.mosek.com/9.2/pythonfusion/index.html`.

[89] Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

[90] Ding-Xuan Zhou. Density problem and approximation error in learning theory. In *Abstract and Applied Analysis*, volume 2013. Hindawi, 2013.

[91] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[92] Robert E Gaunt, Alastair M Pickett, and Gesine Reinert. Chi-square approximation by stein's method with application to pearson's statistic. *The Annals of Applied Probability*, 27(2): 720–756, 2017.

[93] Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

[94] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [N/A] We only train and test on simulated data.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Not applicable
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We only train and test on simulated datasets.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Our simulation is relatively easy and will not consume much computation power.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Review on Kernel Mean Embedding

We first review the definition of kernel mean embedding (KME).

Definition A.1 (Kernel Mean Embedding)**.** Given a distribution Q, and a positive definite kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, i.e., given $n \in \mathbb{N}, a_1, \ldots, a_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k\left(z_i, z_j\right) \geq 0$$

holds for any $z_1, \ldots, z_n \in \mathcal{Z}$, the KME of Q is defined by:

$$\mu_Q := \int k(z, \cdot) Q(\mathrm{d}z).$$

We have the following property.

Proposition A.2. *[69] For a positive definite kernel k defined on compact space $\mathcal{Z} \times \mathcal{Z}$, let $\mathcal{H}$ be the induced RKHS. If $\mathbb{E}_{x\sim P}[k(z,z)] < +\infty$, then $\mu_P \in \mathcal{H}$ for all $Q \in \mathcal{P}$.*

It is easy to verify that for a bounded kernel, $\mathbb{E}_{z\sim Q}[k(z,z)] < +\infty$ for all distribution Q. Therefore, as a corollary of Proposition A.2, for any bounded kernel, KME is well-defined in RKHS, i.e., $\mu_P \in \mathcal{H}$.

The following proposition provides the properties of KME that we utilize in Section 4.1. Specifically, Proposition A.3 (i) establishes the equivalence between MMD and the norm distance in KME. Proposition A.3 (ii) bridges the expectation to inner product in RKHS in Eq. (18).

$$\mathbb{E}_{x\sim P}[f(x)] = \int_x \langle f, k(x, \cdot)\rangle \mathrm{d}P(x) = \langle f, \mu_P \rangle. \tag{18}$$

Proposition A.3. *Assume the condition in Proposition A.2 for the existence of the kernel mean embeddings μ_{Q_1}, μ_{Q_2} is satisfied. Then,*

(i) [77, 15] $D_{MMD}(Q_1, Q_2) = \|\mu_{Q_1} - \mu_{Q_2}\|_{\mathcal{H}}$;

(ii) [69] $\mathbb{E}_{z\sim Q_1}[f(z)] = \langle f, \mu_{Q_1}\rangle_{\mathcal{H}}$.

Finally, we review the result in [86] used in Section 5.

Theorem A.4. *[86, Theorem 6.1] Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be the RKHS induced by $k(z,z') = \exp(-\|z - z'\|_2^2/\sigma^2)$ on $\mathcal{Z} = [0,1]^d$. Suppose that $\ell(f^*, \cdot)$ is in the Sobolev space $H^d(\mathbb{R}^d)$. Then, there exists a constant C independent on R but dependent on f, d, and σ such that*

$$\inf_{\|g\|_{\mathcal{H}} \leq R} \|\ell(f^*, \cdot) - g\|_\infty \leq C(\log R)^{-\frac{d}{16}}, \ \forall\, R \geq 1.$$

B Proof of Theorem 4.1

Starting from the bound (8) in Theorem 3.1, we separate our proof into two parts. For $\mathbb{P}[P \notin \mathcal{K}]$, we set

$$\eta = \frac{K}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)}),$$

noting that:

Proposition B.1. *[89, Proposition A.1] Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a continuous positive definite kernel on nonempty compact set $\mathcal{Z} \subset \mathbb{R}^d$ with $\sup_{z\in\mathcal{Z}} \sqrt{k(z,z)} \leq K < \infty$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$,*

$$\left\|\mu_{\hat{P}} - \mu_P\right\|_{\mathcal{H}} \leq \frac{K}{\sqrt{n}}(1 + \sqrt{2\log(1/\delta)}).$$

For the second term $\mathbb{P}[P_{\text{dro}}\ell(f^*, \cdot) - P\ell(f^*, \cdot) > \varepsilon | P \in \mathcal{K}]$ in (8), conditional on $P \in \mathcal{K}$ we have that

$$P_{\text{dro}}\ell(f^*, \cdot) - P\ell(f^*, \cdot) = \max_{Q\in\mathcal{K}} \mathbb{E}_{z\sim Q}[\ell(f^*, z)] - \mathbb{E}_{z\sim P}[\ell(f^*, z)] = \max_{Q\in\mathcal{K}} \langle \ell(f^*, \cdot), \mu_Q - \mu_P\rangle_{\mathcal{H}},$$

where the second line follows from (18). Then, we have

$$P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) = \max_{Q\in\mathcal{K}}\langle \ell(f^*,\cdot), \mu_Q - \mu_P\rangle_{\mathcal{H}} \le \max_{Q\in\mathcal{K}} \|\ell(f^*,\cdot)\|_{\mathcal{H}}\|\mu_Q - \mu_P\|_{\mathcal{H}} \le 2M\eta,$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the last inequality holds since for all $P \in \mathcal{K}$,

$$\|\mu_Q - \mu_P\|_{\mathcal{H}} \le \|\mu_Q - \mu_{\hat{P}}\|_{\mathcal{H}} + \|\mu_{\hat{P}} - \mu_P\|_{\mathcal{H}} \le 2\eta$$

by the triangle inequality. Hence,

$$\mathbb{P}[P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) > \varepsilon | P \in \mathcal{K}] \le \mathbb{P}[\max_{P\in\mathcal{K}} \|\ell(f^*,\cdot)\|_{\mathcal{H}}\|\mu_Q - \mu_P\|_{\mathcal{H}} > \varepsilon | P \in \mathcal{K}] = 0.$$

provided that $\varepsilon \ge 2M\eta$. Therefore, with probability at least $1-\delta$,

$$P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot) \le 2M\eta = \frac{2KM}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)}).$$

This completes our proof.

C Proof of Theorem 5.1

For notational simplicity, we denote

$$g_R := \arg\inf_{\|g\|_{\mathcal{H}}\le R} \|\ell(f^*,\cdot) - g\|_\infty$$

for all $R \ge 1$, where g_R is well-defined since for fixed R, $h(g) = \|\ell(f^*,\cdot) - g\|_\infty$ is a continuous function over a compact set $\{g : \|g\|_{\mathcal{H}} \le R\}$. Then, by Theorem 3.1 and Proposition B.1, for any $\delta > 0$, once we set $\eta = \frac{1}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)})$, we have that

$$\mathbb{P}[P\ell(f^*_{\text{dro}},\cdot) - P\ell(f^*,\cdot) > \varepsilon] \le \delta + \mathbb{P}[P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) > \varepsilon | P \in \mathcal{K}].$$

By definition, we obtain

$$\begin{aligned}
P_{\text{dro}}\ell(f^*,\cdot) - P\ell(f^*,\cdot) &= \max_{Q\in\mathcal{K}}\Big\{\int \ell(f^*,z)Q(\mathrm{d}z) - \int \ell(f^*,z)P(\mathrm{d}z)\Big\} \\
&\le \max_{Q\in\mathcal{K}}\Big\{\int g_R Q(\mathrm{d}z) - \int g_R P(\mathrm{d}z)\Big\} + 2\|\ell(f^*,\cdot) - g_R\|_\infty \\
&\le \max_{Q\in\mathcal{K}}\{\|g_R\|_{\mathcal{H}}\|\mu_Q - \mu_P\|_{\mathcal{H}}\} + 2\|\ell(f^*,\cdot) - g_R\|_\infty \\
&\le 2(\eta R + \|\ell(f^*,\cdot) - g_R\|_\infty),
\end{aligned}$$

where the second equality follows from the Cauchy-Schwarz inequality. We have

$$\|\ell(f^*,\cdot) - g_R\|_\infty = I(\ell(f^*,\cdot), R) \le C(\log R)^{-d/16}$$

according to [86, Theorem 6.1] (and see [90, Proposition 18] for details on the constants). Then, taking infimum over $R \ge 1$ and combining the equations above yield the desired result.

D Extensions to Other DRO Variants

Our framework can be generalized to DROs using other statistical distances, including the Wasserstein distance and ϕ-divergences. Below, we show and derive results for these DROs, and discuss their implications and comparisons with MMD DRO. Moreover, we further develop results under the presence of distributional shift, which is a situation where our derived bounds are of potential additional value as we will discuss.

D.1 1-Wasserstein DRO

We focus on the 1-Wasserstein distance, which is defined as:

Definition D.1 (1-Wasserstein distance)**.** For two distributions $Q_1, Q_2 \in \mathcal{P}$ on domain $\mathbb{R}^d$, the 1-Wasserstein distance $D_{\mathrm{W}} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is defined by

$$D_{\mathrm{W}}(Q_1, Q_2) := \inf \left\{ \int_{\mathcal{Z}^2} \|z_1 - z_2\| \, \Pi(\mathrm{d}z_1, \, \mathrm{d}z_2) : \begin{array}{l} \Pi \text{ is a joint distribution of } z_1 \text{ and } z_2 \\ \text{with marginals } Q_1 \text{ and } Q_2, \text{ respectively} \end{array} \right\},$$

where $\|\cdot\|$ is a given norm in $\mathbb{R}^d$.

Similar to (12), we define 1-Wasserstein DRO as

$$\min_{f \in \mathcal{F}} \max_{Q : D_{\mathrm{W}}(Q, \hat{P}) \leq \eta} \mathbb{E}_{z \sim Q}[\ell(f, z)] \tag{19}$$

and denote $f^*_{\text{W-dro}}$ as the solution to (19). Also define $\|\cdot\|_{\text{Lip}}$ as the Lipschitz norm. We have the following generalization bound:

Theorem D.2 (Generalization Bounds for 1-Wasserstein DRO)**.** *Suppose that P is a light-tailed distribution in the sense that there exists $a > 1$ such that*

$$A := \mathbb{E}_{z \sim P}[\exp(\|z\|^a)] < \infty.$$

Then, there exists constants c_1 and c_2 that only depends on a, A, and d such that for any given $0 < \delta < 1$, if we choose ball size $\eta = \left(\frac{\log(c_1/\delta)}{c_2 n}\right)^{1/\max\{d,2\}}$ and $n \geq \frac{\log(c_1/\delta)}{c_2}$, then the 1-Wasserstein DRO solution $f^*_{\text{W-dro}}$ satisfies

$$\mathbb{E}_P \ell(f^*_{\text{W-dro}}, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) \leq 2\|\ell(f^*, \cdot)\|_{\text{Lip}} \left(\frac{\log(c_1/\delta)}{c_2 n}\right)^{1/\max\{d,2\}}$$

with probability at least $1 - \delta$.

To cast in our main framework described in Sections 3 and 4, the generalization bound in Theorem D.2 depends only on $\ell(f^*, \cdot)$ but not other candidates in the hypothesis class. Theorem D.2 is obtained via analyzing the two components in Theorem 3.1, which respectively correspond to: (1) the ambiguity set contains true distribution P with high confidence and (2) compatibility of the Wasserstein distance with $\ell(f^*, \cdot)$. For the first component, we use the concentration of $D_{\mathrm{W}}(P, \hat{P})$ established in [91, Theorem 2]:

Lemma D.3. *[91, Theorem 2] Suppose that P is a light-tailed distribution in the sense that there exists $a > 1$ such that*

$$A := \mathbb{E}_{z \sim P}[\exp(\|z\|^a)] < \infty.$$

Then, there exists constants c_1 and c_2 that only depends on a, A, and d such that for all $\delta \geq 0$, if $n \geq \frac{\log(c_1/\delta)}{c_2}$, then

$$D_W(P, \hat{P}) \leq \left(\frac{\log(c_1/\delta)}{c_2 n}\right)^{1/\max\{d,2\}}$$

with probability at least $1 - \delta$.

For the second component, we use the bound

$$\mathbb{E}_{Q_1}[\ell(f^*, \cdot)] - \mathbb{E}_{Q_2}[\ell(f^*, \cdot)] \leq \|\ell(f^*, \cdot)\|_{\text{Lip}} D_{\mathrm{W}}(Q_1, Q_2)$$

which follows from the dual representation of $D_{\mathrm{W}}(\cdot, \cdot)$ [32]. With these, a similar analysis as that for Theorem 4.1 yields Theorem D.2. The detailed proof is deferred to Section F.1.

D.2 ϕ-divergence DRO

Next we consider the use of ϕ-divergence, defined as:

Definition D.4 (ϕ-divergence)**.** Let Q_1 and Q_2 be two probability distributions over a space Ω such that $Q_1 \ll Q_2$, i.e., Q_1 is absolutely continuous with respect to Q_2. Then, for a convex function $\phi : [0, \infty) \to (-\infty, \infty]$ such that $\phi(x)$ is finite for all $x > 0, \phi(1) = 0$, and $\phi(0) = \lim_{t\to 0^+} \phi(t)$ (which could be infinite), the ϕ-divergence of Q_1 from Q_2 is defined as

$$D_\phi(Q_1 \| Q_2) = \int_\Omega \phi\Big(\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\Big)\mathrm{d}Q_2.$$

Using the ϕ-divergence for DRO entails a bit more intricacies than Wasserstein DRO due to the absolute continuity requirement in defining the ϕ-divergence. In particular, if the true distribution P is continuous, then setting the center of a ϕ-divergence ball as the empirical distribution $\hat{P}$ would make the ball never cover the truth, which violates the conditions required in developing our generalization bounds. To remedy, one can set the ball center using kernel density estimate or a parametric distribution, albeit with the price of additional approximation errors coming from the "smoothing" of the empirical distribution. In the following, we first present a general bound for ϕ-divergence DRO that leaves open the choice of the ball center, which we denote P_0, and the function ϕ. We then present a more explicit bound for the case of χ^2-divergence and when the true distribution P is finitely discrete, for which we could use the empirical distribution $\hat{P}$ as the ball center and hence avoid the additional smoothing error. We would leave the other explicit bounds, including the use of the remedial smoothing distributions in the continuous case, to separate future work.

Like above, we define ϕ-divergence DRO as

$$\min_{f\in\mathcal{F}} \max_{Q: D_\phi(Q\|P_0)\le\eta} \mathbb{E}_{z\sim Q}[\ell(f, z)] \tag{20}$$

where P_0 is a baseline distribution constructed from iid samples $\{z_i\}_{i=1}^n$. We denote $f^*_{\phi\text{-dro}}$ as the solution to (20). We have the following result:

Theorem D.5 (General Generalization Bounds for ϕ-divergence DRO)**.** *Suppose that for a given $0 < \delta < 1$, we construct ambiguity set $\mathcal{K} = \{Q : D_\phi(Q\|P_0) \le \eta\}$ such that*

(i) (high-confidence ambiguity set) $\mathbb{P}[P \in \mathcal{K}] \ge 1 - \delta/2$;

(ii) (bound of DRO above the baseline) $\max_{Q\in\mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_{P_0}\ell(f^*, \cdot) \le C_1(\ell(f^*, \cdot), \sqrt{\eta}, \phi)$;

(iii) (approximation error) $\mathbb{E}_{P_0}\ell(f^*, \cdot) - \mathbb{E}_P\ell(f^*, \cdot) \le C_2(\ell(f^*, \cdot), n, \delta)$ with probability at least $1 - \delta/2$.

where C_1 and C_2 are quantities that depend on the respective arguments. Then, the ϕ-divergence DRO solution $f^*_{\phi-dro}$ satisfies

$$\mathbb{E}_P\ell(f^*_{\phi-dro}, \cdot) - \mathbb{E}_P\ell(f^*, \cdot) \le C_1(\ell(f^*, \cdot), \sqrt{\eta}, \phi) + C_2(\ell(f^*, \cdot), n, \delta)$$

with probability at least $1 - \delta$.

Theorem D.5 is derived by expanding the analysis for Theorem 3 to incorporate the error between the DRO optimal value over its baseline P_0, and also the approximation error induced by P_0 over P. This additional decomposition facilitates the analysis of specific ϕ-divergences. The proof is deferred to Section F.2. Below, we show an example of applying it to the χ^2-divergence with P_0 set to be the empirical distribution in the finitely discrete case.

Definition D.6 (χ^2-divergence)**.** Let Q_1 and Q_2 be two probability distributions over a space Ω such that $Q_1 \ll Q_2$, i.e., Q_1 is absolutely continuous with respect to Q_2. Then, the χ^2-divergence of Q_1 from Q_2 is defined as

$$D_{\chi^2}(Q_1\|Q_2) = \int_\Omega \Big(\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2} - 1\Big)^2 \Big/ \Big(\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\Big)\mathrm{d}Q_2.$$

With this, we consider χ^2-divergence DRO

$$\min_{f\in\mathcal{F}} \max_{Q:D_\chi^2(Q\|\hat{P})\le\eta} \mathbb{E}_{z\sim Q}[\ell(f,z)] \tag{21}$$

where we now set the empirical distribution $\hat{P}$ as the ball center. We denote $f^*_{\chi^2\text{-dro}}$ as the solution to (21). By utilizing Theorem D.5 and the finite-sample approximation error of χ^2-statistics established in [92], we have the following result:

Theorem D.7 (Generalization Bounds for χ^2-divergence DRO on Discrete Distributions)**.** *Suppose that P is a discrete distribution with m support with $\mathbb{P}[z = z_i] = p_i$. Suppose that (i) $p_{\min} = \min_{1\le i\le m} p_i > 0$; (ii) $\|\ell(f^*,\cdot)\|_\infty < +\infty$. Then, for all $0<\delta<1$, if we choose ball size*

$$\eta = \frac{1}{n}\Big(m + 2\log(4/\delta) + 2\sqrt{m\log(4/\delta)}\Big)$$

for χ^2-divergence DRO, then for all $n \ge \frac{10^6 m^2}{p_{\min}^3\delta^2}$, the χ^2-divergence DRO solution $f^*_{\chi^2\text{-dro}}$ satisfies

$$\mathbb{E}_P\ell(f^*_{\chi^2\text{-dro}},\cdot) - \mathbb{E}_P\ell(f^*,\cdot) \le \|\ell(f^*,\cdot)\|_\infty\Big\{\frac{1}{\sqrt{n}}\sqrt{m + 2\log(4/\delta) + 2\sqrt{m\log(4/\delta)}} + \sqrt{\frac{2\log(2/\delta)}{n}}\Big\}$$

with probability at least $1-\delta$.

The detailed proof is deferred to Section F.3. Once again, the bound in Theorem D.7 involves only $\ell(f^*,\cdot)$ but not other candidates in the hypothesis class. To consider cases using smoothed P_0 such as kernel density estimate and with continuous ground-truth P, we would need to derive bounds on their divergence and analyze the approximation error on the expected loss between P_0 and P. These are both substantial investigation that we delegate to future work.

Note that in our result for Wasserstein DROs, the rates in terms of n is $n^{-1/\max\{d,2\}}$, which is slower than that for MMD DRO in Theorem 4.1. The slow rate of Wasserstein DRO comes from the large ball size needed to confidently cover the true P using the distance (i.e., to control the first term in (8)). This shows a tradeoff between the hypothesis class dependence and the rate in n that differs from ERM. More precisely, Wasserstein DRO has a generalization that favorably relies only on the true loss $\ell(f^*,\cdot)$, but scales inferiorly in terms of n. It is thus most suited when the complexity of hypothesis class outweighs the convergence rate consideration. In the next section, we will describe the distribution shift setting where such a tradeoff can be amplified.

We also note that, unlike Wasserstein DRO, χ^2-divergence DRO retains a rate of $1/\sqrt{n}$. However, this is because Theorem D.7 has focused on the finite-support case, where the number of support points m comes into play in the multiplicative factor instead of the rate. If we consider continuous P and use a smooth baseline distribution, we expect the rate of χ^2-divergence DRO to deteriorate much like the case of Wasserstein DRO.

E Extensions to Distributional Shift Setting

We consider the setting of distributional shift as follows. We have iid samples from the training distribution P_{train}. Our testing distribution P_{test}, however, is different from P_{train}, i.e., $P_{\text{test}} \ne P_{\text{train}}$. Let

$$f^*_{\text{test}} \in \text{argmin}_{f\in\mathcal{F}}\mathbb{E}_{P_{\text{test}}}[\ell(f,\cdot)]$$

be the oracle best solution under P_{test}. Our goal is derive a data-driven solution f_{data} to bound the excess risk for the *testing* objective

$$\mathbb{E}_{P_{\text{test}}}\ell(f_{\text{data}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot). \tag{22}$$

We first provide a general DRO bound as Theorem 3.1 that allows for distribution shift:

Theorem E.1 (A General DRO Bound under Distributional Shift)**.** *Let $\mathcal{Z}\subset\mathbb{R}^d$ be a sample space, P_{train} be a distribution on $\mathcal{Z}$, $\{z_i\}_{i=1}^n$ be iid samples from P_{train}, and $\mathcal{F}$ be the function class. For loss function $\ell:\mathcal{F}\times\mathcal{Z}\to\mathbb{R}$ and DRO $\min_{f\in\mathcal{F}}\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f,\cdot)$, DRO solution f^*_{dro} satisfies that for any $\varepsilon>0$,*

$$\begin{aligned}&\mathbb{P}[\mathbb{E}_{P_{test}}\ell(f^*_{dro},\cdot) - \mathbb{E}_{P_{test}}\ell(f^*_{test},\cdot) > \varepsilon]\\ &\le \mathbb{P}[P_{test}\notin\mathcal{K}] + \mathbb{P}[\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*_{test},\cdot) - \mathbb{E}_{P_{test}}\ell(f^*_{test},\cdot) > \varepsilon | P_{test}\in\mathcal{K}].\end{aligned} \tag{23}$$

Proof. The proof is identical to that of Theorem 3.1, except that we put subscript "test" in P and f^* there. □

Theorem E.1, and its proof that is virtually identical to that of Theorem 3.1, shows that our main framework in Section 3 can be naturally employed under distribution shift. In particular, the ambiguity set $\mathcal{K}$ needs to be constructed from the training set, otherwise there is no restriction on its application. The key now is the choice of ambiguity set size and how it distorts generalization more explicitly when distribution shift happens. Below, we investigate the cases of MMD DRO and 1-Wasserstein DRO whose explicit distribution shift bounds could be readily derived.

First, we define MMD DRO under distribution shift as

$$\min_{f\in\mathcal{F}} \max_{Q:D_{\text{MMD}}(Q,\hat{P}_{\text{train}})\leq\eta} \mathbb{E}_{z\sim Q}[\ell(f,z)] \tag{24}$$

where $\hat{P}_{\text{train}}$ is the empirical distribution constructed by iid samples $\{z_i\}_{i=1}^n$ from P_{train}. We now denote f^*_{dro} as the solution of (25).

Theorem E.2 (Generalization Bounds for MMD DRO under Distributional Shift)**.** *Let $\mathcal{Z}$ be a compact subspace of $\mathbb{R}^d$, $k:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ be a bounded continuous positive definite kernel, and $(\mathcal{H},\langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the corresponding RKHS. Suppose also that (i) $\sup_{z\in\mathcal{Z}}\sqrt{k(z,z)}\leq K$; (ii) $\ell(f^*_{test},\cdot)\in\mathcal{H}$ with $\|\ell(f^*_{test},\cdot)\|_{\mathcal{H}}<+\infty$; (iii) $D_{MMD}(P_{train},P_{test})<+\infty$. Then, for all $\delta\geq 0$, if we choose ball size $\eta = D_{MMD}(P_{train},P_{test})+\frac{K}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)})$ for MMD DRO, then MMD DRO solution f^*_{dro} satisfies*

$$\mathbb{E}_{P_{test}}\ell(f^*_{dro},\cdot)-\mathbb{E}_{P_{test}}\ell(f^*_{test},\cdot)\leq 2\|\ell(f^*_{test},\cdot)\|_{\mathcal{H}}\Big\{D_{MMD}(P_{train},P_{test})+\frac{K}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)})\Big\}$$

with probability at least $1-\delta$.

Compared to Theorem 4.1, Theorem E.2 has the extra condition that $D_{\text{MMD}}(P_{\text{train}},P_{\text{test}})<+\infty$ which signifies the amount of distribution shift. Correspondingly the MMD ball size needs to be chosen to cover this shift amount. We provide the proof that details these modifications in Section F.4.

Next, we define 1-Wasserstein DRO under distribution shift as

$$\min_{f\in\mathcal{F}} \max_{Q:D_{\text{W}}(Q,\hat{P}_{\text{train}})\leq\eta} \mathbb{E}_{z\sim Q}[\ell(f,z)] \tag{25}$$

where $\hat{P}_{\text{train}}$ is the empirical distribution constructed by iid samples $\{z_i\}_{i=1}^n$ from P_{train}. We now denote $f^*_{\text{W-dro}}$ as the solution of (25).

Theorem E.3 (Generalization Bounds for 1-Wasserstein DRO under Distributional Shift)**.** *Suppose that $d_W(P_{train},P_{test})<+\infty$ and P_{train} is a light-tailed distribution in the sense that there exists $a>1$ such that*

$$A:=\mathbb{E}_{z\sim P}[\exp(\|z\|^a)]<\infty.$$

Then, there exists constants c_1 and c_2 that only depends on a, A, and d such that for all $\delta\geq 0$, if we choose ball size $\eta = d_W(P_{train},P_{test})+\left(\frac{\log(c_1/\delta)}{c_2 n}\right)^{1/\max\{d,2\}}$ and $n\geq\frac{\log(c_1/\delta)}{c_2}$, then 1-Wasserstein DRO solution $f^*_{W\text{-}dro}$ satisfies

$$\mathbb{E}_{P_{test}}\ell(f^*_{W\text{-}dro},\cdot)-\mathbb{E}_{P_{test}}\ell(f^*_{test},\cdot)\leq 2\|\ell(f^*_{test},\cdot)\|_{Lip}\Big\{D_W(P_{train},P_{test})+\Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}\Big\}$$

with probability at least $1-\delta$.

Similar to the MMD DRO case, Theorem E.3, compared to Theorem D.2, involves the amount of distribution shift $d_{\text{W}}(P_{\text{train}},P_{\text{test}})<+\infty$ and correspondingly a ball size that is chosen to cover this amount. The detailed proof of Theorem E.3 is deferred to Section F.5.

With the generalization bounds in Theorems E.2 and E.3, we now explain how MMD and Wasserstein DROs can have potential amplified strengths over ERM under distribution shift. To this end, let us consider using the classical analysis of ERM to bound the excess risk for the testing objective (22). This entails a decomposition

$$\begin{aligned}&\mathbb{E}_{P_{\text{test}}}\ell(\hat{f}^*,\cdot)-\mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot)\\ =&[\mathbb{E}_{P_{\text{test}}}\ell(\hat{f}^*,\cdot)-\mathbb{E}_{\hat{P}_{\text{train}}}\ell(\hat{f}^*,\cdot)]+[\mathbb{E}_{\hat{P}_{\text{train}}}\ell(\hat{f}^*,\cdot)-\mathbb{E}_{\hat{P}_{\text{train}}}\ell(f^*,\cdot)]+[\mathbb{E}_{\hat{P}_{\text{train}}}\ell(f^*,\cdot)-\mathbb{E}_{P_{\text{test}}}\ell(f^*,\cdot)]\end{aligned}$$

where the second term is ≤ 0 and the third term depends only on the true loss $\ell(f^*, \cdot)$. The key is to analyze the first term, which we decompose into

$$\mathbb{E}_{P_{\text{test}}}\ell(\hat{f}^*, \cdot) - \mathbb{E}_{\hat{P}_{\text{train}}}\ell(\hat{f}^*, \cdot) = [\mathbb{E}_{P_{\text{test}}}\ell(\hat{f}^*, \cdot) - \mathbb{E}_{P_{\text{train}}}\ell(\hat{f}^*, \cdot)] + [\mathbb{E}_{P_{\text{train}}}\ell(\hat{f}^*, \cdot) - \mathbb{E}_{\hat{P}_{\text{train}}}\ell(\hat{f}^*, \cdot)].$$

If we use the standard approach of uniformly bounding over the hypothesis class $\mathcal{L}$, then from the first component above we would obtain a bound that involves

$$\sup_{f \in \mathcal{F}} \|l(f, \cdot)\|_{\mathcal{H}} D_{\text{MMD}}(P_{\text{train}}, P_{\text{test}}) \tag{26}$$

or

$$\sup_{f \in \mathcal{F}} \|l(f, \cdot)\|_{\text{Lip}} D_{\text{W}}(P_{\text{train}}, P_{\text{test}}). \tag{27}$$

Now, consider the situation where there is a large distribution shift, i.e., $D(P_{\text{train}}, P_{\text{test}})$ is big, to the extent that it dominates the error coming from the training set noise that scales with n. In this case, the quantity (26) or (27) could become the dominant term in the ERM bound. In contrast, in MMD DRO and Wasserstein DRO, the dominant terms are

$$2\|\ell(f^*_{\text{test}}, \cdot)\|_{\mathcal{H}} D_{\text{MMD}}(P_{\text{train}}, P_{\text{test}}), \tag{28}$$

and

$$2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}} D_{\text{W}}(P_{\text{train}}, P_{\text{test}}) \tag{29}$$

respectively. Comparing (28) and (29) with (26) and (27), we see that ERM could amplify the impact of distribution shift $D_{\text{MMD}}(P_{\text{train}}, P_{\text{test}})$ or $D_{\text{W}}(P_{\text{train}}, P_{\text{test}})$ by a factor that involves a uniform bound on $\{\ell(f, \cdot) : f \in \mathcal{F}\}$, while the corresponding factor in DROs only involves the true loss function $\ell(f^*_{\text{test}}, \cdot)$. This reveals a potentially significant benefit of DRO over ERM in the presence of distribution shift.

F Missing Proofs of Additional Results

F.1 Proof of Theorem D.2

Starting from the bound (8) in Theorem 3.1, we separate our proof into two parts. First, to control $\mathbb{P}[P \notin \mathcal{K}]$, we set

$$\eta = \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}.$$

Then, by Lemma D.3, for all $n \geq \frac{\log(c_1/\delta)}{c_2}$, we have that $\mathbb{P}[P \notin \mathcal{K}] \leq \delta$. For the second term $\mathbb{P}[\max_{Q \in \mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) > \varepsilon | P \in \mathcal{K}]$ in (8), conditional on $P \in \mathcal{K}$ we have that

$$\max_{Q \in \mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) \leq \max_{Q \in \mathcal{K}} \|\ell(f^*, \cdot)\|_{\text{Lip}} D_{\text{W}}(Q, P),$$

by using [93]. Then, we have

$$\begin{aligned}\max_{Q \in \mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) &\leq \max_{Q \in \mathcal{K}} \|\ell(f^*, \cdot)\|_{\text{Lip}} D_{\text{W}}(Q, P) \\ &\leq \max_{Q \in \mathcal{K}} \|\ell(f^*, \cdot)\|_{\text{Lip}} (D_{\text{W}}(Q, \hat{P}) + D_{\text{W}}(\hat{P}, P)) \leq 2\|\ell(f^*, \cdot)\|_{\text{Lip}} \eta,\end{aligned}$$

where the last line follows from the triangle inequality. Hence,

$$\mathbb{P}[\max_{Q \in \mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) > \varepsilon | P \in \mathcal{K}] \leq \mathbb{P}[2\|\ell(f^*, \cdot)\|_{\text{Lip}} \eta > \varepsilon | P \in \mathcal{K}] = 0.$$

provided that $\varepsilon \geq 2\|\ell(f^*, \cdot)\|_{\text{Lip}} \eta$. Therefore, with probability at least $1 - \delta$,

$$\max_{Q \in \mathcal{K}} \mathbb{E}_Q \ell(f^*, \cdot) - \mathbb{E}_P \ell(f^*, \cdot) \leq 2\|\ell(f^*, \cdot)\|_{\text{Lip}} \eta = 2\|\ell(f^*, \cdot)\|_{\text{Lip}} \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}.$$

This completes our proof.

F.2 Proof of Theorem D.5

For simplicity, let

$$\mathcal{E}_1 = \{P \in \mathcal{K}\}, \quad \mathcal{E}_2 = \{\mathbb{E}_{P_0}\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) \le C_2(\ell(f^*,\cdot),n,\delta)\}.$$

Similar to proof of Theorem 3.1, we have that

$$\begin{aligned}
&\mathbb{P}[\mathbb{E}_P\ell(f^*_{\text{dro}},\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon] \\
&\le \mathbb{P}[[\mathbb{E}_P\ell(f^*_{\text{dro}},\cdot) - \max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*_{\text{dro}},\cdot)] + [\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot)] > \varepsilon] \\
&\le \mathbb{P}[\mathcal{E}_1^{\complement}] + \mathbb{P}[\{[\mathbb{E}_P\ell(f^*_{\text{dro}},\cdot) - \max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*_{\text{dro}},\cdot)] + [\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot)] > \varepsilon\} \cap \mathcal{E}_1] \\
&\le \delta/2 + \mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon\} \cap \mathcal{E}_1],
\end{aligned}$$

where the last line follows from $\mathbb{P}[\mathcal{E}_1^{\complement}] \le \delta/2$ by definition and $\mathbb{E}_P\ell(f^*_{\text{dro}},\cdot) - \max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*_{\text{dro}},\cdot) \le 0$ if $\mathcal{E}_1$ occurs according to worst case definition of DRO. Now we consider

$$\mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon\} \cap \mathcal{E}_1],$$

where we set $\varepsilon = C_1(\ell(f^*,\cdot),\sqrt{\eta},\phi) + C_2(\ell(f^*,\cdot),n,\delta)$. This can be rewritten as

$$\begin{aligned}
&\mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon\} \cap \mathcal{E}_1] \\
&= \mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon\} \cap \mathcal{E}_1 \cap \mathcal{E}_2] + \mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > \varepsilon\} \cap \mathcal{E}_1 \cap \mathcal{E}_2^{\complement}] \\
&\le \mathbb{P}[\{\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_{P_0}\ell(f^*,\cdot) > \varepsilon - C_2(\ell(f^*,\cdot),n,\delta) \cap \mathcal{E}_1 \cap \mathcal{E}_2\}] + \mathbb{P}[\mathcal{E}_2^{\complement}] \\
&\le 0 + \delta/2,
\end{aligned}$$

where the last line follows from

$$\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_{P_0}\ell(f^*,\cdot) \le C_1(\ell(f^*,\cdot),\sqrt{\eta},\phi) < \varepsilon - C_2(\ell(f^*,\cdot),n,\delta).$$

This completes our proof.

F.3 Proof of Theorem D.7

It suffices to verify assumptions (i)(ii)(iii) in Theorem D.5 with $C_1(\ell(f^*,\cdot),\sqrt{\eta},\phi) = \sqrt{\eta}\|\ell(f^*,\cdot)\|_\infty$, $C_2(\ell(f^*,\cdot),n,\delta) = \|\ell(f^*,\cdot)\|_\infty\sqrt{\frac{2\log(2/\delta)}{n}}$.

First, by [92, (4.2)], we have that

$$\mathbb{P}[\chi^2(P\|\hat{P}) > \eta] \le \mathbb{P}[\chi^2_{(m-1)} > n\eta] + \frac{250m}{\sqrt{n}p_{\min}^{3/2}},$$

where $\chi^2_{(m-1)}$ is a χ^2 random variable with degree of freedom $m-1$. According to [94, Lemma 1],

$$\mathbb{P}[\chi^2_{(m-1)} > (m-1) + 2\sqrt{(m-1)x} + 2x] \le e^{-x}.$$

Therefore, for all $n \ge \frac{10^6 m^2}{p_{\min}^3\delta^2}$,

$$\mathbb{P}[\mathbb{P} \notin \mathcal{K}] = \mathbb{P}[\chi^2(P\|\hat{P}) > \eta] \le \delta/4 + \delta/4 = \delta/2$$

and (i) is satisfied. Second, note that

$$\max_{Q\in\mathcal{K}}\mathbb{E}_Q\ell(f^*,\cdot) - \mathbb{E}_{\hat{P}}\ell(f^*,\cdot) = \max_{Q\in\mathcal{K}}\mathbb{E}_{\hat{P}}[(L-1)\ell(f^*,\cdot)],$$

where $L = \frac{\mathrm{d}Q}{\mathrm{d}\hat{P}}$ is the likelihood ratio. Then, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\mathbb{E}_{\hat{P}}[(L-1)\ell(f^*,\cdot)]\right| &\le \sqrt{\mathbb{E}_{\hat{P}}[((L-1)/\sqrt{L})^2]}\sqrt{\mathbb{E}_{\hat{P}}[(\sqrt{L}\ell(f^*,\cdot))^2]} \\
&\le \sqrt{\mathbb{E}_{\hat{P}}[(L-1)^2/L]}\sqrt{\mathbb{E}_Q[\ell(f^*,\cdot)^2]} \\
&\le \sqrt{\chi^2(Q\|\hat{P})}\sqrt{\|\ell(f^*,\cdot)\|_\infty^2} = \sqrt{\eta}\|\ell(f^*,\cdot)\|_\infty,
\end{aligned}$$

This gives (ii). Third, by Hoeffding's inequality,

$$\mathbb{P}[\mathbb{E}_{P_0}\ell(f^*,\cdot) - \mathbb{E}_P\ell(f^*,\cdot) > t] \le e^{-\frac{nt^2}{2\|\ell(f^*,\cdot)\|_\infty^2}}$$

and (iii) is satisfied. This completes our proof.

F.4 Proof of Theorem E.2

Starting from the bound (23) in Theorem E.1, we separate our proof into two parts. First, to control $\mathbb{P}[P_{\text{test}} \notin \mathcal{K}]$, we set

$$\eta = D_{\text{MMD}}(P_{\text{train}}, P_{\text{test}}) + \frac{K}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)}).$$

Then, by Proposition B.1, we have that

$$\mathbb{P}[P_{\text{test}} \notin \mathcal{K}] = \mathbb{P}[D_{\text{MMD}}(P_{\text{test}}, \hat{P}_{\text{train}}) > \eta] \le \mathbb{P}\Big[D_{\text{MMD}}(P_{\text{train}}, \hat{P}_{\text{train}}) > \frac{K}{\sqrt{n}}(1+\sqrt{2\log(1/\delta)})\Big] \le 1-\delta,$$

For the second term $\mathbb{P}[\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) > \varepsilon | P_{\text{test}} \in \mathcal{K}]$ in (8), conditional on $P_{\text{test}} \in \mathcal{K}$, we have that

$$\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) = \max_{Q\in\mathcal{K}} \langle \ell(f^*_{\text{test}},\cdot), \mu_Q - \mu_{P_{\text{test}}}\rangle_{\mathcal{H}},$$

which follows from (18). Then, we have

$$\begin{aligned}\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) &= \max_{Q\in\mathcal{K}} \langle \ell(f^*_{\text{test}},\cdot), \mu_Q - \mu_{P_{\text{test}}}\rangle_{\mathcal{H}} \\ &\le \max_{Q\in\mathcal{K}} \|\ell(f^*_{\text{test}},\cdot)\|_{\mathcal{H}} \|\mu_Q - \mu_{P_{\text{test}}}\|_{\mathcal{H}} \le 2\|\ell(f^*_{\text{test}},\cdot)\|_{\mathcal{H}}\eta,\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the last inequality holds since for all $P \in \mathcal{K}$,

$$\|\mu_Q - \mu_{P_{\text{test}}}\|_{\mathcal{H}} \le \|\mu_Q - \mu_{\hat{P}_{\text{train}}}\|_{\mathcal{H}} + \|\mu_{\hat{P}_{\text{train}}} - \mu_{P_{\text{test}}}\|_{\mathcal{H}} \le 2\eta$$

by the triangle inequality. Hence,

$$\mathbb{P}[\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) > \varepsilon | P_{\text{test}} \in \mathcal{K}] \le \mathbb{P}[2\|\ell(f^*_{\text{test}},\cdot)||_{\mathcal{H}}\eta > \varepsilon | P_{\text{test}} \in \mathcal{K}] = 0.$$

provided that $\varepsilon \ge 2\|\ell(f^*_{\text{test}},\cdot)||_{\mathcal{H}}\eta$. Therefore, with probability at least $1-\delta$,

$$\begin{aligned}&\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_P\ell(f^*_{\text{test}},\cdot) \le 2\|\ell(f^*_{\text{test}},\cdot)||_{\mathcal{H}}\eta \\ &\le 2\|\ell(f^*_{\text{test}},\cdot)\|_{\mathcal{H}}\Big\{D_{\text{W}}(P_{\text{train}}, P_{\text{test}}) + \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}\Big\}.\end{aligned}$$

This completes our proof.

F.5 Proof of Theorem E.3

Starting from the bound (23) in Theorem E.1, we separate our proof into two parts. First, to control $\mathbb{P}[P_{\text{test}} \notin \mathcal{K}]$, we set

$$\eta = D_{\text{W}}(P_{\text{train}}, P_{\text{test}}) + \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}.$$

Then, by Lemma D.3, for all $n \ge \frac{\log(c_1/\delta)}{c_2}$, we have that

$$\mathbb{P}[P_{\text{test}} \notin \mathcal{K}] = \mathbb{P}[D_{\text{W}}(P_{\text{test}}, \hat{P}_{\text{train}}) > \eta] \le \mathbb{P}\Big[D_{\text{W}}(P_{\text{train}}, \hat{P}_{\text{train}}) > \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}}\Big] \le 1-\delta,$$

For the second term $\mathbb{P}[\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) > \varepsilon | P_{\text{test}} \in \mathcal{K}]$ in (8), conditional on $P_{\text{test}} \in \mathcal{K}$ we have that

$$\max_{Q\in\mathcal{K}} \mathbb{E}_Q\ell(f^*_{\text{test}},\cdot) - \mathbb{E}_{P_{\text{test}}}\ell(f^*_{\text{test}},\cdot) \le \max_{Q\in\mathcal{K}} \|\ell(f^*_{\text{test}},\cdot)||_{\text{Lip}} D_{\text{W}}(Q, P_{\text{test}}),$$

by using [93]. Then, we have

$$\begin{aligned}
&\max_{Q\in\mathcal{K}} \mathbb{E}_Q \ell(f^*_{\text{test}}, \cdot) - \mathbb{E}_{P_{\text{test}}} \ell(f^*_{\text{test}}, \cdot) \\
\leq\ &\max_{Q\in\mathcal{K}} \|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}} D_{\text{W}}(Q, P_{\text{test}}) \\
\leq\ &\max_{Q\in\mathcal{K}} \|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}} (D_{\text{W}}(Q, \hat{P}_{\text{train}}) + D_{\text{W}}(\hat{P}_{\text{train}}, P_{\text{test}})) \leq 2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}}\eta,
\end{aligned}$$

where the last line follows from the triangle inequality. Hence,

$$\mathbb{P}[\max_{Q\in\mathcal{K}} \mathbb{E}_Q \ell(f^*_{\text{test}}, \cdot) - \mathbb{E}_{P_{\text{test}}} \ell(f^*_{\text{test}}, \cdot) > \varepsilon | P_{\text{test}} \in \mathcal{K}] \leq \mathbb{P}[2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}}\eta > \varepsilon | P_{\text{test}} \in \mathcal{K}] = 0.$$

provided that $\varepsilon \geq 2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}}\eta$. Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{Q\in\mathcal{K}} \mathbb{E}_Q \ell(f^*_{\text{test}}, \cdot) - \mathbb{E}_P \ell(f^*_{\text{test}}, \cdot) \leq 2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}}\eta \\
&\leq 2\|\ell(f^*_{\text{test}}, \cdot)\|_{\text{Lip}} \Big\{ D_{\text{W}}(P_{\text{train}}, P_{\text{test}}) + \Big(\frac{\log(c_1/\delta)}{c_2 n}\Big)^{1/\max\{d,2\}} \Big\}.
\end{aligned}$$

This completes our proof.