Hedging against Complexity: Distributionally Robust Optimization with Parametric Approximation

Garud Iyengar
Columbia University

Henry Lam
Columbia University

Tianyu Wang
Columbia University

Abstract

Empirical risk minimization (ERM) and distributionally robust optimization (DRO) are popular approaches for solving stochastic optimization problems that appear in operations management and machine learning. Existing generalization error bounds for these methods depend on either the complexity of the cost function or dimension of the uncertain parameters; consequently, the performance of these methods is poor for high-dimensional problems with objective functions under high complexity. We propose a simple approach in which the distribution of uncertain parameters is approximated using a parametric family of distributions. This mitigates both sources of complexity; however, it introduces a model misspecification error. We show that this new source of error can be controlled by suitable DRO formulations. Our proposed parametric DRO approach has significantly improved generalization bounds over existing ERM / DRO methods and parametric ERM for a wide variety of settings. Our method is particularly effective under distribution shifts. We also illustrate the superior performance of our approach on both synthetic and real-data portfolio optimization and regression tasks.

1 Introduction

The goal of data-driven stochastic optimization is to solve

$$\min_{x\in\mathcal{X}} \{Z(x) := \mathbb{E}_{\xi\sim\mathbb{P}^*}[h(x;\xi)]\}, \tag{1}$$

where $x \in \mathcal{X}$ is the decision, ξ is a random perturbation in the sample space Ξ distributed according to $\mathbb{P}^*$, and $h : \mathcal{X} \times \Xi \to \mathbb{R}$ is the cost function. Here, we assume $\mathbb{P}^*$ is unknown and we only have access to i.i.d. samples $\hat{\xi}_i \sim \mathbb{P}^*$, $i = 1, \ldots, n$. This problem setting arises ubiquitously from machine learning to various applications involving decision making (Shapiro et al. (2014); Birge and Louveaux (2011)).

To tackle the above problem, the commonest method is to replace the unknown $\mathbb{P}^*$ with the empirical measure $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{\xi}_i}$ in (1), leading to the empirical risk minimization (ERM) problem (Hastie et al. (2009)):

$$\min_{x\in\mathcal{X}} \left\{ \hat{Z}(x) := \mathbb{E}_{\hat{\mathbb{P}}_n}[h(x;\xi)] = \frac{1}{n}\sum_{i=1}^n h(x;\hat{\xi}_i), \right\}. \tag{2}$$

A second approach that is surging in popularity over recent years is distributionally robust optimization (DRO), where the unknown $\mathbb{P}^*$ is replaced by the worst-case distribution over a so-called ambiguity set $\mathcal{A}$, giving rise to:

$$\min_{x\in\mathcal{X}} \left\{ \hat{Z}(x) := \max_{\mathbb{P}\in\mathcal{A}} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] \right\}. \tag{3}$$

Here, $\mathcal{A}$ is constructed using the data and, at least intuitively, by selecting a $\mathcal{A}$ that covers the ground-truth $\mathbb{P}^*$ with high confidence, (3) outputs a solution with a worst-case performance guarantee. In order to guarantee a statistically consistent solution, it is common to set $\mathcal{A} = \{\mathbb{P}|d(\mathbb{P}, \hat{\mathbb{P}}_n) \le \varepsilon_n\}$ for some statistical distance d and let ε_n shrink to zero as n increases. This approach has been studied with d set to Wasserstein distance (Esfahani and Kuhn (2018); Blanchet and Murthy (2019)), f-divergence (Ben-Tal et al. (2013)), kernel distance (Staib and Jegelka (2019)) and other variants.

In this paper, we are interested in bounding the generalization error

$$\mathcal{E}(\hat{x}) := Z(\hat{x}) - Z(x^*), \tag{4}$$

where $\hat{x}$ is an approximate solution and $x^* \in \arg\min_{x\in\mathcal{X}} Z(x)$ denotes the oracle best solution. $\mathcal{E}(\hat{x})$ captures the true objective performance of $\hat{x}$ relative to x^*, thus providing a direct measurement of the suboptimality of $\hat{x}$. Bounds on $\mathcal{E}(\hat{x})$ are typically of the form $\mathcal{E}(\hat{x}) \le \frac{B}{n^\alpha}$ where $B, \alpha > 0$ are method-dependent: For

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

ERM, $\alpha = \frac{1}{2}$ and B depends on the complexity of the hypothesis class, i.e., $\{h(x,\xi) : x \in \mathcal{X}\}$, represented by well-known notions such as the Vapnik–Chervonenki (VC) dimension (Vapnik (1999); Bartlett and Mendelson (2002)) and local Rademacher complexity (Bartlett et al. (2005); Xu and Zeevi (2020)). On the other hand, DRO can be analyzed by two mainstream viewpoints. One treats DRO as a regularization of ERM (where the regularizer depends on the choice of d (Duchi and Namkoong (2019); Gotoh et al. (2021); Lam (2019); Blanchet et al. (2019b); Gao (2022)), which gives rise to similar α and B as ERM. In the second approach the ambiguity set $\mathcal{A}$ is constructed as a (nonparametric) confidence region for $\mathbb{P}^*$ resulting in a a worst-case performance bound on $\hat{x}$ (Esfahani and Kuhn (2018); Bertsimas et al. (2018); Delage and Ye (2010); Wiesemann et al. (2014)). This can be converted into a bound for $\mathcal{E}(\hat{x})$ where B depends only on $h(x^*,\cdot)$ instead of the hypothesis class (Zeng and Lam (2022)), but then α typically degrades as $\frac{1}{D_\xi}$ where D_ξ denotes the dimension of the randomness ξ. In other words, in all the existing bounds for ERM and DRO, the generalization error $\mathcal{E}(\hat{x})$ depends on either the complexity of the cost function class or the dimension of the distribution space. Thus, for a high-dimensional problem with complex cost function, both ERM and DRO are likely to have poor performance.

Our main goal in this paper is to propose a simple approach that aims to remove the dependence of the bound on the generalization error $\mathcal{E}(\hat{x})$ on *both* the function complexity and distributional dimension. Our approach operates by replacing the empirical distribution $\hat{\mathbb{P}}_n$ typically used as the center of the ambiguity set in DRO with a suitable parametric distribution. For convenience we call our approach *parametric DRO* (P−DRO). Using the second analysis route of DRO mentioned above, we obtain B that depends only on $h(x^*,\cdot)$ while, because of the use of parametric distribution, we also remove the dependence of D_ξ in α. Of course, all of this come with the price of a model misspecification error due to the use of parametric distributions. The main insight is that by choosing the ambiguity set size properly, the worst-case nature of P−DRO can be leveraged to control the impact of model misspecification, and ultimately exhibit a gracious tradeoff between this latter error and the removal of complexity/dimension dependence.

Under this framework, we demonstrate how the strength of P−DRO is further amplified under distribution shift, i.e., when training and testing data statistically differ, thanks to the hedge on model misspecification provided by P−DRO. Our desirable generalization bound of P−DRO under distribution shift also serves as a propellant of DRO as truly superior against model changes – While previous literature has argued the advantages of DRO in protecting against unexpected distribution shift, the arguments are based on a worst-case bound applied on the attained objective (e.g., Van Parys et al. (2020); Sutter et al. (2021)), which does not imply whether the obtained solution is good relative to other possibilities or the oracle solution. Our generalization bound, on the other hand, reveals how P−DRO can be better than both ERM, conventional DRO, and also parametric analogs of ERM, meaning that our solution is better under the shifted test distribution than other previous approaches. On the other hand, P−DRO requires potentially more computation effort than these other methods, due to the need to suitably discretize the parametric distribution (if it is continuous) for optimization tractability. We will also analyze the price of such a discretization effort.

In the following, we first explain the existing generalization error bounds of ERM and DRO, including notably the key reasoning behind their derivations (Section 2). Then, we present P−DRO and its basic theory (Section 3). We generalize the theory to distribution shift (Section 3.1) and incorporate discretization or Monte Carlo error (Section 3.3). Finally, we present numerical experiments on both synthetic and real data to support the strengths of P−DRO (Section 4).

2 Background

We briefly discuss how existing generalization bounds for $\mathcal{E}(\hat{x})$ for ERM and DRO are derived. This would reveal the related literature and also set the stage for our new bounds for P−DRO. First, it is customary to decompose:

$$\begin{aligned}\mathcal{E}(\hat{x}) &= [Z(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(\hat{x}) - \hat{Z}(x^*)] \\ &\quad + [\hat{Z}(x^*) - Z(x^*)] \\ &\le [Z(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(x^*) - Z(x^*)]. \end{aligned} \tag{5}$$

In the above, $\hat{x}$ denotes either the ERM solution obtained from (2) or DRO solution from (3). $\hat{Z}(\cdot)$ refers to the corresponding *estimated* objective function, namely the sample average objective depicted in (2) and the worst-case objective in (3). The inequality in (5) that removes the middle term in the decomposition thus follows from the optimality of $\hat{x}$ in minimizing $\hat{Z}(\cdot)$ in either ERM or DRO. From (5), one can proceed to bound the two terms and, in a sense, optimizing the overall generalization bound relies on a balancing between the bounds on these two terms.

ERM. In ERM, the second term $\hat{Z}(x^*) - Z(x^*)$, which relies only on the oracle optimal solution x^* among the decision space $\mathcal{X}$, can be bounded easily using standard bounds for sample mean. On the other hand, the first term $Z(\hat{x}) - \hat{Z}(\hat{x})$ depends on the random solution $\hat{x}$ and is bounded by its supremum $\sup_{x\in\mathcal{X}} |Z(x) - \hat{Z}(x)|$ (or localized versions). Using tools from empirical processes (van der Vaart et al. (1996)), we obtain in general a bound for $\mathcal{E}(\hat{x})$ of the form $O(\sqrt{\frac{MZ(x^*)\text{Comp}(\mathcal{H})\log n}{n}})$ (Vapnik (1999); Boucheron et al. (2005)), where Comp$(\mathcal{H})$ is some *complexity* measure of the hypothesis class $\mathcal{H}$ =

$\{h(x;\cdot)|x \in \mathcal{X}\}$. For example, the VC Dimension. and $M = \sup_{x\in\mathcal{X}} \|h(x;\cdot)\|_\infty$.

DRO from regularization perspective. Analyzing $\mathcal{E}(\hat{x})$ for DRO can take two viewpoints. One is its equivalence to a *variability regularization* on ERM. To explain this, let us now put subscript "ERM" and "DRO" under $\hat{Z}$ to denote their respective estimated objective function. Lam (2016); Duchi and Namkoong (2019); Gao et al. (2022) show that, for small enough ball size ε, we have roughly speaking:

$$\hat{Z}_{DRO}(x) = \hat{Z}_{ERM}(x) + \mathcal{V}_d(x)\sqrt{\varepsilon} + O(\varepsilon),\ \forall x \in \mathcal{X}. \quad (6)$$

Here $\mathcal{V}_d(x)$ is a variability measure of the cost function h that depends on the statistical distance d used in the ambiguity set. For example, $\mathcal{V}_d(x)$ is the Lipschitz norm of $h(x;\cdot)$ if d is 1-Wasserstein (Blanchet et al. (2019a); Gao et al. (2022)), and $\sqrt{\text{Var}_{\mathbb{P}^*}[h(x,\xi)]}$ if d is an f-divergence (Lam (2016); Duchi and Namkoong (2019); Duchi et al. (2021)). (6) can be used to bound the second term $\hat{Z}(x^*) - Z(x^*)$ in (5) by connecting to ERM. Moreover, with ε properly chosen (depending on the hypothesis class complexity), (6) can be converted into the bound

$$Z(x) \le \hat{Z}_{DRO}(x) + O\left(\frac{1}{n}\right),\ \forall x \in \mathcal{X}$$

by using an empirical Bernstein inequality (Maurer and Pontil (2009)), which can be used to bound the first term $Z(\hat{x}) - \hat{Z}(\hat{x})$ in (5) as well. Putting these together arrives at a bound for $\mathcal{E}(\hat{x})$ given by $O(\mathcal{V}_d(x^*)\sqrt{\frac{\text{Comp}(\mathcal{H})}{n}})$. Comparing with ERM, this DRO bound bears the constant $\mathcal{V}_d(x^*)$ instead of $MZ(x^*)$, but both bounds require Comp$(\mathcal{H})$.

DRO from robust bound perspective. As another perspective to understand DRO, if the ball size ε is chosen large enough such that

$$\mathbb{P}[d(\mathbb{P}^*, \hat{\mathbb{P}}_n) \le \varepsilon] \ge 1 - \delta \quad (7)$$

i.e., $\mathcal{G}$ covers the ground-truth $\mathbb{P}^*$ with high probability $1-\delta$, then the first term $Z(\hat{x}) - \hat{Z}(\hat{x})$ in (5) is non-positive with probability at least $1-\delta$ (Ben-Tal et al. (2013); Bertsimas et al. (2018)). Note that this choice of ε does not depend on the cost function h. At the same time, the second term $\hat{Z}(x^*) - Z(x^*)$ depends on ε, but not Comp$(\mathcal{H})$. These altogether give rise to an overall bound that only depends on $\mathcal{H}$ through $h(x^*,\cdot)$ (Zeng and Lam (2022)). However, for (7) to hold, we typically need to choose ε to scale in $O(n^{-\frac{1}{D_\xi}})$, which in turn would degrade the bound for the second term. This is the case if we use Wasserstein (Esfahani and Kuhn (2018)) and f-divergence, the latter requiring a modification of $\hat{\mathbb{P}}_n$ to a smoothed distribution estimator due to absolute continuity requirement in defining the divergence (Jiang and Guan (2018)). The only exception is Maximum Mean Discrepancy (MMD) that can retain ε to be $O(1/\sqrt{n})$, but then the second term bound requires strong reproducing kernel assumption on $h(x^*,\cdot)$, i.e. $h(x^*,\cdot) \in \mathcal{H}$ in Zeng and Lam (2022). Overall, if the assumption is mild, then we would have a bound for $\mathcal{E}(\hat{x})$ given by $O(n^{-\frac{1}{D_\xi}})$.

Overview of our bound. The bounds discussed above are shown in the first column of Table 1. As we can see, they either depend on the hypothesis class complexity Comp$(\mathcal{H})$ or the distributional dimension D_ξ. Our approach `P-DRO`, which uses a suitably fit parametric model in the DRO ball center, replaces both Comp$(\mathcal{H})$ and D_ξ with a potentially much smaller *parametric complexity* Comp(Θ). However, in doing so, we incur a model misspecification term $\mathcal{E}_{apx}$. The tradeoff between Comp(Θ) and $\mathcal{E}_{apx}$ are shown in Table 1 (shown at the bottom of the second column): When sample size n is moderate, the gain in Comp(Θ) over Comp$(\mathcal{H})$ could be significant and outwash the loss from $\mathcal{E}_{apx}$. Moreover, if we simply apply the same parametric model into ERM, we obtain a bound that depends less desirably on $\mathcal{E}_{apx}$ (shown at the top of the second column).

3 Main Results

Given i.i.d. sample $\{\hat{\xi}_i\}_{i=1}^n$ and a class of parametric distributions $\mathcal{P}_\Theta$ parametrized by Θ, our `P-DRO` solves (3) where $\mathcal{A} = \{\mathbb{P}|d(\mathbb{P},\hat{\mathbb{Q}}) \le \varepsilon\}$ is now centered around the parametric distribution $\hat{\mathbb{Q}} \in \mathcal{P}_\Theta$. This $\hat{\mathbb{Q}}$ is estimated from the sample via a parametric fit. As a special case, we can apply the same $\hat{\mathbb{Q}}$ to ERM, giving rise to `P-ERM` (i.e., setting $\varepsilon = 0$ in `P-DRO`).

To analyze `P-DRO`, we first make the following general assumption:

Assumption 1 (Oracle Estimator). $\hat{\mathbb{Q}} \in \mathcal{P}_\Theta$ *satisfies*

$$d(\mathbb{P}^*, \hat{\mathbb{Q}}) \le \mathcal{E}_{apx}(\mathbb{P}^*, \mathcal{P}_\Theta) + \frac{Comp(\Theta)\log(1/\delta)}{n^\alpha} =: \Delta(\delta, \Theta),$$

with probability $1-\delta$, *for some* $\alpha > 0$, *Comp*(Θ) *is the complexity of* $\mathcal{P}_\Theta$, *and* $\mathcal{E}_{apx}(\mathbb{P}^*, \mathcal{P}_\Theta)$ *(sometimes abbreviated to* $\mathcal{E}_{apx}$*) is a non-negative function such that* $\mathcal{E}_{apx}(\mathbb{P}^*, \mathcal{P}_\Theta) = 0$ *if* $\mathbb{P}^* \in \mathcal{P}_\Theta$.

Assumption 1 holds under a wide range of parametric models, though its verification is case-by-case. Here we discuss two important settings (We leave the formal definitions of popular metrics d including Wasserstein distance and f-divergence, such as Kullback-Leibler (KL), χ^2 and Hellinger (H^2), in the Appendix A):

Example 1. d = *1-Wasserstein distance,* $\mathcal{P}_\Theta = \{\mathcal{N}(\mu,\Sigma)|\mu \in \mathbb{R}^{D_\xi}\}$ *with known* Σ, *and marginals*

Method	No Distribution Shift		Additional Error Due to Distribution Shift	
	Standard	Parametric	Standard	Parametric
ERM	$\sqrt{\frac{MZ(x^*)\text{Comp}(\mathcal{H})}{n}}$	$\mathcal{V}_d(x^*)\sqrt{\frac{\text{Comp}(\Theta)}{n}+\mathcal{E}_{apx}}+M\mathcal{E}_{apx}^{\frac{3}{4}}$	$d(\mathbb{P}^{tr},\mathbb{P}^{te})M^{\frac{3}{4}}(\frac{\text{Comp}(\mathcal{H})}{n})^{\frac{1}{4}}$	$M(d(\mathbb{P}^{tr},\mathbb{P}^{te}))^{\frac{3}{4}}$
DRO with metric d	$\mathcal{V}_d(x^*)\cdot\frac{1}{n^{1/D_\xi}}$ $\mathcal{V}_d(x^*)\sqrt{\frac{\text{Comp}(\mathcal{H})}{n}}$	$\mathcal{V}_\mathbf{d}(\mathbf{x}^*)\sqrt{\frac{\mathbf{Comp}(\mathbf{\Theta})}{\mathbf{n}}+\mathcal{E}_{\mathbf{apx}}}$	0 $d(\mathbb{P}^{tr},\mathbb{P}^{te})\mathcal{V}_d^{\frac{1}{2}}(x^*)M^{\frac{1}{2}}(\frac{\text{Comp}(\mathcal{H})}{n})^{\frac{1}{4}}$	**0**

Table 1: Generalization error of different methods w/o distribution shift. In the case with distribution shift, we show the additional term besides $d(\mathbb{P}^{tr},\mathbb{P}^{te})\mathcal{V}_d(x^*)$, which would be paid for across all methods.

of the random variable ξ is subGaussian with parameter with parameter σ, i.e. $\mathbb{E}[\exp(v^\top(\xi-\mathbb{E}[\xi]))]\le\exp\left(\frac{\|v\|^2\sigma^2}{2}\right),\forall v\in\mathbb{R}^{D_\xi}$. *Then Assumption 1 holds for* $\hat{\mathbb{Q}}=\mathcal{N}(\frac{1}{n}\sum_{i=1}^n\hat{\xi}_i,\Sigma)$, $\mathcal{E}_{apx}=W_1(\mathbb{P}^*,\mathbb{Q}^*)$ *with* $\mathbb{Q}^*=\mathcal{N}(\mathbb{E}[\xi],\Sigma)$, $\alpha=\frac{1}{2}$ *and* $Comp(\Theta)=\sqrt{D_\xi}\sigma$.

We explain the rationale behind this example. It follows by $W_1(\mathbb{P}^*,\hat{\mathbb{Q}})\le W_1(\mathbb{P}^*,\mathbb{Q}^*)+W_1(\mathbb{Q}^*,\hat{\mathbb{Q}})$, where we bound $W_1(\mathbb{Q}^*,\hat{\mathbb{Q}})\le W_2(\mathbb{Q}^*,\hat{\mathbb{Q}})$ and use the computation of W_2 for two Gaussian distributions $W_2(\mathbb{Q}^*,\hat{\mathbb{Q}})=\sqrt{\sum_{j=1}^{D_\xi}|\frac{1}{n}\sum_{i=1}^n(\hat{\xi}_i))_j-\mathbb{E}[\xi]_j|^2}$ in Dowson and Landau (1982). Then we apply subGaussian concentration inequality (Wainwright (2019)) to all D_ξ components and obtain $W_2(\mathbb{Q}^*,\hat{\mathbb{Q}})\le\sigma\sqrt{\frac{D_\xi\log(1/\delta)}{n}}$.

Example 2 (Extracted from Theorem 13 in Liang (2021))**.** *d = KL-divergence, $\mathcal{P}_\Theta$ is the class of all distributions governing $g_\theta(Z)$ for some random variable Z and function g_θ parametrized by $\theta\in\Theta$. Then Assumption 1 holds for $\hat{\mathbb{Q}}$ as the distribution of $g_{\hat{\theta}_n}(Z)$ and*

$$\mathcal{E}_{apx}=\sup_\theta\inf_\omega\left\|\log\frac{p_*}{p_\theta}-f_\omega\right\|_\infty+B\inf_\theta\left\|\log\frac{p_\theta}{p_*}\right\|_\infty^{\frac{1}{2}},$$
$$Comp(\Theta)=\sqrt{\texttt{Pdim}(\mathcal{F})},\alpha=\frac{1}{2},$$

where p_* and p_θ are the densities of $\mathbb{P}^*$ and $g_\theta(Z)$ if we consider a GAN estimator with the discriminator class $\mathcal{F}=\{f_\omega(x):\mathbb{R}^{D_\xi}\to\mathbb{R}\}$ realized by a neural network with weight parameter ω, and generator class $\mathcal{G}=\{g_\theta(z):\mathbb{R}^{D_\xi}\to\mathbb{R}^{D_\xi}\}$ realized by a neural network with weight parameter θ:

$$\hat{\theta}_n\in\arg\min_{\theta:g_\theta\in\mathcal{G}}\max_{\substack{\omega:f_\omega\in\mathcal{F},\\ \|f_\omega\|_\infty\le B}}\{\mathbb{E}_Z f_\omega(g_\theta(Z))-\hat{\mathbb{E}}_n f_\omega(X)\},$$

with $\texttt{Pdim}(\mathcal{F})$ as the Pseudo dimension of $\mathcal{F}$.

Here $\mathcal{E}_{apx}$ reflects the expressiveness of the generator and discriminator in Example 2 and Comp(Θ) describes the statistical complexity of the discriminator. Note that α is dimension-independent in both examples above, and this is also generally the case for most interesting metrics (we leave more details in the Appendix B).

Next, to state our main result, we consider two main types of distances d. First, the Integral Probability Metric (IPM) (Müller (1997)) is defined as

$$d(\mathbb{P},\mathbb{Q}):=\sup_{\{f:\mathcal{V}_d(f)\le1\}}\left|\mathbb{E}_\mathbb{P}[f]-\mathbb{E}_\mathbb{Q}[f]\right|,$$

for appropriately defined variability measure $\mathcal{V}_d(f)$. Examples include the 1-Wasserstein distance ($\mathcal{V}_d(f)=\|f\|_{\text{Lip}}$), Total Variation Distance ($\mathcal{V}_d(f)=2\|f\|_\infty$) and MMD ($\mathcal{V}_d(f)=\|f\|_\mathcal{H}$) (Zhao and Guan (2015)). Second, we consider f-divergences beyond the IPM class. Here, we can still link it to IPM, in particular the Total Variation Distance d_{TV}, if the following holds with some constant C_d:

$$d_{TV}(\mathbb{P},\mathbb{Q})\le C_d\sqrt{d(\mathbb{P},\mathbb{Q})}.\tag{8}$$

For example, KL-divergence holds with $C_{KL}=\frac{1}{\sqrt{2}}$.

With these, letting x^{P-DRO} be the solution to P−DRO, we have the following result.

Theorem 1 (Generalization bounds for P−DRO)**.** *Suppose Assumption 1 holds and the size of the ambiguity set $\varepsilon\ge\Delta(\delta,\Theta)$. Then, with probability at least $1-\delta$, the generalization error of P−DRO satisfies the following:*

(a) When d is an IPM metric,

$$\mathcal{E}(x^{P-DRO})\le2\mathcal{V}_d(x^*)\varepsilon.$$

(b) When d is a non-IPM satisfying (8), including χ^2, KL, H^2,

$$\mathcal{E}(x^{P-DRO})\le4C_d\|h(x^*;\cdot)\|_\infty\sqrt{\varepsilon}.$$

(c) When d is the χ^2-divergence, we can improve the bound to

$$2\sqrt{\varepsilon Var_{\mathbb{P}^*}[h(x^*;\xi)]}+2\varepsilon^{\frac{3}{4}}\|h(x^*;\cdot)\|_\infty.$$

Theorem 1 gives the bounds on $\mathcal{E}(x^{P-DRO})$ (excluding constant factors) with probability at least $1-\delta$ for the following examples:

(1) 1-Wasserstein distance: $\|h(x^*;\cdot)\|_{\text{Lip}}\Delta(\delta,\Theta)$

(2) KL-divergence: $\|h(x^*;\cdot)\|_\infty\sqrt{\Delta(\delta,\Theta)}$. In the Appendix D.1.2, we further show an improvement to $\sqrt{\text{Var}_{\mathbb{P}^*}[h(x^*;\cdot)]\Delta(\delta,\Theta)}$ under mild conditions.

In the above, $\Delta(\delta,\Theta)$ can be plugged in from Assumption 1 and the quantities discussed right after it.

The key ideas in proving Theorem 1 are as follows. Consider the first and second terms in the right hand side in the decomposition (5). Based on Assumption 1, our choice of ε ensures $\mathbb{P}[d(\mathbb{P}^*,\hat{\mathbb{Q}})\le\varepsilon]\ge 1-\delta$. Then under the event $d(\mathbb{P}^*,\hat{\mathbb{Q}})\le\varepsilon$, we have $\mathbb{P}^*\in\mathcal{A}$ and $\mathbb{E}_{\mathbb{P}^*}[g(\xi)]\le\sup_{\mathbb{P}\in\mathcal{A}}[g(\xi)]$ for any measurable function g. This implies that the first term $Z(\hat{x})-\hat{Z}(\hat{x})$ in (5) is non-positive with probability at least $1-\delta$. This observation holds for all three cases in Theorem 1. On the other hand, when d is IPM, the second term $\hat{Z}(x^*)-\hat{Z}(x^*)$ can be written as

$$\begin{aligned}\hat{Z}(x^*)-Z(x^*) &= \max_{d(\mathbb{P},\hat{\mathbb{Q}})\le\varepsilon}\mathbb{E}_{\mathbb{P}}[h(x^*;\cdot)]-\mathbb{E}_{\mathbb{P}^*}[h(x^*;\cdot)]\\ &\le \mathcal{V}_d(x^*)\max_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{Q}})\le\varepsilon} d(\mathbb{P},\mathbb{P}^*)\\ &\le \mathcal{V}_d(x^*)(d(\mathbb{P},\hat{\mathbb{Q}})+d(\hat{\mathbb{Q}},\mathbb{P}^*))\\ &\le 2\mathcal{V}_d(x^*)\varepsilon.\end{aligned}$$

When d is a non-IPM satisfying (8), we have

$$\hat{Z}(x^*)\le\max_{d_{TV}(\mathbb{P},\hat{\mathbb{Q}})\le C_d\sqrt{\varepsilon}}\mathbb{E}_{\mathbb{P}}[h(x^*;\cdot)],$$

which allows us to reduce to the previous case. Moreover, to obtain the improved result for the χ^2-divergence in part (c) in particular, we borrow the Cauchy-Schwarz inequality to relate to $\sqrt{\text{Var}_{\mathbb{P}^*}[h(x^*;\xi)]}$:

$$|\mathbb{E}_{\mathbb{P}}[h(x^*;\xi)]-\mathbb{E}_{\mathbb{Q}}[h(x^*;\xi)]|\le\sqrt{2\chi^2(\mathbb{P},\mathbb{Q})\text{Var}_{\mathbb{P}}[h(x^*;\xi)]}. \tag{9}$$

(9) helps us obtain the inequality $\sup_{\mathbb{P}\in\mathcal{A}}\mathbb{E}_{\mathbb{P}}[h(x^*;\xi)]-\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\le\sqrt{2\varepsilon\text{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}$ when d is the χ^2-divergence, and we only need to bound the term $|\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]-\mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]|$ by applying (9) again.

Lastly, in Appendix D.1.2, under additional mild conditions, when d is the KL divergence or Hellinger distance, we improve the result in part (b) of Theorem 1 to $\mathcal{E}(x^{P-DRO})\le c_1\sqrt{\varepsilon\text{Var}_{\mathbb{P}^*}[h(x^*;\xi)]}+c_2\varepsilon^{\frac{3}{4}}\|h(x^*;\cdot)\|_\infty$ with probability at least $1-\delta$ for some constants c_1,c_2.

Next, letting x^{P-ERM} be the solution to P−ERM, we have the following result:

Theorem 2 (Generalization bounds for P−ERM)**.** *Suppose Assumption 1 holds and the cost function $|h(x;\xi)|\le M,\forall x,\xi$. Then with probability at least $1-\delta$, the generalization error of P−ERM satisfies the following:*

(a) When d is an IPM metric:

$$\mathcal{E}(x^{P-ERM})\le 2\left(\sup_{x\in\mathcal{X}}\mathcal{V}_d(x)\right)\Delta(\delta,\Theta).$$

(b) When d is a non-IPM satisfying (8), including χ^2, KL, H^2,

$$\mathcal{E}(x^{P-ERM})\le 4C_dM\sqrt{\Delta(\delta,\Theta)}.$$

(c) When d is (modified) χ^2-divergence, we can improve the bound to:

$$\sqrt{2\Delta(\delta,\Theta)}\sqrt{Var_{\mathbb{P}^*}[h(x^*;\xi)]}+2M\left(\Delta(\delta,\Theta)\right)^{\frac{3}{4}}.$$

Compared with Theorem 1, the bound for $\mathcal{E}(x^{P-ERM})$ in Theorem 2 involves the uniform quantity $\sup_{x\in\mathcal{X}}\mathcal{V}_d(x)$ or M, which can be much larger than $\mathcal{V}_d(x^*)$ or $\|h(x^*;\cdot)\|_\infty$ for $\mathcal{E}(x^{P-DRO})$ in Theorem 1. This amplifies the model error $\mathcal{E}_{apx}$ when using P−ERM or, equivalently, demonstrates the power of P−DRO in curbing the impact of model error. Note that for χ^2-divergence, the first term in the improved bound for P−ERM is the same as that for P−DRO when $\varepsilon=\Delta(\delta,\Theta)$, but P−ERM still incurs the uniform quantity M in the second term.

We briefly outline the proof for Theorem 2. Consider the decomposition (5) and, without the worst-case machinery of DRO here, we bound the two terms by $\sup_{x\in\mathcal{X}}|Z(x)-\hat{Z}(x)|$, which leads to the appearance of $\sup_{x\in\mathcal{X}}\mathcal{V}_d(x)$. The improved χ^2 result follows by replacing the uniform bound $\sup_{x\in\mathcal{X}}\sqrt{\text{Var}_{\mathbb{P}^*}[h(x;\xi)]}$ with

$$\sqrt{\text{Var}_{\mathbb{P}^*}[h(x^{P-ERM};\xi)]}\le\sqrt{\text{Var}_{\mathbb{P}^*}[h(x^*;\xi)]}+2M(\chi^2(\mathbb{P}^*,\hat{\mathbb{Q}}))^{\frac{3}{4}}.$$

The main results of this section are summarized under "No Distribution Shift" in Table 1. As discussed in the Introduction, our P−DRO bounds do not depend on $\text{Comp}(\mathcal{H})$ and D_ξ but instead the parametric complexity $\text{Comp}(\Theta)$ and model error $\mathcal{E}_{apx}$. In general, P−DRO compares favorably with existing ERM/DRO when the hypothesis class is complex or distributional dimension is high, and when the data size is small so that the model error $\mathcal{E}_{apx}$ becomes relatively less profound compared to $\text{Comp}(\Theta)/n$. Our experimental results in Section 4 will support these stipulations.

Note that in Table 1, we suppress the dependency of α and focus on the comparison of major terms, i.e. $\text{Comp}(\Theta),\text{Comp}(\mathcal{H}),\mathcal{E}_{apx}$. From the examples satisfying Assumption 1, to bound $d(\mathbb{P}^*,\hat{\mathbb{Q}})$, $\alpha=1$ can hold when d is χ^2-divergence (Example 2) and $\alpha=\frac{1}{2}$ can hold when d is 1-Wasserstein distance (Example 1), which implies the results shown in the "Parametric" column under "No Distribution Shift" in Table 1.

3.1 Generalization to Distribution Shift

We extend our framework to the distribution shift setting. We have i.i.d. sample from the training distribution $\mathbb{P}^{tr}$ but the test distribution $\mathbb{P}^{te}\neq\mathbb{P}^{tr}$, and $\hat{x}$ is computed using the

sample from $\mathbb{P}^{tr}$. We are interested in the generalization error $\mathcal{E}^{te}(\hat{x}) = \mathbb{E}_{\mathbb{P}^{te}}[h(\hat{x};\xi)] - \min_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{te}}[h(x;\xi)]$. We have the following results:

Corollary 1 (Generalization bounds for P−DRO under distribution shift)**.** *Suppose Assumption 1 holds, and the radius* $\varepsilon \geq \Delta(\delta,\Theta) + d(\mathbb{P}^{te},\mathbb{P}^{tr})$*, then* $\mathcal{E}(x^{P-DRO})$ *is bounded by the same term in Theorem 1.*

Corollary 2 (Generalization bounds for P−ERM under distribution shift)**.** *Suppose Assumption 1 holds, then* $\mathcal{E}(x^{P-ERM})$ *is bounded by the same term in Theorem 2 except replacing* $\Delta(\delta,\Theta)$ *with* $\Delta(\delta,\Theta) + d(\mathbb{P}^{te},\mathbb{P}^{tr})$.

The proof of the two results follows a similar structure as Theorem 2. The right half of Table 1 summarizes the additional errors incurred by the various methods because of distribution shift. We see that P−ERM under distribution shift suffers from the product of $d(\mathbb{P}^{tr},\mathbb{P}^{te})$ and uniform quantities over $\mathcal{X}$ including M and $\sup_{x\in\mathcal{X}} \mathcal{V}_d(x)$.

In the literature, the evaluation metric in distribution shift is called "discrepancy metric", which can be defined as $\mathsf{disc}_L(\mathcal{H};\mathbb{P}^{te},\mathbb{P}^{tr}) = \sup_{h_1,h_2\in\mathcal{H}} |\mathbb{E}_{\mathbb{P}^{te}} L(h_1,h_2) - \mathbb{E}_{\mathbb{P}^{tr}} L(h_1,h_2)|$ for some loss function $L : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ (Mansour et al. (2009); Ben-David et al. (2010); Zhang et al. (2019)). This metric depends on the hypothesis class $\mathcal{H}$. Instead, the metric $d(\mathbb{P}^{te},\mathbb{P}^{tr})$ to measure the distribution shift in our generalization error bound does not consider the interactions between $\mathcal{H}$ and $\mathbb{P}^{tr},\mathbb{P}^{te}$. Meanwhile, we do not need to get access to samples of $\mathbb{P}^{te}$, but we need the value (or an upper bound) of $d(\mathbb{P}^{te},\mathbb{P}^{tr})$. Lee and Raginsky (2018) uses the same discrepancy measure and obtains a similar generalization error bound when $\hat{\mathbb{Q}}$ is taken to be the empirical distribution and d is the p-Wasserstein distance, which inherits the curse of dimensionality in standard Wasserstein-DRO approaches. This challenge is also shown in the numerical results in Section 4.

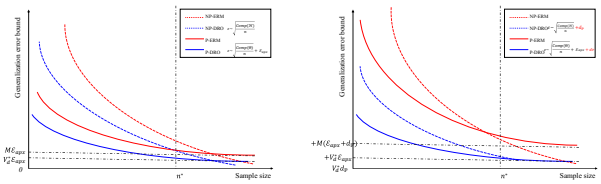
As mentioned, we do not assume any specific structure in the distribution shift but only $d(\mathbb{P}^{te},\mathbb{P}^{tr})$. This distinguishes our result from those associated with some specific types of distribution shift, such as group-based approaches (Sagawa et al. (2020)), latent covariate shifts (Duchi et al. (2020)), and conditional shifts (Sahoo et al. (2022)). It would be interesting to extend the P−DRO idea to these specific types such that the approach is more realistic to downstream tasks, e.g. representing $\mathbb{P}^{te}_{Y|x}$ by some parametric distributions and then robustifying in Sahoo et al. (2022).

3.2 Error Tradeoffs Compared to Existing Bounds

We discuss several implications of P−DRO regarding generalization. It is designed to hedge against function class complexity and distributional dimension, while paying a controllable price of model misspecification in our parametric approximation. Illustrated in Figure 1, when the sample size is not too large, i.e. when $n \leq n^*$, by replacing $\mathrm{Comp}(\mathcal{H})$ in ERM or $1/D_\xi$ in DRO with $\mathrm{Comp}(\Theta)$ plus $\mathcal{E}_{apx}$, P−DRO can enjoy better generalization. Besides, under distribution shift, both ERM and P−ERM are further negatively impacted by the amplification of the impact of function class complexity and indicate the strength of P−DRO.

We point out that our P−DRO framework hinges on the availability of a parametric model with low $\mathcal{E}_{apx}$ measured by the metric d. While this presumption may not hold in all cases, fortunately there is a rich literature in statistics for selecting and estimating parametric models: information-based model selection (Anderson and Burnham (2004)), and decision-driven parameter calibration (Ban et al. (2018)). Our approach is not to create new methods for parametric model estimation; rather *takes advantage of* this rich existing literature. More precisely, P−DRO turns the P−ERM solutions obtained from directly using these models into consistently better solutions via robustification – the error in P−ERM has an extra uniform term $M\mathcal{E}_{apx}^{\frac{3}{4}}$ in addition to the error in P−DRO when there is no distribution shift (see Table 1).

Figure 1 also demonstrates that the performance of P−DRO may be dominated by others. If $\mathrm{Comp}(\mathcal{H}) \approx \mathrm{Comp}(\Theta)$, and the parametric class $\mathcal{P}_\Theta$ provides a poor approximation for $\mathbb{P}^*$, i.e. $\mathcal{E}_{apx}$ is large, the ambiguity size ε has to be set large in order to cover the true distribution. In such setting, P−DRO underperforms against nonparametric approaches even for small n. In the limit $n \to \infty$, the error of nonparametric approaches converges to 0; however, the error of P−DRO is lower bounded by $\mathcal{E}_{apx}$.



(a) Without Shift (b) With Shift

Figure 1: Concept of model performance when $\mathrm{Comp}(\mathcal{H}) \gg \mathrm{Comp}(\Theta)$, where $d_{\mathbb{P}} = d(\mathbb{P}^{tr},\mathbb{P}^{te})$ and n^* increases with $\mathrm{Comp}(\mathcal{H})$ and decreases with $\mathrm{Comp}(\Theta), \mathcal{E}_{apx}$.

Moreover, when the distribution center $\hat{\mathbb{Q}}$ is continuous, the inner maximization term $\sup_{\mathbb{P}\in\mathcal{A}} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$ in (3) becomes an infinite-dimensional optimization problem even after dual reformulation. For example, for 1-Wasserstein distance (Esfahani and Kuhn (2018)) and f-divergence (Bayraksan and Love (2015)), the inner prob-

lem $\sup_{\mathbb{P}\in\mathcal{A}} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$ would be reformulated as:

$$\inf_{\lambda\geq 0}\{\lambda\varepsilon + \mathbb{E}_{\xi_0\sim\hat{\mathbb{Q}}}[\sup_{\xi\in\Xi}\{h(x;\xi) - \lambda\|\xi_0 - \xi\|\}]\} \quad \text{(1-W)}$$

$$\inf_{\lambda\geq 0,\mu\in\mathbb{R}}\{\mu + \lambda\varepsilon + \mathbb{E}_{\xi\sim\hat{\mathbb{Q}}}[(\lambda f)^*(h(x;\xi)-\mu)]\} \quad (f)$$

where the objective involves a high-dimensional integral over $\hat{\mathbb{Q}}$ instead of the empirical distribution. This makes the problem even harder to evaluate and optimize than DRO based on the empirical distribution. There are two approaches to handle this issue. One is through Monte Carlo and sample average approximation that reduce the problem to a structure resembling DRO based on the empirical distribution. Another is through stochastic approximation. In the next subsection we discuss the first approach and how its error can be controlled. We leave the discussion of the second approach in the Appendix D.5.

3.3 Incorporation of Monte Carlo Errors

Suppose for tractability purpose we generate Monte Carlo sample $\tilde{\xi}_i \sim \hat{\mathbb{Q}}$, $i = 1,\ldots,m$ to construct $\hat{\mathbb{Q}}_m := \frac{1}{m}\sum_{i=1}^m \delta_{\tilde{\xi}_i}$, which approximates the ball center of the ambiguity set $\mathcal{A}$ in (3), i.e., we now use $\mathcal{A} = \{\mathbb{P}|d(\mathbb{P},\hat{\mathbb{Q}}_m) \leq \varepsilon\}$. Let x^{P-DRO_m} be the corresponding solution. We investigate the required Monte Carlo size m such that $\mathcal{E}(x^{P-DRO_m}) \approx \mathcal{E}(x^{P-DRO})$ in Theorem 1.

The approximated reformulation now incurs both the statistical generalization error from the data and the Monte Carlo sampling error, and we would like the latter error to be dominated by the former. We assume $h(x;\xi) \in [0,M], \forall x,\xi$ throughout this subsection.

Theorem 3 (Generalization bounds for Wasserstein P−DRO with Monte Carlo errors)**.** *Suppose Assumption 1 holds and the size of the ambiguity set satisfies* $\frac{\varepsilon}{2} \geq \Delta(\delta,\Theta)$. *If our Monte Carlo size* $m \geq C(\frac{2}{\varepsilon})^{D_\xi}$ *for some constant* C, *when* d *is 1-Wasserstein distance, then with probability at least* $1-\delta$, *we have:*

$$\mathcal{E}(x^{P-DRO_m}) \leq 2\|h(x^*;\cdot)\|_{Lip}\varepsilon.$$

Note that when $\mathcal{E}_{apx} \approx 0$ in $\Delta(\delta,\Theta)$ and we set $\varepsilon = \Delta(\delta,\Theta)$, the required Monte Carlo size $m \approx n^{\alpha D_\xi}$, which depends on the distribution dimension. A key observation in proving Theorem 3 is that we can still establish $\mathbb{P}^*[d(\mathbb{P}^*,\hat{\mathbb{Q}}_m) \leq \varepsilon] \geq 1-\delta$ since $W_1(\mathbb{P}^*,\hat{\mathbb{Q}}_m) \leq W_1(\mathbb{P}^*,\hat{\mathbb{Q}}) + W_1(\hat{\mathbb{Q}},\hat{\mathbb{Q}}_m) \leq \varepsilon$ for large m.

However, the argument of Theorem 3 does not hold more generally since, for instance, $d(\mathbb{P}^*||\mathbb{Q}_m) = \infty$ for any m and continuous distribution $\mathbb{P}^*$ for general f-divergence d. Leveraging on the equivalence between DRO and regularization, we provide another result below:

Theorem 4 (Generalization bounds for general P−DRO with Monte Carlo errors)**.** *Suppose Assumption 1 holds and the size of the ambiguity set* $\varepsilon \geq \Delta(\delta,\Theta)$, *when* d *is* χ^2*-divergence or 1-Wasserstein distance, if the Monte Carlo size* $m \geq C\left(\frac{M}{\mathcal{V}_d(x^*)\varepsilon}\right)^k \mathrm{Comp}(\mathcal{H})\log m$ *for some constant* C *and* k, *then with probability at least* $1-\delta$, $\mathcal{E}(x^{P-DRO_m}) \leq 2\mathcal{E}_d$, *where* $\mathcal{E}_d$ *is the corresponding generalization error upper bound in Theorem 1.*

The key idea here is to control the following term $\forall x \in \mathcal{X}$:

$$\left|\sup_{d(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] - \sup_{d(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]\right|,$$

via the variability regularization property of DRO so that (3.3) can be dominated by $\mathcal{E}_d$ when $m \approx \mathrm{Comp}(\mathcal{H})n^\alpha$, which is now independent of the distribution dimension but depends on hypothesis class complexity. In Appendix D.3.5, we further show the metric d in Theorem 4 can be extended to p-Wasserstein distance for $p \in [1,2]$.

Computational Issues. Theorems 3 and 4 provide the required Monte Carlo sizes such that $\mathcal{E}(x^{P-DRO_m}) \approx \mathcal{E}(x^{P-DRO})$. With this use of Monte Carlo, P−DRO can be viewed as translating the statistical errors entailed by the distribution dimension to the model error and additional computational effort. The latter involves two aspects, one is the Monte Carlo sampling of the parametric model $\hat{\mathbb{Q}}$ in lieu of the original data, which is considered acceptable since this is typically cheap for common models. In other words, the m in x^{P-DRO_m} can be much bigger than the original data size n drawn from $\mathbb{P}^*$. Second is the optimization complexity for DRO. Since our optimization model reduces to the same formulation as DRO based on the empirical distribution of m data points after the Monte Carlo sampling, we can borrow the same existing procedures for empirical-based DRO. This is conceptually attractive, but we should caution that solving the latter is not always easy to do, and certainly more difficult than solving ERM. On the other hand, we can leverage recently proposed large-scale DRO procedures specially designed for f-divergence in Levy et al. (2020); Jin et al. (2021) and Wasserstein distance in Sinha et al. (2018) via variants of stochastic gradient method. We hope the current and future active investigation of large-scale DRO computation would make the procedure for solving P−DRO much more efficient.

Similarly, we also provide the required Monte Carlo size of P−ERM such that $\mathcal{E}(x^{P-ERM_m}) \approx \mathcal{E}(x^{P-ERM})$ in Appendix D.4, where m also depends on the hypothesis class complexity.

Other Related Work. We conclude our theoretical discussion by pointing out our differences with other work that uses parametric models in DRO. First, despite the popularity of such models in statistics and machine learning, they have not been investigated in DRO until recently. Sutter et al. (2020) establish the optimality of DRO methods

by using a Pareto-optimality argument for KL divergence, based on the criterion of out-of-sample "disappointment" instead of the generalization error. Shapiro et al. (2021) formulates and derives asymptotic results similar to variability regularization for so-called Bayesian risk optimization under parametric uncertainty. Michel et al. (2021, 2022) propose ambiguity sets that only contain parametric distributions, but without theoretical generalization guarantee under misspecified parametric distribution class. Besides, they evaluate model performances via loss robustness instead of the excess risk that we investigate.

Compared with previous work, our framework differs by: 1) focusing on the generalization error measured by the excess risk against the oracle solution; 2) providing finite-sample theoretical guarantees; 3) demonstrating our advantages both with and without distribution shift; 4) generality in accommodating most of the commonly used distance metrics including Wasserstein distance and f-divergence.

4 Numerical Studies

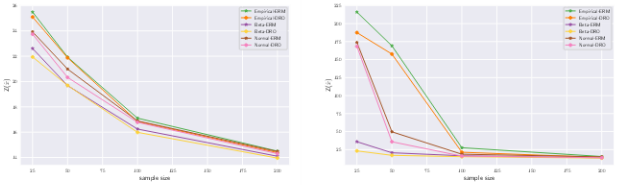
We present experiments to illustrate the properties of our models under different scenarios on both synthetic and real-world datasets. Existing ERM and DRO models in literature using the empirical distribution are denoted `NP-ERM` and `NP-DRO` and serve as our benchmarks. We tune the ambiguity size ε through cross validation in DRO methods and set Monte Carlo size $m = 50n$ for each parametric model under sample size n (unless noted otherwise). Because of page limit, detailed setups and full experimental results can be found in the Appendix E.

Synthetic Example We consider the problem of minimizing the following objective:

$$h(x;\xi) = \left|\min\{0, \xi^\top x - \mu\}\right|^\alpha = \left(\mu - \xi^\top x\right)_+^\alpha, \quad (10)$$

where ξ are asset return and μ is the target return. The vector $x \in \mathcal{X}$ represents the allocation weights, where $\mathcal{X} = \{\sum_i x_i = 1, x_i \geq -\tau\}$. This objective is called the downside risk when $\alpha = 2$ in practical portfolio optimization in Sortino et al. (2001). Here $\text{Comp}(\mathcal{H})$ grows with τ and α and is provably large. We vary $\tau \in \{2, 10\}$ and $\alpha \in \{1, 2, 4\}$. Our base case is that each marginal $(\xi)_i$ is fully parameterized, following a location variant of Beta distribution $Beta(\alpha_i, 2)$ from the domain $[0, 1]$ to $[-r, r]$ with $\alpha_i \in [\alpha_L, \alpha_U]$. We use χ^2-divergence as the DRO metric here.

Besides the empirical measure $\hat{\mathbb{P}}_n$ (Empirical-*), $\hat{\mathbb{Q}}$ in (3) is fit with the location variant of the Beta distribution class (Beta-*), where $\mathcal{E}_{apx} = 0$ in the base case. We also fit the data with the normal class (Normal-*) and find that this type of `P-DRO` approach still enjoys relatively good performance. We show the results of one base case setup in Figure 2 (a). Since $\text{Comp}(\mathcal{H})$ is much larger than $\text{Comp}(\Theta)$ here (shown in Appendix E.2), it is natural to see that parametric models perform much better than the nonparametric counterpart. Besides although `P-DRO` outperforms `P-ERM` with statistical significance ($p < 0.001$) under all chosen sample size, the absolute margins between `P-DRO` and `P-ERM` are not very obvious especially under large sample size. We show variants where distribution is misspecified (in Appendix E.2) or under distribution shift reported in Figure 2 (b). Here we find that `P-DRO` enjoys larger gains than other models, which is consistent with our theoretical argument. Other complexity setups (α, τ) in the Appendix show similar results and the gap of the performance gain of `P-DRO` compared with other methods is larger with growing (α, τ).



(a) Base Case (b) Distribution Shift

Figure 2: Average $Z(\hat{x})$ across different ERM-DRO models varying n with $(\tau, \alpha) = (2, 2)$.

Real Data I: Portfolio Optimization We now consider portfolio allocation in the real datasets and continue to use the objective (10) with $\alpha = 2$ and $\tau \in \{2, 10\}$. We use four real-world datasets from Kenneth French website with $D_\xi \in \{6, 10, 25, 30\}$. Note that the asset returns are neither stationary nor generated from some simple parametric families. Therefore, any approach would face distribution shift and model misspecification. We test our method against the benchmarks under the commonly implemented "rolling-sample" approach to evaluate the empirical out-of-sample cost $\hat{h} = \frac{1}{N}\sum_{i=1}^N (\mu - \hat{r}_i)_+^2$ of N out-of-sample returns $\{\hat{r}_i\}_{i=1}^n$. We still fit parametric models with the location variant of Beta distributions and normal distributions.

Table 2 shows that parametric models (beta, normal) perform better than directly implementing the empirical distribution, especially under large D_ξ. The performance of `NP-DRO` / `NP-ERM` is very sensitive to the choice of τ, and thus significantly dominated by parametric approaches in that regime. `P-DRO` can reduce the problem of misspecification from `P-ERM`. Besides, we find that fitting with beta parametric models can generate better decisions than the normal in this case.

Real Data II: Regression on LDW Data Finally, we consider a regression problem with a relatively large

Table 2: Comparison of performance $\hat{h}$ under different models in the portfolio allocation problem with $\tau = 2$ (the term in bracket shows the multiples of the empirical $\hat{h}$ with $\tau = 10$ for the corresponding model).$^+$ means the DRO model outperforms the ERM counterpart with statistical significance; * means P−DRO outperform empirical-DRO with statistical significance ($p < 0.001$).

	empirical		beta		normal	
dataset / method	ERM	DRO	ERM	DRO	ERM	DRO
10-Industry	36.26 (1.00)	33.35 (1.00)$^+$	31.27 (1.00)	**30.64 (1.00)**	35.4 (1.00)	31.88 (1.00)$^+$
6-FF	28.91 (1.02)	27.98 (1.01)	35.93 (1.00)	35.81 (1.00)	28.75 (1.01)	**27.93 (1.00)**
30-Industry	210.07 (9.97)	195.1 (9.58)$^+$	35.26 (1.00)	**34.33 (1.00)***	84.58 (1.03)	62.06 (1.01)$^{+*}$
25-FF	60.86 (2.90)	53.39 (2.94)$^+$	37.62 (1.00)	**36.94 (1.00)***	48.58 (1.11)	37.41 (1.04)$^{+*}$

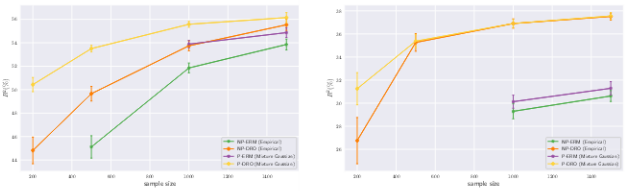
dataset. We choose the benchmark dataset psid[1], which is an observational dataset from economic surveys with $\texttt{dim}(x) = 8$ pretreated variables. The goal is to predict residents' earnings in 1978 given those features. Given n samples $\{(x_i, y_i)\}_{i=1}^n$ where the feature vector $x \in \mathbb{R}^{\texttt{dim}(x)}$ and the label $y \in \mathbb{R}$, we can express loss minimization under DRO to be:

$$\min_{h \in \mathcal{H}} \max_{d(\mathbb{P}, \hat{\mathbb{P}}) \leq \varepsilon} \mathbb{E}_{(x,y)\sim\mathbb{P}}[\ell(y; h(x))], \tag{11}$$

where $\mathcal{H}$ is the class of linear functions with quadratic interactions between features, i.e. $[1, x_1, \ldots, x_8, x_1x_2, \ldots, x_ix_j, \ldots] \in \mathbb{R}^{37}$ as our function class $\mathcal{H}$, and ℓ is the squared loss. For DRO models, we choose d as 2-Wasserstein distance with ℓ_2 norm in (11), i.e. the result of Theorem 1 in Blanchet et al. (2019b). We fit $\{(x_i, y_i)\}_{i=1}^n$ with a mixture of Gaussian distributions for joint (x, y), where each subpopulation is based on each possible combination of category variables in x.

In Figure 3 (a), we only report out-of-sample R^2[2] for different methods averaged over 50 independent runs in each dataset for each sample size. Under this setup, all models enjoy better performance as the sample size grows. However, due to high function complexity, ERM models are dominated by DRO models, especially when the sample size is not too large. And P−DRO enjoys much better and statistically significant performance than NP−DRO under smaller sample size, which is consistent with the performance trends with larger n in Figure 1.

In the case of distribution shift, we consider one type of marginal distribution shift on the feature vector in Figure 3 (b). We focus on the case where individuals in the training dataset are above 25 years old but the ones in the test datasets are below that. And we tune the ambiguity size ε from an separate validation dataset to approximate the extent of distribution shift also from the same candidate hyperparameter set before. Under such case, P−DRO is slightly better than NP−DRO but not statistically significant, but they both have significantly superior results than ERM models under distribution shift.



(a) Without Shift (b) With Age Shift

Figure 3: Comparison of average R^2 (%) under different models in LDW Datasets (ERM models without points in small sample size represents instability results, i.e. $R^2 < -1$ with too large estimated $\|\theta\|$). P−DRO is statistically significant than NP−DRO ($p < 0.001$) in all sample sizes without shift and $n = 200$ with age shift.

5 Conclusion and Future Direction

We develop a distributionally robust framework that builds on the parametric distribution, investigate its generalization error properties and demonstrate its superiority against existing benchmarks. In the future, we plan to study the following directions. First is to extend to more complex generative models and incorporate data-driven model selection to balance between nonparametric and parametric approaches under different regimes like Figure 1. Second, it is natural to consider contextual optimization with better robustification of the contextual or policy distribution estimators. We can also incorporate other social concerns such as fairness and causality to make more interpretable ambiguity sets through the lens of parametric approximation. Besides, desirable representation of unknown distributions can also apply to other areas of data-driven decision-making such as online problems and offline policy learning.

Acknowledgements

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710

[1] available at https://users.nber.org/ rdehejia/nswdata2.html

[2] Under squared loss ℓ, $Z(\hat{x}) = K(1 - R^2)$ for some $K > 0$.

and IIS-1849280. We also thank the anonymous referees and meta-reviewer for the valuable feedback.

References

D Anderson and K Burnham. Model selection and multimodel inference. *Second. NY: Springer-Verlag*, 63 (2020):10, 2004.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Gah-Yi Ban, Noureddine El Karoui, and Andrew EB Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov): 463–482, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

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.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

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

John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

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

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

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

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

Xi Chen, Qihang Lin, and Guanglin Xu. Distributionally robust optimization with confidence bands for probability density functions. *INFORMS Journal on Optimization*, 4(1):65–89, 2022.

Etienne de Klerk, Daniel Kuhn, and Krzysztof Postek. Distributionally robust optimization with polynomial densities: theory, models and algorithms. *Mathematical Programming*, 181(2):265–296, 2020.

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

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2007.

DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

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

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

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

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.

Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Ruiwei Jiang and Yongpei Guan. Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782, 2021.

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

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

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Tengyuan Liang. How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406, 2021.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 1999.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Martin Mevissen, Emanuele Ragnoli, and Jia Yuan Yu. Data-driven distributionally robust polynomial optimization. *Advances in Neural Information Processing Systems*, 26, 2013.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Modeling the second player in distributionally robust optimization. In *International Conference on Learning Representations*, 2021.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Distributionally robust models with parametric likelihood ratios. In *International Conference on Learning Representations*, 2022.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Alexander Shapiro, Enlu Zhou, and Yifan Lin. Bayesian distributionally robust optimization. *arXiv preprint arXiv:2112.08625*, 2021.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Frank A Sortino, Stephen Satchell, and Frank Sortino. *Managing downside risk in financial markets*. Butterworth-Heinemann, 2001.

Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Tobias Sutter, Andreas Krause, and Daniel Kuhn. Robust generalization despite distribution shift via minimum discriminating information. *Advances in Neural Information Processing Systems*, 34, 2021.

AW van der Vaart, A.W. van der Vaart, A. van der Vaart, and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405. URL https://books.google.com/books?id=0CenCW9qmp4C.

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

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

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

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

Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

Yibo Zeng and Henry Lam. Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2022.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. *arXiv preprint arXiv:1711.02771*, 2017.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine learning*, pages 7404–7413. PMLR, 2019.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. *Available on Optimization Online*, 2(5):1–40, 2015.

A Basic Definition

Definition 1 (Wasserstein distance)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions supported on Ξ, the p-Wasserstein distance $W_p : \Xi \times \Xi \to \mathbb{R}$ is defined by:*

$$W_p(\mathbb{P},\mathbb{Q}) = \inf_{\Pi \in \mathcal{M}(\Xi\times\Xi)} \left\{ \left(\int_{\Xi\times\Xi} \|x-y\|^p \Pi(dx,dy) \right)^{\frac{1}{p}} : \Pi_x = \mathbb{P}, \Pi_y = \mathbb{Q} \right\},$$

where Π is the joint distribution of x and y, Π_x and Π_y are the corresponding marginal distributions of Π.

We also need the standard measure concentration result of 1-Wasserstein distance in our analysis.

Lemma 1 (Measure concentration, from Theorem 2 in Fournier and Guillin (2015))**.** *Suppose $\mathbb{P}$ is a light-tailed distribution such that $A := \mathbb{E}_{\mathbb{P}}[\exp(\|\xi\|^a)] < \infty$ for some $a > 1$. Then there exists some constants c_1, c_2 only depending on a, A and D_ξ such that $\forall \delta \geq 0$, if $n \geq \frac{\log(c_1/\delta)}{c_2}$, then $W_1(\mathbb{P}, \hat{\mathbb{P}}_n) \leq (\frac{\log(c_1/\delta)}{c_2 n})^{1/\max\{D_\xi, 2)\}}$.*

Definition 2 (f-divergence)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions and $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$. For a convex function $f : [0,\infty) \to (-\infty, \infty]$ such that $f(x)$ is finite $\forall x > 0$, $f(1) = 0$, f-divergence of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:*

$$D_f(\mathbb{P},\mathbb{Q}) = \int f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right].$$

Remark 1. *In terms of specific metrics d in Assumption 1 in the main text, we define them as:*

Name	*Notation in terms of $d(\mathbb{P},\mathbb{Q})$*	*$f(t)$ in Definition 2*
Total Variation (TV) distance	$d_{TV}(\mathbb{P},\mathbb{Q})$	$\frac{\lvert t-1\rvert}{2}$
χ^2-divergence	$\chi^2(\mathbb{P},\mathbb{Q})$	$\frac{(t-1)^2}{2}$
Modified χ^2-divergence	$\chi^2(\mathbb{Q},\mathbb{P})$	$\frac{(t-1)^2}{2t}$
KL divergence	$KL(\mathbb{P},\mathbb{Q})$	$t\log t - (t-1)$
squared Hellinger distance	$H^2(\mathbb{P},\mathbb{Q})$	$(\sqrt{t}-1)^2$

The following inequality shows that (modified) χ^2-divergence, KL-divergence and squared Hellinger distance satisfy (8).

Lemma 2 (Pinsker's inequality)**.** *For distributions $\mathbb{P},\mathbb{Q}$, under our definitions of specific f-divergences above, we have:*

$$d_{TV}(\mathbb{P},\mathbb{Q}) \leq \sqrt{\frac{1}{2}KL(\mathbb{P},\mathbb{Q})} \leq \frac{\sqrt{\chi^2(\mathbb{P},\mathbb{Q})}}{2}.$$

The following result shows that (modified) χ^2-divergence can also be represented as similar forms like IPM in the main text with $\mathcal{V}_d(x) = \sqrt{\text{Var}_{\mathbb{P}^*}[h(x;\xi)]}$ when d is taken as χ^2-divergence.

Lemma 3 (Pseudo IPM property for (modified) χ^2-divergence)**.** *For distributions $\mathbb{P},\mathbb{Q}$, under our definitions of specific f-divergences above, we have:*

$$\left|\mathbb{E}_{\xi\sim\mathbb{P}}[g(\xi)] - \mathbb{E}_{\xi\sim\mathbb{Q}}[g(\xi)]\right| \leq \sqrt{2\min\{\chi^2(\mathbb{P},\mathbb{Q})Var_{\xi\sim\mathbb{P}}[g(\xi)], \chi^2(\mathbb{Q},\mathbb{P})Var_{\xi\sim\mathbb{Q}}[g(\xi)]\}}.$$

Proof. This result follows directly from the definition of χ^2-divergence and the Cauchy-Schwarz inequality. Denote $M^* = \mathbb{E}_{\mathbb{Q}}[g(\xi)]$. Then we have:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[g(\xi)] - \mathbb{E}_{\mathbb{Q}}[g(\xi)] = \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right)(g(\xi) - M^*)\right] &\leq \sqrt{\mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right)^2}\sqrt{\text{Var}_{\mathbb{Q}}[g(\xi)]} \\ &= \sqrt{2\chi^2(\mathbb{P},\mathbb{Q})\text{Var}_{\mathbb{Q}}[g(\xi)]}.\end{aligned}$$

$$\mathbb{E}_{\mathbb{P}}[g(\xi)] - \mathbb{E}_{\mathbb{Q}}[g(\xi)] = \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right)(g(\xi) - M^*)\right] \geq -\sqrt{2\chi^2(\mathbb{P},\mathbb{Q})\text{Var}_{\mathbb{Q}}[g(\xi)]}.$$

The other side follows by considering the term $\mathbb{E}_{\mathbb{Q}}[g(\xi)] - \mathbb{E}_{\mathbb{P}}[g(\xi)]$. □

Following this result, all the properties for our derived generalization error bounds hold both for χ^2-divergence and modified χ^2-divergence.

In the following proofs, if not specially noted, all C with different superscripts and subscripts are denoted as some constants independent of problem-dependent complexity terms. Besides, regarding the statement in Section 3.3 in the main text, we ignore the polynomial dependence on $\log(1/\delta)$ for the required Monte Carlo size for each case.

B Parametric Estimators under Assumption 1

Back to Assumption 1 in the main text, distribution estimation is a fundamental and longstanding topic in statistics and machine learning. Compared with nonparametric approaches, learning distributions in the parametric regime reduces to finite-dimensional estimation of some parameters $\hat{\theta}$. The classical Maximum Likelihood Estimator (MLE) and Methods of Moment can provide some finite-sample guarantee of $\|\hat{\theta} - \theta^*\|$ with and without distribution misspecification (Spokoiny (2012); Boucheron et al. (2013)) in general. Under complex parametric classes where computing likelihood is intractable, minimum distance estimators (Bernton et al. (2019); Briol et al. (2019)) and generative models such as Generative Adversarial Network (GAN) (Goodfellow et al. (2020)) are efficiently implemented in practice to represent complex distributions. These methods can provide generalization guarantees to bound the distribution distance between $\mathbb{P}^*$ and the output estimator $\hat{\mathbb{Q}}$ with some distribution complexity measures (e.g., Zhang et al. (2017); Liang (2021)).

Other Examples for Assumption 1 in the Main Text. We talk about two examples where d is the Wasserstein distance or KL-divergence. We illustrate some additional estimators $\hat{\mathbb{Q}}$, and pairing distribution metrics d, that satisfy Assumption 1.

(1) d is the squared Hellinger distance, $\mathcal{P}_\Theta$ is the class of all distributions governing $g_\theta(Z)$ for some random variable Z and function g_θ parametrized by $\theta \in \Theta$. Then Assumption 1 holds for $\hat{\mathbb{Q}}$ as the distribution of $g_{\hat{\theta}_n}(Z)$ and

$$\mathcal{E}_{apx} = \sup_\theta \inf_\omega \left\| \frac{\sqrt{p_*} - \sqrt{p_\theta}}{\sqrt{p_*} + \sqrt{p_\theta}} - f_\omega \right\|_\infty + B \inf_\theta \left\| \frac{\sqrt{p_*} - \sqrt{p_\theta}}{\sqrt{p_*} + \sqrt{p_\theta}} \right\|_\infty,$$
$$\text{Comp}(\Theta) = \sqrt{\texttt{Pdim}(\mathcal{F})}, \alpha = \frac{1}{2},$$

where p_* and p_θ are the density of $\mathbb{P}^*$ and $g_\theta(Z)$ if we consider GANs estimator with the discriminator class $\mathcal{F}$ and generator class $\mathcal{G}$:

$$\hat{\theta}_n \in \arg\min_{\theta: g_\theta \in \mathcal{G}} \max_{\substack{\omega: f_\omega \in \mathcal{F}, \\ \|f_\omega\|_\infty \le B}} \{\mathbb{E}_Z f_\omega(g_\theta(Z)) - \hat{\mathbb{E}}_n f_\omega(X)\},$$

and $\texttt{Pdim}(\mathcal{F})$ is the Pseudo dimension of $\mathcal{F}$, which is shown in Liang (2021).

(2) d is χ^2-divergence and $\mathbb{P}^* \in \mathcal{P}_\Theta$ itself is a location variant of Beta distribution. See Proposition 1 for the specific result of $\text{Comp}(\Theta)$ and α.

(3) d is 1-Wasserstein distance. We consider the following two different mixture models:

- First, we consider the special case of standard mixture Gaussian models $\mathbb{P}^* \in \mathcal{P}_\Theta = \{\frac{1}{2}\mathcal{N}(\mu, \Sigma) + \frac{1}{2}\mathcal{N}(-\mu, \Sigma) | \mu \in \mathbb{R}^{D_\xi}\}$ with known $\Sigma := \sigma I_{D_\xi \times D_\xi}$.
  Then Assumption 1 holds for $\hat{\mathbb{Q}} := \frac{d}{=} \frac{1}{2}\mathcal{N}(\hat{\mu}, \Sigma) + \frac{1}{2}\mathcal{N}(-\hat{\mu}, \Sigma)$ with the output of EM algorithm with $\hat{\mu}$. Besides, $\text{Comp}(\Theta) = \sqrt{D_\xi}\sigma$ and $\alpha = \frac{1}{2}$, which is implied by $W_1(\mathbb{P}^*, \hat{\mathbb{Q}}) \le \|\hat{\mu} - \mu^*\|_2 = O(\sigma\sqrt{\frac{D_\xi \log(1/\delta)}{n}})$ by $\|\hat{\mu} - \hat{\mu}\|_2^2 = O(\frac{\sigma^2 D_\xi \log(1/\delta)}{n})$ in Theorem 6 of Xu and Zeevi (2020) and Corollary 2 of Balakrishnan et al. (2017) under some additional mild conditions.
- Second, consider $\mathbb{P}^* := \sum_{k=1}^K p_k^* \mathbb{P}_k^*$ for some unknown probability p_k with distribution $\mathbb{P}_k^*$ in terms of each group. We define $\mathcal{P}_\Theta = \{\sum_{k=1}^K p_k \mathcal{N}(\mu_k, \Sigma) | (p_1, \ldots, p_K) \in \Delta_K, \mu_k \in \mathbb{R}^{D_\xi}, \forall k \in [K]\}$[3] for some known Σ. Besides, we are given group labels $\{g_i\}_{i=1}^n$ associated with $\{\xi_i\}_{i=1}^n$, where each $g_i \in [K]$.
  Then Assumption 1 holds for $\hat{\mathbb{Q}} \overset{d}{=} \sum_{k\in[K]} \hat{p}_k \mathcal{N}(\hat{\mu}_k, \Sigma)$, where $\hat{p}_k = \frac{\sum_{i=1}^n \mathbb{I}_{\{g_i = k\}}}{n}, \hat{\mu}_k = \frac{\sum_{i=1}^n \xi_i \mathbb{I}_{\{g_i = k\}}}{n\hat{p}_k}, \forall k \in [K]$. $\mathcal{E}_{apx} = W_1(\mathbb{P}^*, \mathbb{Q}^*)$ with $\mathbb{Q}^* := \frac{d}{=} \sum_{k\in[K]} p_k^* \mathcal{N}(\mathbb{E}_{\xi\sim\mathbb{P}_k^*}[\xi], \Sigma)$, $\alpha = \frac{1}{2}$, $\text{Comp}(\Theta) = C\sqrt{D_\xi}\sigma K$ with some constant C depending on $\mathbb{P}^*$ (e.g. scales with $\frac{1}{\min_{k\in[K]} p_k^*}$ and $\max_{i,j\in[K]} \|\mathbb{E}_{\xi\sim\mathbb{P}_i^*}[\xi] - \mathbb{E}_{\xi\sim\mathbb{P}_j^*}[\xi]\|$).

[3] Δ_K represents K-dimensional probability simplex. Here $\mathcal{P}_\Theta$ corresponds to the model in our last numerical example.

Remark 2. *The derivation of the term in the second case above follows by:*

$$W_1(\mathbb{P}^*, \hat{\mathbb{Q}}) \le W_1(\mathbb{P}^*, \mathbb{Q}^*) + W_1(\mathbb{Q}^*, \tilde{\mathbb{Q}}) + W_1(\tilde{\mathbb{Q}}, \hat{\mathbb{Q}}),$$

where $\tilde{\mathbb{Q}} :\overset{d}{=} \sum_{k\in[K]} \hat{p}_k \mathcal{N}(\mathbb{E}_{\xi\sim\mathbb{P}_k^*}[\xi], \Sigma)$.

As a complement, we also mention some smooth nonparametric estimators $\hat{\mathbb{Q}}$ instead of the empirical distribution $\hat{\mathbb{P}}_n$ used in the DRO literature if the density of $\mathbb{P}^*$ is smooth. A series of suggestions include the histogram density estimate (Mevissen et al. (2013); de Klerk et al. (2020)) and kernel density estimate (KDE) (Jiang and Guan (2018); Zhao and Guan (2015); Chen et al. (2022)). For example, with the Wasserstein distance, the ambiguity size $\varepsilon = O((nh_n^{\dim(\xi)})^{-\frac{1}{2}} \vee h_n^2)$ (h_n is the bandwidth of KDE, where the density of $\hat{\mathbb{Q}}$ is $f(\xi) = \frac{1}{nh_n}\sum_{i=1}^n K(\frac{\xi-\xi_i}{h_n})$ for some kernel function $K(\cdot)$) can include $\mathbb{P}^*$ in the ambiguity set $\mathcal{A}$, where this size ε can be slightly smaller than directly implementing the empirical distribution in terms of n. However, the nature of nonparametric approaches determines that they cannot bypass the curse of dimensionality.

C Formal Derivation of Existing Empirical Approaches

In the following, we detail some steps on how existing approaches are derived specifically in Section 2 in the main text. In each approach, we look at the generalization error $Z(\hat{x}) - Z(x^*)$ where $\hat{x}$ is a data-driven minimizer. In particular, denoting $\hat{Z}(\cdot)$ as the data-driven objective, we use the decomposition roadmap:

$$Z(\hat{x}) - Z(x^*) = [Z(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(\hat{x}) - \hat{Z}(x^*)] + [\hat{Z}(x^*) - Z(x^*)]$$

where the middle term $[\hat{Z}(\hat{x}) - \hat{Z}(x^*)]$ is at most 0 by definition of $\hat{x}$ as a minimizer of $\hat{Z}(\cdot)$. Thus, we would focus on bounding w.h.p.

$$[Z(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(x^*) - Z(x^*)]. \tag{C.1}$$

The traditional approach is to replace the true distribution in the problem with the empirical distribution, i.e. $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ and obtain the empirical optimization objective of (2) with solution x^{N-ERM}.

Lemma 4 (Adapted from Boucheron et al. (2005))**.** *Consider x^{N-ERM} as the minimizer of $\min_x \hat{Z}(x)$ in (2), denote $M := \sup_{x\in\mathcal{X}} \|h(x;\cdot)\|_\infty$, then we have the following generalization error of x^{N-ERM} with probability at least $1-\delta$:*

$$Z(x^{N-ERM}) - Z(x^*) \le \log(1/\delta)\left[\sqrt{\frac{MZ(x^*)Comp(\mathcal{H})\log n}{n}} + \frac{Comp(\mathcal{H})M}{n}\right]. \tag{C.2}$$

This result is minimax optimal w.r.t. the function complexity Comp($\mathcal{H}$), e.g. the case of VC($\mathcal{H}$) shown in Section 5 of Boucheron et al. (2005).

As discussed in the main text, DRO follows another route of setup that relies on optimizing the worst-case objective over an ambiguity set $\mathcal{A} = \{\mathbb{P} : d(\mathbb{P}, \hat{\mathbb{P}}_n) \le \varepsilon\}$ constructed from some distance / divergence metric $d(\cdot, \hat{\mathbb{P}}_n)$ (f-divergence, Wasserstein distance, MMD etc). Denote the optimal solution to (3) as x^{N-DRO}. To evaluate the quality of a solution, by letting $\hat{x} := x^{N-DRO}$ and $Z^{N-DRO}(\cdot) := \hat{Z}(\cdot)$ in (C.1), the second term will lead to an error $O(\frac{\varepsilon\mathcal{V}_d(x^*)}{\sqrt{n}})$. The key lies in the first term, i.e. $Z(x^{N-DRO}) - Z^{N-DRO}(x^{N-DRO})$. In the literature, there are two interpretations to bound the first term, one using the equivalence of DRO with regularization, and one using the confidence guarantee in bounding $Z(Z^{N-DRO})$ with the worst-case objective if the ambiguity set is chosen large enough to be a confidence region for P^*. For convenience, in the sequel we sometimes call the first interpretation as the regularization perspective and the second as the pessimism perspective.

Theorem 5. *For some certain metric, when the ambiguity size $\varepsilon = \Omega((\frac{Comp(\mathcal{H})}{n})^{\frac{1}{2\beta}})$ with $\beta \in \{\frac{1}{2}, 1\}$, we have the following generalization error of x^{N-DRO} with probability at least $1-\delta$:*

$$\begin{aligned} Z(x^{N-DRO}) - Z(x^*) \le \log(1/\delta)\Bigg[&\varepsilon^\beta \mathcal{V}_d(x^*) + \frac{Comp(\mathcal{H})(\sup_{x\in\mathcal{X}} \mathcal{V}_d(x))}{n} \\ &+ \frac{\mathcal{E}_1(x^*)}{\sqrt{n}} + \mathcal{E}_2(x^*, \varepsilon)\Bigg], \end{aligned} \tag{C.3}$$

where β is a constant depending on different metrics d used here. $\mathcal{E}_1(x^*)$ only depends on $h(x^*;\xi)$ and $\mathbb{P}^*$, $\mathcal{E}_2(x^*,\varepsilon) \approx \mathcal{V}_d(x^*)\varepsilon^{2\beta}$, which is of order $\frac{1}{n}$.

When the ambiguity size $\varepsilon = \Omega(n^{-\frac{1}{\alpha(D_\xi)}})$, we may also have:

$$Z(x^{DRO}) - Z(x^*) \leq \frac{\mathcal{V}_d(x^*)\log(1/\delta)}{n^{\frac{1}{\alpha(D_\xi)}}}, \tag{C.4}$$

where $\alpha(D_\xi)$ is some function of the domain dimension of the distribution $\mathbb{P}$.

Proof. For the bound (C.3), it is a combination of the following results:

- Variability regularization in the form:

$$Z(x) \leq Z^{N-DRO}(x) + \frac{\text{Comp}(\mathcal{H})\sup_{x\in\mathcal{X}}\mathcal{V}_d(x)}{n} \tag{C.5}$$

- DRO expansions in the form:

$$Z^{N-DRO}(x^*) \leq \hat{Z}_n(x^*) + \varepsilon^\beta \mathcal{V}_d(x^*) + \mathcal{E}_2(x^*,\varepsilon).$$

- Standard concentration bound for the empirical mean:

$$|\hat{Z}_n(x^*) - Z(x^*)| \leq \frac{\mathcal{E}_1(x^*)}{\sqrt{n}}.$$

For the bound (C.4), it is achieved by using a ball size ε large enough to cover the true $\mathbb{P}^*$ with probability at least $1-\delta$. Under such event, we have:

$$Z(x^{N-DRO}) - Z^{N-DRO}(x^{N-DRO}) \leq 0 \tag{C.6}$$

$$Z^{N-DRO}(x^*) - Z(x^*) \leq \mathcal{V}_d(x^*)\varepsilon. \tag{C.7}$$

Typically, to ensure that the ball size is large enough to cover the true distribution with probability at least $1-\delta$, the ambiguity size is usually needed to depend on D_ξ. □

This result unifies several streams from previous literature. Besides we denote $r_n^* \leq \frac{\text{VC}(\mathcal{H})\log(\frac{n}{\text{VC}(\mathcal{H})})}{n}$ as the fixed point of some sub-root Rademacher Complexity. For example:

Example 3 (Gao (2022)). *In the case of 1-Wasserstein distance, $\beta = 1$, $\varepsilon = \sqrt{\frac{\tau\log(N(\mathcal{H},\frac{1}{n},n)/\delta)}{n}}$ (or $\sqrt{\frac{\tau\log(1/\delta)}{n}} + \sqrt{r_n^*} + \frac{1}{n\sqrt{r_n^*}}$) to ensure that the variability regularization bound (C.5) holds with probability at least $1-\delta$, where τ only depends on $\mathbb{P}^*$ with $\mathcal{V}_{W_1}(x) = \|h(x;\cdot)\|_{Lip}$.*

Example 4 (Duchi and Namkoong (2019)). *In the case of χ^2-divergence, $\beta = \frac{1}{2}$, $\varepsilon = \frac{\log(N(\mathcal{H},\frac{1}{n},n)/\delta)}{n}$ (or $\frac{M\log(1/\delta)}{n} + r_n^*$). Moreover, the second term in the RHS of (C.5) is also of order r_n^* with $\mathcal{V}_{\chi^2}(x) = \sqrt{Var_{\mathbb{P}^*}[h(x;\cdot)]}$.*

D Proofs and Explanations

D.1 Proof before Section 3.1

D.1.1 Proof of Theorem 1.

In the case (a) when d is an IPM metric, we have:

$$\begin{aligned} Z(x^{P-DRO}) - Z(x^*) &\overset{(a)}{\leq} \sup_{d(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] \\ &\overset{(b)}{\leq} \sup_{d(\mathbb{P},\mathbb{P}^*)\leq 2\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] \\ &\overset{(c)}{\leq} 2\mathcal{V}_d(x^*)\varepsilon, \end{aligned} \tag{D.1}$$

where (a) follows from the fact that when $\varepsilon \geq \Delta(\delta, \Theta)$, by Assumption 1, we have $\mathbb{P}^* \in \mathcal{A}(\hat{\mathbb{Q}}; d, \varepsilon)$ (i.e. $d(\mathbb{P}^*, \hat{\mathbb{Q}}) \leq \varepsilon$ with probability at least $1-\delta$). Therefore, the term $Z(x) - \hat{Z}(x)$ in (C.1) is non-positive with probability at least $1-\delta$. Furthermore, (b) follows from the triangular property of distance, $\forall \mathbb{P} \in \mathcal{A}(\hat{\mathbb{Q}}; d, \varepsilon)$, $d(\mathbb{P}, \mathbb{P}^*) \leq d(\mathbb{P}, \hat{\mathbb{Q}}) + d(\hat{\mathbb{Q}}, \mathbb{P}^*) \leq 2\varepsilon$. And (c) follows from the fact that d is IPM with $d(\mathbb{P}, \mathbb{Q}) = \sup_{f: V_d(f) \leq 1} \left| \mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f] \right|$.

In the case (b) when d satisfies the inequality (8), we have:

$$\begin{aligned}
\mathcal{E}(x^{P-DRO}) &\leq \sup_{d(\mathbb{P}, \hat{\mathbb{Q}}) \leq \varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*; \xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*; \xi)] \\
&\leq \sup_{d_{TV}(\mathbb{P}, \hat{\mathbb{Q}}) \leq C_d \sqrt{\varepsilon}} \mathbb{E}_{\mathbb{P}}[h(x^*; \xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*; \xi)] \\
&\leq 4 C_d \sqrt{\varepsilon} \|h(x^*; \cdot)\|_\infty,
\end{aligned}$$

where the first inequality follows by $\mathbb{P}^* \in \mathcal{A}(\hat{\mathbb{Q}}; d, \varepsilon)$ with probability at least $1-\delta$. The second inequality follows by the fact that $d_{TV}(\mathbb{P}, \hat{\mathbb{Q}}) \leq C_d \sqrt{d(\mathbb{P}, \hat{\mathbb{Q}})}$ such that $\{\mathbb{P} : d(\mathbb{P}, \hat{\mathbb{Q}}) \leq \varepsilon\} \subseteq \{\mathbb{P} : d_{TV}(\mathbb{P}, \hat{\mathbb{Q}}) \leq C_d \sqrt{\varepsilon}\}$. The remaining parts follow the same in (D.1) above since TV distance is IPM.

Specially, in the case (c), when d is (modified) χ^2-divergence, we have:

$$\begin{aligned}
\mathcal{E}(x^{P-DRO}) &\overset{(a)}{\leq} \sup_{\chi^2(\mathbb{P}, \hat{\mathbb{Q}}) \leq \varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*; \xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*; \xi)] \\
&\overset{(b)}{\leq} \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*; \xi)] + \sqrt{2\varepsilon \mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*; \xi)]} - \mathbb{E}_{\mathbb{P}^*}[h(x^*; \xi)] \\
&\overset{(c)}{\leq} 2\sqrt{\varepsilon \mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*; \xi)]} \\
&\overset{(d)}{\leq} 2\sqrt{\varepsilon \mathrm{Var}_{\mathbb{P}^*}[h(x^*; \xi)]} + 2^{\frac{5}{4}} \varepsilon^{\frac{3}{4}} \left[(\mathrm{Var}_{\mathbb{P}^*}[h^2(x^*; \xi)])^{\frac{1}{4}} + 2^{\frac{1}{4}} \|h(x^*; \cdot)\|_\infty^{\frac{1}{2}} (\mathrm{Var}_{\mathbb{P}^*}[h(x^*; \xi)])^{\frac{1}{4}} \right] \\
&\leq 2\sqrt{\varepsilon \mathrm{Var}_{\mathbb{P}^*}[h(x^*; \xi)]} + 4\varepsilon^{\frac{3}{4}} \|h(x^*; \cdot)\|_\infty,
\end{aligned}$$

where (a) still follows from the fact that χ^2-divergence satisfies Assumption 1 w.h.p. $(b), (c)$ follows from Lemma 3 for two pairs $(\mathbb{P}, \hat{\mathbb{Q}})$ and $(\mathbb{P}^*, \hat{\mathbb{Q}})$. And (d) follows by:

$$\begin{aligned}
\mathrm{Var}_{\hat{\mathbb{Q}}}[h] - \mathrm{Var}_{\mathbb{P}^*}[h] &\leq \left| \mathbb{E}_{\hat{\mathbb{Q}}}[h^2] - \mathbb{E}_{\mathbb{P}^*}[h^2] \right| + 2\|h\|_\infty \left| \mathbb{E}_{\hat{\mathbb{Q}}}[h] - \mathbb{E}_{\mathbb{P}^*}[h] \right| \\
&\leq \sqrt{2\chi^2(\mathbb{P}^*, \hat{\mathbb{Q}}) \mathrm{Var}_{\mathbb{P}^*}[h^2]} + 2\|h\|_\infty \sqrt{2\chi^2(\mathbb{P}^*, \hat{\mathbb{Q}}) \mathrm{Var}_{\mathbb{P}^*}[h]}. \quad \square
\end{aligned} \tag{D.2}$$

D.1.2 Proof of Improved Results of Theorem 1 for General f-divergence

In fact, the result in Theorem 1 can be improved for general f-divergence from $\|h(x^*; \cdot)\|_\infty$ to $\sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*; \cdot)]}$, without requiring (8) as long as some mild conditions hold for the cost function $h(x; \cdot)$ and sample size n. We present some technique non-asymptotical results for general f-divergence to bound $\hat{Z}(x^*) - Z(x^*)$ with the help of duality in DRO under f-divergence. These types of results hold for any f-divergence DRO problem with continuous or discrete ball center $\hat{\mathbb{Q}}$. Before that, we first present some known asymptotical results.

Remark 3 (Adapted from Theorem 1 in Duchi et al. (2021)). *For general f-divergence, we can obtain the following equality asymptotically under some mild conditions for h and ε:*

$$\sup_{\mathbb{P} \in \mathcal{A}(\hat{\mathbb{Q}}; d_f, \varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x; \xi)] = \mathbb{E}_{\hat{\mathbb{Q}}}[h(x; \xi)] + \sqrt{\frac{2\varepsilon}{f''(1)} Var_{\hat{\mathbb{Q}}}[h(x; \xi)]} + O_p\left(\frac{1}{n}\right).$$

$$\inf_{\mathbb{P} \in \mathcal{A}(\hat{\mathbb{Q}}; d_f, \varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x; \xi)] = \mathbb{E}_{\hat{\mathbb{Q}}}[h(x; \xi)] - \sqrt{\frac{2\varepsilon}{f''(1)} Var_{\hat{\mathbb{Q}}}[h(x; \xi)]} + O_p\left(\frac{1}{n}\right).$$

Theorem 6. *When the sample size n is large enough and $\varepsilon \to 0$ when $n \to \infty$, for general metric d_f in f-divergence and $\|h(x^*;\cdot)\|_\infty < \infty$, we have:*

$$\sup_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};d_f,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] \le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] + C(f)\sqrt{Var_{\hat{\mathbb{Q}}}[h(x;\xi)]\varepsilon}, \tag{D.3}$$

$$\inf_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};d_f,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] \ge \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - C(f)\sqrt{Var_{\hat{\mathbb{Q}}}[h(x;\xi)]\varepsilon}, \tag{D.4}$$

where $C(f)$ only depends on the metric d_f and $\mathbb{P}^*$.

Then the result in the case (b) in Theorem 1 can be improved to:

$$\begin{aligned}
\mathcal{E}(x^{P-DRO}) &\le \sup_{d_f(\mathbb{P},\hat{\mathbb{Q}})\le\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] \\
&\le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] + C(f)\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon} \\
&\le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \inf_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};d_f,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] + C(f)\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon} \\
&\le 2C(f)\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon},
\end{aligned}$$

where the second and fourth inequality follows by the result in Theorem 6. And the first and third inequality is a result of $\mathbb{P}[\mathbb{P}^* \in \mathcal{A}(\hat{\mathbb{Q}};d_f,\varepsilon)] \ge 1-\delta$ such that $\inf_{d_f(\mathbb{P},\hat{\mathbb{Q}})\le\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] \le \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] \le \sup_{d_f(\mathbb{P},\hat{\mathbb{Q}})\le\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)]$ with probability at least $1-\delta$. After that, we can use the same argument before to bound $\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}$.

Proof of Theorem 6. We first show (D.3). We have:

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};d_f,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] &\le \min_{\lambda\ge 0,\mu} \lambda\mathbb{E}_{\hat{\mathbb{Q}}}\left[f^*\left(\frac{h(x^*;\xi)-\mu}{\lambda}\right)\right] + \lambda\varepsilon + \mu \\
&\le \hat{\lambda}\mathbb{E}_{\hat{\mathbb{Q}}}\left[f^*\left(\frac{h(x;\xi)-\hat{\mu}}{\hat{\lambda}}\right)\right] + \sqrt{\frac{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon}{f''(1)}} + \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] \\
&\le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] + \left(\frac{1}{\sqrt{f''(1)}} + \frac{\sqrt{f''(1)}(f^*)''(0)\tilde{C}(f,\mathbb{P}^*)}{2}\right)\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon},
\end{aligned}$$

where the first inequality above is based on the weak duality condition, i.e. Theorem 1 in Ben-Tal et al. (2013)[4], and the second inequality above is given by $\hat{\lambda} = \sqrt{\frac{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}{f''(1)\varepsilon}}$, $\hat{\mu} = \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]$ as the feasible dual solution, and the third inequality plugging in the value of $\hat{\lambda}$ and $\hat{\mu}$, then we take the Taylor expansion up to the second order for f^* with a proxy of Maclaurin remainder $\tilde{C}(f,\mathbb{P}^*) \to 1$ when $\varepsilon \to 0$:

$$\begin{aligned}
\hat{\lambda}\mathbb{E}_{\hat{\mathbb{Q}}}\left[f^*\left(\frac{h(x^*;\xi)-\hat{\mu}}{\hat{\lambda}}\right)\right] &\le \hat{\lambda}\mathbb{E}_{\hat{\mathbb{Q}}}\left[f^*(0) + (f^*)'(0)\left(\frac{h(x^*;\xi)-\hat{\mu}}{\hat{\lambda}}\right) + \frac{(f^*)''(0)\tilde{C}(f,\mathbb{P}^*)}{2}\left(\frac{h(x^*;\xi)-\hat{\mu}}{\lambda}\right)^2\right], \\
&= \hat{\lambda}\mathbb{E}_{\hat{\mathbb{Q}}}\left[\frac{(f^*)''(0)\tilde{C}(f,\mathbb{P}^*)}{2}\left(\frac{h(x^*;\xi)-\hat{\mu}}{\hat{\lambda}}\right)^2\right] = \frac{(f^*)''(0)\tilde{C}(f,\mathbb{P}^*)\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}{2\hat{\lambda}},
\end{aligned}$$

where the equality holds by $f^*(0) = 1$ and $\hat{\mu} = \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]$.

Then if ε is small, we let $C(f,\mathbb{P}^*) \ge \frac{1}{\sqrt{f''(1)}} + \frac{\sqrt{f''(1)}(f^*)''(0)\tilde{C}(f,\mathbb{P}^*)}{2}$ and obtain (D.3). For (D.4), we only need to consider $-h(x;\cdot)$ and plug in the result of (D.3). □

We now show several common divergences satisfying (8) and give some concrete values for the result above. We focus on the counterpart w.r.t. (D.3).

[4] Although strong duality holds generally in this problem, we only need weak duality in our proof.

Example 5 (KL divergence). *We take $f(t) = t\log t-(t-1)$. Then $f^*(t) = e^t-1$ with $f''(1) = 1$. We use Taylor inequality $e^t - 1 \le t + t^2$ when $t \in (-1, 1)$, i.e. we need $\left|\frac{h(x^*;\xi)-\hat{\mu}}{\hat{\lambda}}\right| \le 1$, which implies when $\sqrt{\frac{Var_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}{\varepsilon}} = \hat{\lambda} \ge 2\|h(x^*;\cdot)\|_\infty$, i.e. $\varepsilon \le \frac{Var_{\hat{\mathbb{P}}}[h(x^*;\xi)]}{4\|h(x^*;\cdot)\|_\infty^2}$, then we have:*

$$\sup_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};KL,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] \le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] + 3\sqrt{Var_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon}.$$

Example 6 (Hellinger distance). *Similarly, we apply $f(t) = (\sqrt{t}-1)^2$ and $f''(1) = \frac{1}{2}$. Then for $t < 1$, $f^*(t) = \frac{t}{1-t} = \frac{1}{1-t} - 1 \le t + 2t^2$ when $t \in [-\frac{1}{2}, \frac{1}{2}]$. Thus if $\varepsilon \le \frac{Var_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}{2\|h(x^*;\cdot)\|_\infty^2}$, we have:*

$$\sup_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};H^2,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] \le \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] + (2+\sqrt{2})\sqrt{Var_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\varepsilon}.$$

Therefore, following the same argument in the case (c) of Theorem 1, if $\Delta(\delta,\Theta) \le \varepsilon \le \frac{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]}{c_0 2\|h(x^*;\cdot)\|_\infty^2}$, $\mathcal{E}(x^{P-DRO})$ can be improved to $c_1\sqrt{\varepsilon \mathrm{Var}_{\mathbb{P}^*}[h(x^*;\xi)]} + c_2\varepsilon^{\frac{3}{4}}\|h(x^*;\cdot)\|_\infty$ for KL divergence and Hellinger distance with probability at least $1-\delta$.

D.1.3 Proof of Theorem 2

In the case (a) when d is an IPM metric, we have:

$$\begin{aligned}
\mathcal{E}(x^{P-ERM}) &\le |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^{P-ERM};\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^{P-ERM};\xi)]| + |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]| \\
&\overset{(a)}{\le} 2\sup_{x\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}^*}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]\right| \\
&\overset{(b)}{\le} 2\sup_{x\in\mathcal{X}} \mathcal{V}_d(x) d(\mathbb{P}^*, \hat{\mathbb{Q}}).
\end{aligned}$$

where (a) follows from the uniform bound $\forall x \in \mathcal{X}$, (b) follows from the fact that d is IPM such that $d(\mathbb{P},\mathbb{Q}) = \sup_{\mathcal{V}_d(f)\le 1}|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]|$. □

In the case (b) where d satisfies the inequality (8), similarly, we have:

$$\begin{aligned}
\mathcal{E}(x^{P-ERM}) &\le |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^{P-ERM};\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^{P-ERM};\xi)]| + |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]| \\
&\le 2\sup_{x\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}^*}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]\right| \\
&\le 4Md_{TV}(\mathbb{P}^*, \hat{\mathbb{Q}}) \le 4C_d M\sqrt{d(\mathbb{P}^*, \hat{\mathbb{Q}})}.
\end{aligned}$$

Specially, in the case (c), when d is (modified) χ^2-divergence, following the previous decomposition and Lemma 3, we have:

$$\begin{aligned}
\mathcal{E}(x^{P-ERM}) &\le |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^{P-ERM};\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^{P-ERM};\xi)]| + |\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]| \\
&\le \sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\left(\sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^{P-ERM};\xi)]} + \sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\xi)]}\right) \\
&\le 2\sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\xi)]} + \sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\sqrt{|\mathbb{E}_{\mathbb{P}^*}[h^2(x^{P-ERM};\xi)] - \mathbb{E}_{\mathbb{P}^*}[h^2(x^*;\xi)]|} \\
&\overset{(a)}{\le} 2\sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\xi)]} + \sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\sqrt{4M^2 d_{TV}(\hat{\mathbb{Q}},\mathbb{P}^*)} \\
&\overset{(b)}{\le} 2\sqrt{2\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*)}\sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\xi)]} + 2M(\chi^2(\hat{\mathbb{Q}},\mathbb{P}^*))^{\frac{3}{4}},
\end{aligned}$$

where the inequality (a) follows by the following argument:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^*}[h^2(x^{P-ERM};\xi)] - \mathbb{E}_{\mathbb{P}^*}[h^2(x^*;\xi)] &\le \mathbb{E}_{\mathbb{P}^*}\left[(h(x^{P-ERM};\xi) + h(x^*;\xi))((h(x^{P-ERM};\xi) - h(x^*;\xi))\right] \\
&\le 2M\mathbb{E}_{\mathbb{P}^*}[h(x^{P-ERM};\xi) - h(x^*;\xi)] \le 4M^2 d_{TV}(\hat{\mathbb{Q}},\mathbb{P}^*),
\end{aligned}$$

where the last inequality is from the definition of TV distance. And the inequality (b) follows from Lemma (2).

For each part above, we then use Assumption 1 to obtain the result. □

D.2 Proof and Detailed Analysis in Section 3.1

Here, we illustrate the main parts in the generalization error in the right part of Table 1. Besides the notations in the main text, we denote $x^{tr} \in \arg\min_{x\in\mathcal{X}} \mathbb{E}^{tr}[h(x;\xi)]$.

We first illustrate some discussion of the generalization error of existing approaches for ERM and DRO under the empirical distribution $\hat{\mathbb{P}}_n$. Again we focus on the generalization error

$$Z^{te}(\hat{x}) - Z^{te}(x^*)$$

and will use the decomposition

$$Z^{te}(\hat{x}) - Z^{te}(x^*) = [Z^{te}(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(\hat{x}) - \hat{Z}(x^*)] + [\hat{Z}(x^*) - Z^{te}(x^*)]$$

where the middle term $[\hat{Z}(\hat{x}) - \hat{Z}(x^*)]$ is at most 0 by definition of $\hat{x}$ as a minimizer of $\hat{Z}(\cdot)$. Thus, we would focus on bounding w.h.p.

$$[Z^{te}(\hat{x}) - \hat{Z}(\hat{x})] + [\hat{Z}(x^*) - Z^{te}(x^*)] \tag{D.5}$$

where the second term above can be further decomposed as:

$$(\hat{Z}(x^*) - Z^{tr}(x^*)) + (Z^{tr}(x^*) - Z^{te}(x^*)) \le \hat{Z}(x^*) - Z^{tr}(x^*) + \mathcal{V}_d(x^*) d(\mathbb{P}^{tr}, \mathbb{P}^{te}),$$

where the term $Z^{tr}(x^*) - Z^{te}(x^*)$ can also be decomposed using Lemma 3 such that:

$$Z^{tr}(x^*) - Z^{te}(x^*) \le \sqrt{2\chi^2(\mathbb{P}^{tr}, \mathbb{P}^{te}) \mathrm{Var}_{\mathbb{P}^{te}}[h(x^*;\xi)]}.$$

The term $Z^{tr}(x^*) - Z^{te}(x^*)$ cannot be avoided through all the methods based on the error decomposition, which can be regarded as the "best decision distance" under distribution shift. We make the following simplifications in the analysis part of existing regularization approaches.

- We use χ^2-divergence between $\mathbb{P}^{te}$ and $\mathbb{P}^{tr}$ to evaluate the extent of distribution shift.
- $\mathrm{Var}_{tr}[h(x^{tr};\xi)] \approx \mathrm{Var}_{te}[h(x^*;\xi)]$. The "variablity" does not change across shift, i.e. $\mathcal{V}_d(x^{tr})$ under $\mathbb{P}^{tr}$ is the same order as $\mathcal{V}_d(x^{te})$ under $\mathbb{P}^{te}$.

ERM: Consider x^{ERM} as the minimizer of $\min_{x\in\mathcal{X}} \hat{\mathbb{E}}^{tr}[h(x;\xi)]$ where $\hat{\mathbb{E}}^{tr}[\cdot]$ denotes the empirical expectation from the training set. Now, consider

$$Z^{te}(x^{ERM}) - \hat{Z}^{tr}(x^{ERM}) = [Z^{te}(x^{ERM}) - Z^{tr}(x^{ERM})] + [Z^{tr}(x^{ERM}) - \hat{Z}^{tr}(x^{ERM})],$$

where the first term can be further bounded by:

$$\begin{aligned} Z^{te}(x^{ERM}) - Z^{tr}(x^{ERM}) &= \mathbb{E}^{tr}[(\frac{d\mathbb{P}^{te}}{d\mathbb{P}^{tr}} - 1) h(x^{ERM};\xi)] \le \sqrt{2\chi^2(\mathbb{P}^{te}, \mathbb{P}^{tr}) \mathrm{Var}_{tr}[h(x^{ERM};\xi)]} \\ &\le \sqrt{2\chi^2(\mathbb{P}^{te}, \mathbb{P}^{tr}) \left(\mathrm{Var}_{tr}[h^2(x^{tr};\xi)] + M\sqrt{\frac{\mathrm{Comp}(\mathcal{H}) M}{n}} \right)}. \end{aligned}$$

where the last inequality is bounded by $\mathbb{E}^{tr}[h^2(x^{ERM};\xi)] - \mathbb{E}^{tr}[h^2(x^{tr};\xi)] \le 2M(Z^{tr}(x^{ERM}) - Z^{tr}(x^{tr}))$, which then reduces to Lemma 4. For the N−ERM case, denoting the generalization error in Lemma 4 as $\mathcal{E}_N(n, \mathbb{P}^{tr}, \mathcal{H}, x^{tr})$ (e.g., in Lemma 4 where there is no distribution shift, we have $\mathbb{P}^{tr} = \mathbb{P}^*, x^{tr} = x^*$). Besides, denote $\mathcal{E}(\mathbb{P}^{tr}, \mathbb{P}^{te}, \mathcal{H}) := |Z^{te}(x^{ERM}) - Z^{tr}(x^{ERM})| + |Z^{te}(x^*) - Z^{tr}(x^*)|$ bounded before. which characterizes the distribution shift effects on the optimization models over the complexity class. Thus, overall we have the generalization error:

$$Z^{te}(x^{N-ERM}) - Z^{te}(x^*) \le \mathcal{E}(\mathbb{P}^{tr}, \mathbb{P}^{te}, \mathcal{H}) + \mathcal{E}_N(n, \mathbb{P}^{tr}, \mathcal{H}, x^{tr}),$$

where we would incur an additional term $\mathcal{E}(\mathbb{P}^{tr}, \mathbb{P}^{te}, \mathcal{H}) = \sqrt{d(\mathbb{P}^{te}, \mathbb{P}^{tr})(\mathcal{V}_d^2(x^*) + M\sqrt{\frac{\text{Comp}(\mathcal{H})M}{n}})}$ comparing with the case without distribution shift. In addition to $d(\mathbb{P}^{te}, \mathbb{P}^{tr})(\mathcal{V}_d(x^*)$, the additional error of N−ERM due to distribution shift is $d(\mathbb{P}^{te}, \mathbb{P}^{tr})M^{\frac{3}{4}}(\frac{\text{Comp}(\mathcal{H})}{n})^{\frac{1}{4}}$.

We turn to consider the DRO perspective, i.e. $x^{N-DRO} \in \arg\min_{x\in\mathcal{X}} \max_{d(\mathbb{Q},\hat{\mathbb{P}}_n^{tr})\le\varepsilon} \mathbb{E}_{\mathbb{Q}}[h(x;\xi)]$, where $\hat{\mathbb{P}}_n^{tr}$ is the empirical distribution of the training set.

DRO from the regularization perspective: Similarly, by choosing the ambiguity size ε properly such that it satisfies the condition in order for (C.3) to hold and denote RHS of (C.3) to be $\mathcal{E}^{DRO}(n, \mathbb{P}^{tr}, \mathcal{H}, x^*; \mathcal{A}) := |Z^{te}(x^{N-DRO}) - Z^{tr}(x^{N-DRO})| + |Z^{te}(x^*) - Z^{tr}(x^*)|$ without distribution shift. By the error decomposition in the shifted case, following similar analysis, we would get the following bound w.h.p.:

$$Z^{te}(x^{N-DRO}) - Z^{te}(x^*) \le \sqrt{d(\mathbb{P}^{te}, \mathbb{P}^{tr})\left(\mathcal{V}_d^2(x^*) + M\mathcal{V}_d(x^*)\sqrt{\frac{\text{Comp}(\mathcal{H})}{n}}\right)} + \mathcal{E}^{DRO}(n, \mathbb{P}^{tr}, \mathcal{H}, x^*; \mathcal{A}).$$

In addition to $d(\mathbb{P}^{te}, \mathbb{P}^{tr})\mathcal{V}_d(x^*)$, the additional error of N−DRO due to distribution shift is $d(\mathbb{P}^{te}, \mathbb{P}^{tr})\mathcal{V}_d^{\frac{1}{2}}(x^*)M^{\frac{1}{2}}(\frac{\text{Comp}(\mathcal{H})}{n})^{\frac{1}{4}}$.

DRO from the pessimism perspective: **For those d in the ambiguity set which is IPM**, by choosing the radius $\varepsilon \ge d(\mathbb{P}^{te}, \mathbb{P}^{tr}) + O(n^{-\frac{1}{\alpha(D_\xi)}})$ to cover the test distribution $\mathbb{P}^{te}$ properly, we could get

$$Z^{te}(x^{N-DRO}) - Z^{te}(x^*) \le \mathcal{V}_d(x^*)(n^{-\frac{1}{\alpha(D_\xi)}} + d(\mathbb{P}^{te}, \mathbb{P}^{tr})), \tag{D.6}$$

where $\alpha(D_\xi)$ is some function of the dimension measure of the distribution $\mathbb{P}$. where dim is some dimension measure of the distribution $\mathbb{P}^{tr}$. Thee bound (D.6) is achieved by using a ball size ε large enough such that it covers the true $\mathbb{P}^{te}$ w.h.p., where in the case of distribution shift for the empirical distribution under Wasserstein distance, this would be typically $\frac{1}{n^{1/D_\xi}} + d(\mathbb{P}^{tr}, \mathbb{P}^{te})$. Formal type of the results have been established in Theorem E.3 in Zeng and Lam (2022). Note that this requires knowledge of the extent of distribution shift, i.e., $d(\mathbb{P}^{tr}, \mathbb{P}^{te})$ (or an upper bound of it).

Now, comparing the pessimism-based DRO with ERM or regularization-based DRO, if the upper bound of $d(\mathbb{P}^{tr}, \mathbb{P}^{te})$ is not too loose, the former avoids the product of a distance between $\mathbb{P}^{tr}$ and $\mathbb{P}^{te}$ and a term that depends on the complexity measure of the loss function class. Nonetheless, pessimism-based DRO would depend on the dimension of the training distribution, but in this case possibly acceptable as the dominant quantity can be the distribution shift amount $d(\mathbb{P}^{tr}, \mathbb{P}^{te})$.

Although triangular inequality does not hold for general f-divergence, we can apply some "pseudo distance" decomposition in some f-divergence below when the support of $\mathbb{P}^{te}, \mathbb{P}^{tr}, \hat{\mathbb{Q}}$ is the same.

Lemma 5 ("Almost" **triangular inequality** for some f-divergence)**.** *Considering the relationship between $\mathbb{P}^{te}, \mathbb{P}^{tr}, \hat{\mathbb{Q}}$ (under the same support), we have:*

$$\chi^2(\mathbb{P}^{te}, \hat{\mathbb{Q}}) \le 2\left\|\frac{d\mathbb{P}^{tr}}{d\hat{\mathbb{Q}}}\right\|_\infty \chi^2(\mathbb{P}^{te}, \mathbb{P}^{tr}) + 2\chi^2(\mathbb{P}^{tr}, \hat{\mathbb{Q}}).$$

$$KL(\mathbb{P}^{te}, \hat{\mathbb{Q}}) \le KL(\mathbb{P}^{te}, \mathbb{P}^{tr}) + \left\|\frac{d\mathbb{P}^{te}}{d\mathbb{P}^{tr}}\right\|_\infty KL(\mathbb{P}^{tr}, \hat{\mathbb{Q}}).$$

Proof. Proof of Lemma 5. Since we are only considering the continuous distribution class, we denote the density of $\mathbb{P}^{te}, \mathbb{P}^{tr}, \hat{\mathbb{Q}}$ as f, g, h respectively.

For χ^2-divergence, we have:

$$\int \frac{(f-h)^2}{h}d\mu \le \int \frac{2(f-g)^2 + 2(g-h)^2}{h}d\mu \le 2\left\|\frac{g}{h}\right\|_\infty \int \frac{(f-g)^2}{g}d\mu + 2\int \frac{(g-h)^2}{h}d\mu.$$

For KL-divergence, we have:

$$\int f\ln\frac{f}{h}d\mu = \int f\left(\ln\frac{f}{g} + \ln\frac{g}{h}\right)d\mu \le \int f\ln\frac{f}{g}d\mu + \left\|\frac{f}{g}\right\|_\infty \int g\ln\frac{g}{h}d\mu. \square$$

This types of inequalities means that we can derive similar upper bounds in (D.6) for the nonparametric estimators such as KDE for smooth and absolute continuous densities to obtain $\mathcal{V}_d(x^*)(n^{-\frac{1}{\alpha(D_\xi)}} + d(\mathbb{P}^{te}, \mathbb{P}^{tr}))$ for some large n.

We next analyze the parameteric methods in Section 3.1.

Proofs of Corollary 1 and 2. These results directly follow assuming the exact or "almost" triangular inequality holds for the metric $d(\mathbb{P}^{te}, \hat{\mathbb{Q}}) \leq c_1 d(\mathbb{P}^{te}, \mathbb{P}^{tr}) + c_2 d(\mathbb{P}^{tr}, \hat{\mathbb{Q}})$ under appropriate conditions.

For the P−DRO problem (i.e. Corollary 1), ignoring the constant, if $\varepsilon \geq \Delta(\delta, \Theta) + d(\mathbb{P}^{te}, \mathbb{P}^{tr})$, when d is an IPM metric, by (D.5), with probability at least $1-\delta$, we have:

$$\begin{aligned}\mathcal{E}(x^{P-DRO}) &\leq |Z^{te}(x^{P-DRO}) - \hat{Z}(x^{P-DRO})| + \hat{Z}(x^*) - Z^{te}(x^*) \\ &\leq 0 + \max_{d(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x^*;\xi)] - \mathbb{E}^{te}[h(x^*;\xi)] \\ &\leq 2\mathcal{V}_d(x^*)\varepsilon,\end{aligned}$$

where the second inequality holds due to $\mathbb{P}[\mathbb{P}^{te} \in \mathcal{A}(\hat{\mathbb{Q}}; d, \varepsilon)] \geq 1-\delta$ such that $Z^{te}(\cdot) \leq \hat{Z}(\cdot)$ with probability at least $1-\delta$. And the third inequality returns to the case (a) in the proof of Theorem 1. The other cases of the metric d under (b) and (c) follow similarly. This result follows the same proof structure compared with the previous DRO from the pessimism perspective, and does not pay for additional terms due to the distribution shift besides $d(\mathbb{P}^{te}, \mathbb{P}^{tr})$.

For the P−ERM case (i.e. Corollary 2), when d is an IPM metric, we have:

$$\begin{aligned}\mathcal{E}(x^{P-ERM}) &\leq 2\sup_{x\in\mathcal{X}} |Z^{te}(x) - \hat{Z}(x)| \\ &\leq 2(\sup_{x\in\mathcal{X}} \mathcal{V}_d(x)) d(\mathbb{P}^{te}, \hat{\mathbb{Q}}) \leq 2(\sup_{x\in\mathcal{X}} \mathcal{V}_d(x))(d(\mathbb{P}^{te}, \mathbb{P}^{tr}) + d(\mathbb{P}^{tr}, \hat{\mathbb{Q}})).\end{aligned}$$

Therefore, in Corollary 2, with respect to all the sub cases in Theorem 2 by replacing $\Delta(\delta, \Theta)$ with $\Delta(\delta, \Theta) + d(\mathbb{P}^{te}, \mathbb{P}^{tr})$, in addition to the $\mathcal{V}_d(x^*)d(\mathbb{P}^{te}, \mathbb{P}^{tr})$ term, the additional error of P−ERM still need to pay due to distribution shift is at least $Md(\mathbb{P}^{te}, \mathbb{P}^{tr})^{\frac{3}{4}} + \sqrt{2d(\mathbb{P}^{te}, \mathbb{P}^{tr})}\sqrt{\text{Var}_{te}[h(x^*;\xi)]}$ by part (c).

D.3 Required Monte Carlo Size for P−DRO in Section 3.3

In this part, we denote $x^{P-DRO_m} \in \arg\min_{x\in\mathcal{X}} \max_{d(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$. We generally would like to investigate the required sample size m such that $\mathcal{E}(x^{P-DRO_m}) \approx \mathcal{E}(x^{P-DRO})$ in Theorem 1. The idea is to let the Monte Carlo sampling error dominated by the statistical generalization error in each P−DRO case. In the main text, for simplicity of the notation, we ignore the dependency of $\log(1/\delta)$, but the sample size is at most polynomial level of that term, i.e. $m \approx \log(1/\delta))^k$ for some constant k.

we concretize the complexity term appearing in the main text (e.g. Table 1, Theorem 4) $\text{Comp}(\mathcal{H}) \approx \frac{\text{Comp}_n(\mathcal{H})}{(\log n)^k}$ for some constant k as the log of Covering number. And we mainly use $\text{Comp}_m(\mathcal{H})$ in the proof here. We follow the notations of covering number from Section 2.2.2 in Maurer and Pontil (2009); Duchi and Namkoong (2019). That is to say, for $\varepsilon > 0$, a function class $\mathcal{H}$ and an integer n, the "empirical ℓ_∞ covering number" $\mathcal{N}_\infty(\mathcal{H}, \varepsilon, n)$ is defined to be:

$$\mathcal{N}_\infty(\mathcal{H}, \varepsilon, n) = \sup_{\boldsymbol{\xi}\in\Xi^n} \mathcal{N}(\mathcal{H}(\boldsymbol{\xi}), \varepsilon, \|\cdot\|_\infty), \tag{D.7}$$

where $\mathcal{H}(\boldsymbol{\xi}) = \{(h(\xi_1), \ldots, h(\xi_n)) : h \in \mathcal{H}\} \subseteq \mathbb{R}^n$ and for $A \subseteq \mathbb{R}^n$, the number $\mathcal{N}(A, \varepsilon, \|\cdot\|_\infty)$ is the smallest cardinality $|A'|$ of a set $A' \subseteq A$ such that $A \subset \cup_{x_0\in A'}\{x : \|x - x_0\|_\infty \leq \varepsilon\}$. **We denote $\text{Comp}_n(\mathcal{H}) := \log \mathbf{N}_\infty(\mathcal{H}, \frac{\mathbf{1}}{\mathbf{n}}, \mathbf{n})$ below**, which usually scales as $(\log n)^k$ in n for $\mathcal{H}$ in practice.

We also suppress the $\log(1/\delta)$ dependency inside $\text{Comp}_m(\mathcal{H})$. That is to say, when we use the argument "with probability at least $1-\delta, \cdots \leq \sqrt{\text{Comp}_m(\mathcal{H})}$", we are referring to "$\cdots \leq \sqrt{\text{Comp}_m(\mathcal{H}) + \log(1/\delta)}$".

We first present the proof of generalization bounds for Wasserstein P−DRO with Monte Carlo errors, i.e. Theorem 3 where $m \approx n^{\alpha D_\xi}$.

D.3.1 Proof of Theorem 3.

In this case, we assume each distribution in $\mathcal{P}_\Theta$ satisfies the conditions in Lemma 1.

Comparing it with Theorem 1, we only need to show that $\mathbb{P}^* \in \mathcal{A}(\hat{\mathbb{Q}}_m; W_1, \varepsilon)$ with probability at least $1-\delta$. And other parts follow directly by the case (a) in Theorem 1 only need to consider $\mathcal{V}_d(x) = \|h(x;\cdot)\|_{\text{Lip}}$. By triangular inequality of Wasserstein distance, we have:

$$\begin{aligned} W_1(\mathbb{P}^*, \hat{\mathbb{Q}}_m) &\le W_1(\mathbb{P}^*, \hat{\mathbb{Q}}) + W_1(\hat{\mathbb{Q}}, \hat{\mathbb{Q}}_m) \\ &\le \frac{\varepsilon}{2} + \left(\frac{C}{m}\right)^{\frac{1}{D_\xi}} \log(1/\delta) \le \varepsilon, \end{aligned}$$

where the last inequality holds when $\frac{\varepsilon}{2} \ge \Delta(\delta, \Theta)$ in Assumption 1 and $\left(\frac{C}{m}\right)^{\frac{1}{D_\xi}} \log(1/\delta) \le \frac{\varepsilon}{2}$, i.e. $m \ge C(\frac{2\log(1/\delta)}{\varepsilon})^{D_\xi}$ for some constant C. Then the subsequent steps are analogous to proof of the first part in Theorem 1, only replacing $\hat{\mathbb{Q}}$ with $\mathbb{P}^*$. □

D.3.2 Statement of Theorem 4.

Below, we present our specific results as well as their proofs with respect to **χ^2-divergence** and **general Wasserstein distance** where $m \approx \text{Comp}(\mathcal{H})n^\alpha$ which is independent with D_ξ but dependent with the **complexity term** $\text{Comp}(\mathcal{H})$. Then, the general case in Theorem 4 in the main text is made up of the following sub results with more specific conditions.

Theorem 7 (Generalization bounds for χ^2 P−DRO with Monte Carlo errors)**.** *Suppose Assumption 1 holds and the cost function $h(x;\xi) \in [0, M], \forall x, \xi$ with $Var_{\mathbb{P}^*}[h(x^*;\cdot)] > 0$. The size of the ambiguity set $\varepsilon \ge \Delta(\delta, \Theta)$. If Monte Carlo size $m \ge C_0 \left(\frac{LM}{\sqrt{Var_{\mathbb{P}^*}[h(x^*;\cdot)]\varepsilon}}\right)^2 Comp_m(\mathcal{H})$ for some numerical constant C_0, when d is χ^2-divergence, then with probability at least $1-\delta$, we have:*

$$\mathcal{E}(x^{P-DRO_m}) \le \begin{cases} 2\mathcal{E}_{\chi^2} + C_1\sqrt{\frac{\varepsilon}{L}}M, & \text{if } Var_{\hat{\mathbb{Q}}}[h(x^{P-DRO_m};\xi)] \le 2\varepsilon M^2 \\ 2\mathcal{E}_{\chi^2}, & \text{otherwise} \end{cases},$$

where $L \ge 1$ and $\mathcal{E}_{\chi^2}$ is the generalization error upper bound in the case (c) of Theorem 1 in the main text.

Remark 4. *Although this result depends on another term L, due to "incomplete" exact variance regularization of χ^2-divergence, when $Var_{\mathbb{P}^*}[h(x;\xi)]$ is sufficiently large, as long as the required Monte Carlo size $m \ge C_0 \left(\frac{M}{\sqrt{Var_{\mathbb{P}^*}[h(x^*;\cdot)\varepsilon}}\right)^2 Comp_m(\mathcal{H})$, $\mathcal{E}(x^{P-DRO_m}) \le 2\mathcal{E}_{\chi^2}$.*

On the other hand, even if the variance is not enough, as long as $\sqrt{\frac{\varepsilon}{L}}M \le \mathcal{E}_{\chi^2} \le \sqrt{Var_{\mathbb{P}^*}[h(x^*;\cdot)]\varepsilon}$, i.e. $L \ge \frac{M^2}{Var_{\mathbb{P}^*}[h(x^*;\cdot)]}$ and therefore $m \ge \left(\frac{LM}{\sqrt{Var_{\mathbb{P}^*}[h(x^*;\cdot)]\varepsilon}}\right)^6 Comp_m(\mathcal{H})$ for some numerical constant C_0, we still have $\mathcal{E}(x^{P-DRO_m}) \le 3\mathcal{E}_{\chi^2}$. The dependence of the required sample size m is also independent with the distribution dimension and we hope to bridge the variance gap in our later work in this region.

Similarly, following similar proof structure, we can also obtain a dimension-free required Monte Carlo sample size for the Wasserstein case. This may be better than the result in Corollary 3 of main text where the degrading effects of $\text{Comp}(\mathcal{H})$ is smaller than that of D_ξ to the generalization error. We consider 1-Wasserstein distance here.

Theorem 8 (Generalization bounds for 1-Wasserstein P−DRO with Monte Carlo errors)**.** *Suppose Assumption 1 holds and the cost function $|h(x;\xi)| \le M, \forall x, \xi$ with $\|h(x^*;\cdot)\|_{Lip} > 0$ with some proper conditions of $\mathcal{H}$ in Lemma 8. The size of the ambiguity set $\varepsilon \ge \Delta(\delta, \Theta)$. If Monte Carlo size $m \ge C_0(\frac{M}{\|h(x^*;\cdot)\|_{Lip}\varepsilon})^2 Comp_m(\mathcal{H})$ for some numerical constant C_0, when d is 1-Wasserstein distance, then with probability at least $1-\delta$, we have: $\mathcal{E}(x^{P-DRO_m}) \|h(x^*;\cdot)\|_{Lip}\varepsilon$.*

Since the overall proof of these two results are a bit involved, we move them to the next two subsections. We first present two uniform concentration inequalities under $\mathcal{H}$ for the empirical mean and variance here under the assumptions listed here, i.e. $0 \le h(x;\xi) \le M, \forall x, \xi$.

Lemma 6 (Uniform Hoeffding Inequality, based on Maurer and Pontil (2009)). *Under the problem setup, with probability at least $1-\delta$, we have:*

$$\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] \leq C_1 M \sqrt{\frac{Comp_m(\mathcal{H})}{m}}, \tag{D.8}$$

where C_1 is some numerical constant independent with the function complexity and sample size.

Lemma 7 (Uniform Variation Concentration Inequality). *Under the problem setup, with probability at least $1-\delta$, we have:*

$$\sqrt{Var_{\hat{\mathbb{Q}}_m}[h(x;\xi)]} \geq \sqrt{1-\frac{1}{m}}\sqrt{Var_{\hat{\mathbb{Q}}}[h(x;\xi)]} - \frac{2M^2}{m} - C_2 M\sqrt{\frac{Comp_m(\mathcal{H})}{m}}, \tag{D.9}$$

where C_2 is some numerical constant independent with the function complexity and sample size.

Proof. These properties are directly based on the variance concentration inequality (adapted from Lemma A.1 in Duchi and Namkoong (2019)), $\forall x \in \mathcal{X}$, when $m \geq 3$, we have with probability at least $1-\delta$:

$$\sqrt{\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]} \geq \sqrt{1-\frac{1}{m}}\sqrt{\text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]} - \frac{2M^2}{m} - M\sqrt{\frac{2\log(1/\delta)}{m}}. \tag{D.10}$$

$$\sqrt{\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]} \leq \sqrt{1+\frac{1}{m}}\sqrt{\text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]} + M\sqrt{\frac{2\log(1/\delta)}{m}}. \tag{D.11}$$

Then the second term of RHS in (D.10), we taking (D.10) for the uniform covering number version to obtain (D.9). □

D.3.3 Proof of Theorem 7.

The proof is divided into the following three steps. For simplicity, we denote $\hat{Z}(x) = \sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$ and our discrete approximation in practice $\hat{Z}_m(x) = \sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$. We will show that $\sup_{x\in\mathcal{X}} |\hat{Z}(x) - \hat{Z}_m(x)|$ is small so that we can leverage on results in Theorem 1. Before beginning our main proof, we first state an decomposition result for the empirical variance compared with the true variance under with probability at least $1-\delta, \forall x \in \mathcal{X}$:

$$\begin{aligned}\left|\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] - \text{Var}_{\mathbb{P}^*}[h(x;\xi)]\right| &= \left|\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] - \text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]\right| + \left|\text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \text{Var}_{\mathbb{P}^*}[h(x;\xi)]\right| \\ &\leq M^2\left(C_1\sqrt{\frac{\text{Comp}_m(\mathcal{H})}{m}} + 3\sqrt{2\varepsilon}\right),\end{aligned} \tag{D.12}$$

where the first term in the inequality follows from the uniform Hoeffding inequality. And the second term in the inequality follows by the value of ε such that $\chi^2(\mathbb{P}^*,\hat{\mathbb{Q}}) \leq \varepsilon$ and $\|h\|_\infty \leq M$ in (D.2).

Step 1: Variance Regularization. Following Lemma 3, we have:

$$\sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] \leq \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] + \sqrt{2\varepsilon \text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}, \tag{D.13}$$

$$\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] \leq \sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] \leq \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] + \sqrt{2\varepsilon \text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}, \tag{D.14}$$

We now show the condition for n so that the exact equality of RHS holds in (D.14). Note that $\sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]$ is equivalent to the value of the following optimization problem:

$$\max_{p\in\mathbb{R}_+^m} \sum_{i=1}^m p_i h(x;\xi_i), \text{ s.t. : } \sum_{i=1}^m (p_i - \frac{1}{m})^2 \leq \frac{2\varepsilon}{m}, \sum_{i=1}^m p_i = 1.$$

The the maximizing value of this problem is attainable to RHS in the second inequality of (D.14) whenever $\sqrt{2\varepsilon}\frac{h(x;\xi)-\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}{\sqrt{\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}} \geq -1$. Since $h(x;\xi) \in [0,M], \forall x, \xi$, it is sufficient to satisfy the following inequality:

$$\text{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)] \geq 2\varepsilon M^2, \forall x \in \mathcal{X}. \tag{D.15}$$

In general, based on the analysis of the equality condition of Cauchy inequality, we reach the following result as a more refined **variance-dependent** lower bound of $\sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}}_m)\le\varepsilon}\mathbb{E}_{\mathbb{P}}[h(x;\xi)]$:

$$\sup_{\chi^2(\mathbb{P},\hat{\mathbb{Q}}_m)\le\varepsilon}\mathbb{E}_{\mathbb{P}}[h(x;\xi)]\ge\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]+\sqrt{\Delta\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}, \tag{D.16}$$

as long as $\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\ge\Delta M^2$.

For any integer $L\ge 1$ and a given $\hat{\mathbb{Q}}$ output from the distribution estimator, we split the decision space $\mathcal{X}$ into the following regions $\mathcal{X}_0\cup\mathcal{X}_1\cup\ldots\mathcal{X}_{L+1}\cup\mathcal{X}_{L+2}$, where $\mathcal{X}_{L+2}=\{x\in\mathcal{X}:\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]\ge 2\varepsilon M^2\}$, and:

$$\mathcal{X}_\ell=\left\{x\in\mathcal{X}:\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]\in\left[\frac{\ell-1}{L}2\varepsilon M^2,\frac{\ell}{L}2\varepsilon M^2\right)\right\},\forall\ell\in\{1,\ldots,L+1\}.$$

Before conducting detailed analysis of this variance regularization effect, we first let the Monte Carlo size satisfy for (D.8):

$$MC_1\sqrt{\frac{\mathrm{Comp}_m(\mathcal{H})}{m}}:=\Delta_{\mathbb{E}}\le\frac{2\varepsilon M}{L}, \tag{D.17}$$

so that with probability at least $1-\delta$, by (D.12) we have:

$$\left|\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]-\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\right|\le\frac{2\varepsilon M^2}{L},\forall x\in\mathcal{X} \tag{D.18}$$

Step 2: Monte Carlo Error Decomposition to bound $\hat{Z}(x)-\hat{Z}_m(x)$. We consider the decision variable under different regimes.

(a) $\forall x\in\mathcal{X}_{L+2}$, by (D.18), (D.15) holds and therefore by (D.13) and (D.16), we obtain:

$$\begin{aligned}\hat{Z}(x)-\hat{Z}_m(x)&\le\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]+\sqrt{2\varepsilon\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}-\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]-\sqrt{2\varepsilon\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}\\&=(\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]-\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)])+\sqrt{2\varepsilon}(\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}-\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}).\end{aligned} \tag{D.19}$$

For the first term RHS in (D.19), we let the Monte Carlo size m satisfies $C_1M\sqrt{\frac{\mathrm{Comp}_m(\mathcal{H})}{m}}:=\Delta_{\mathbb{E}}\le\frac{2\varepsilon M}{L}$ in (D.8) in Lemma 6. Then, combining (D.8) and (D.9) into (D.19), we obtain with probability at least $1-\delta$, $\forall x\in\mathcal{X}$:

$$\begin{aligned}\hat{Z}(x)-\hat{Z}_m(x)&\le\Delta_{\mathbb{E}}+\sqrt{2\varepsilon}\left(C_2M\sqrt{\frac{\mathrm{Comp}_m(\mathcal{H})}{m}}+(1-\sqrt{1-\frac{1}{m}})\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}+\frac{2M^2}{m}\right)\\&\le\Delta_{\mathbb{E}}+C_2M\sqrt{\frac{2\varepsilon\mathrm{Comp}_m(\mathcal{H})}{m}}+\frac{\sqrt{2\varepsilon}(2M^2+\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)])}{m}\\&\le C_3\Delta_{\mathbb{E}}+\frac{3\sqrt{2\varepsilon}M^2}{m}\le C_3'\Delta_E,\end{aligned}$$

where C_3,C_3' is another numerical constant independent with the function complexity and sample size. The second inequality follows by the IPM property of TV distance.

(b) $\forall x\in\mathcal{X}_i,i\in\{1,\ldots,L+1\}$, we have $\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\ge\max\{\frac{i-2}{L}2\varepsilon M^2,0\}$.

(b.1) If $i\ge 2$, by (D.13) and (D.16) as well as the definition of $\mathcal{X}_i$, we have:

$$\begin{aligned}\hat{Z}(x)-\hat{Z}_m(x)&\le\left(\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]-\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\right)+\sqrt{\frac{i}{L}2\varepsilon\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}-\sqrt{\frac{i-2}{L}2\varepsilon\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}\\&\le\Delta_{\mathbb{E}}+\sqrt{\frac{2(i-2)\varepsilon}{L}}\left(\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}-\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}\right)+\sqrt{\frac{4\varepsilon}{L}\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}\\&\le\Delta_{\mathbb{E}}+\sqrt{\frac{2(i-2)\varepsilon}{L}}\frac{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]-\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}{\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]}+\sqrt{\mathrm{Var}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]}}+\sqrt{\frac{4\varepsilon}{L}\frac{2i}{L}M^2}\\&\le\Delta_{\mathbb{E}}+\sqrt{\frac{2(i-2)\varepsilon}{L}}\frac{2\varepsilon M^2/L}{2\sqrt{2(i-2)\varepsilon M^2/L}}+C_4'\sqrt{\frac{\varepsilon}{L}}M\le\Delta_{\mathbb{E}}+C_4\frac{\varepsilon}{L}M+C_4'\sqrt{\frac{\varepsilon}{L}}M.\end{aligned} \tag{D.20}$$

(b.2) If $i = 1$, by (D.13), by definition of $\mathcal{X}_1$, we have:

$$\begin{aligned}\hat{Z}(x) - \hat{Z}_m(x) &\leq \left(\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\right) + \sqrt{\frac{2\varepsilon}{L}\text{Var}_{\hat{\mathbb{Q}}}[h(x;\xi)]} \\ &\leq \Delta_{\mathbb{E}} + \frac{2\varepsilon M}{L}.\end{aligned} \tag{D.21}$$

In general, combining different subcases in (b), with probability at least $1-\delta$, $\forall x \in \mathcal{X}\backslash\mathcal{X}_{L+2}$, if we let $\frac{\varepsilon}{L} \leq \sqrt{\frac{\varepsilon}{L}} \leq 1$, then we finally have:

$$\hat{Z}(x) - \hat{Z}_m(x) \leq C_0\sqrt{\frac{\varepsilon}{L}}M. \tag{D.22}$$

Step 3: Generalization Error Decomposition. Plugging the solution x^{P-DRO_m} into (D.22), we have:

$$\begin{aligned}&Z(x^{P-DRO_m}) - \hat{Z}_m(x^{P-DRO_m}) \\ &\leq \hat{Z}(x^{P-DRO_m}) - \hat{Z}_m(x^{P-DRO_m}) \leq \begin{cases}\Delta_{\mathbb{E}} + C_0\sqrt{\frac{\varepsilon}{L}}M & \text{if } x \notin \mathcal{X}_{L+2} \\ C_0'\Delta_{\mathbb{E}} & \text{otherwise}\end{cases}.\end{aligned} \tag{D.23}$$

for some constant C_0 when L is large, where the first inequality follows by Assumption 1 and Theorem 1, with probability at least $1-\delta$, $\mathbb{P}^* \in \mathcal{A}(\hat{\mathbb{Q}}; \chi^2, \varepsilon)$ when $\varepsilon \geq \Delta(\delta, \Theta)$. Finally:

$$\hat{Z}_m(x^*) - Z(x^*) \leq (\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]) + \sqrt{2\varepsilon\text{Var}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)]},$$

where the first term is bounded by Bernstein inequality, with probability at least $1-\delta$:

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] &\leq (\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]) + (\mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)] - \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)]) \\ &\leq \sqrt{\frac{2\text{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]\log(1/\delta)}{m}} + \frac{\|h(x^*;\cdot)\|_\infty \log(1/\delta)}{3m} + \sqrt{2\varepsilon\text{Var}_{\mathbb{P}^*}[h(x^*;\xi)]} \\ &\leq \|h(x^*;\cdot)\|_\infty\left(\sqrt{\frac{2\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{3m}\right) + \sqrt{2\varepsilon\text{Var}_{\mathbb{P}^*}[h(x^*;\xi)]}\end{aligned} \tag{D.24}$$

And for the second term, by (D.11), we have with probability at least $1-\delta$:

$$\sqrt{2\varepsilon\text{Var}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)]} \leq \sqrt{2\varepsilon\text{Var}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]} + 2\|h(x^*;\cdot)\|_\infty\sqrt{\frac{\varepsilon\log(1/\delta)}{m}}. \tag{D.25}$$

Therefore, based on the RHS of (D.23),(D.24) and (D.25), following the same decomposition procedure,

$$Z(x^{P-DRO_m}) - Z(x^*) \leq (Z(x^{P-DRO_m}) - \hat{Z}_m(x^{P-DRO_m})) + (\hat{Z}_m(x^*) - Z(x^*)).$$

Therefore, comparing the preliminary results in the case (c) of Theorem 1, as long as m is large enough, we can attain the similar generalization bound then. □

D.3.4 Proof of Theorem 8.

Before providing the proof, we first present the following lemma:

Lemma 8 (Based on Theorem 6.3 of Esfahani and Kuhn (2018) and Corollary 2 of Gao et al. (2022)). *Assume $h(x;\xi)$ is Lipschitz continuous and convex w.r.t. ξ. If Ξ is unbounded and there exists $\xi_0 \in \mathcal{Z}$ such that $\limsup_{\|\tilde{\xi}-\xi_0\|\to\infty} \frac{h(x;\xi)-h(x;\xi_0)}{\|\tilde{\xi}-\xi_0\|} = \|h(x;\cdot)\|_{Lip}$, then for any $\hat{\mathbb{P}}$ we have:*

$$\sup_{W_1(\mathbb{P},\hat{\mathbb{P}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] = \mathbb{E}_{\hat{\mathbb{P}}}[h(x;\xi)] + \varepsilon\|h(x;\cdot)\|_{Lip}.$$

We follow similar arguments in Step 3 of proof in Theorem 7. That is,

$$\mathbb{E}_{\mathbb{P}^*}[h(\hat{x};\xi)] \overset{(a)}{\leq} \sup_{W_1(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(\hat{x};\xi)] \overset{(b)}{=} \sup_{W_1(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(\hat{x};\xi)] + (\mathbb{E}_{\hat{\mathbb{Q}}}[h(\hat{x};\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(\hat{x};\xi)]), \tag{D.26}$$

where (a) is given by $W_1(\mathbb{P}^*,\hat{\mathbb{Q}}) \leq \varepsilon$ if we take the ambiguity size ε to cover the true distribution with probability at least $1-\delta$. And (b) is given by applying Lemma 8 with the ball center is $\hat{\mathbb{Q}}$ and $\hat{\mathbb{Q}}_m$ respectively.

On the other hand, we have:

$$\begin{aligned}\sup_{W_1(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}_m}[h(x^*;\xi)] &\leq \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)] + \varepsilon\|h(x^*;\cdot)\|_{\text{Lip}} \\ &\leq \mathbb{E}_{\mathbb{P}^*}[h(x^*;\xi)] + (\mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x^*;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x^*;\xi)]) + 2\varepsilon\|h(x^*;\cdot)\|_{\text{Lip}}.\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\mathcal{E} &\leq 2\|h(x^*;\cdot)\|_{\text{Lip}}\varepsilon + 2\sup_{x\in\mathcal{X}}\left|\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\right| \\ &\leq 2\|h(x^*;\cdot)\|_{\text{Lip}}\varepsilon + 2C_1 M\sqrt{\frac{\text{Comp}_m(\mathcal{H})}{m}}.\end{aligned}$$

Letting the second term smaller than $\|h(x^*;\cdot)\|_{\text{Lip}}\varepsilon$, we obtain the sample size required in Theorem 8. □

D.3.5 Generalization Result in p-Wasserstein distance

More generally, this argument to control the following Monte Carlo error term $\forall x \in \mathcal{X}$:

$$\left|\sup_{d(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] - \sup_{d(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)]\right|,$$

can be extended to the case of p-Wasserstein distance with $p \in (1,2]$ with the help of the following Lemma.

Lemma 9 (Adapted from Lemma 1 in Gao et al. (2022)). *Under some mild first-order conditions for $h(x;\cdot) \in \mathcal{H}$ (i.e., Assumption 1 and 2 in Gao et al. (2022)), given p-Wasserstein distance with $p \in (1,2]$, for any distribution $\hat{\mathbb{Q}}$, there exists $C > 0$ such that:*

$$\left|\sup_{W_p(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \varepsilon\mathcal{V}_{\hat{\mathbb{Q}},q}(h(x;\cdot))\right| \leq C\varepsilon^p.$$

where $\mathcal{V}_{\hat{\mathbb{Q}},q}(h(x;\cdot))$ is the L_q norm of the random variable $\frac{\partial h(x;\xi)}{\partial\xi}$ under the measure $\hat{\mathbb{Q}}$ with $\frac{1}{p}+\frac{1}{q}=1$.

Note that Lemma 1 in Gao et al. (2022) is originally meant for problems with the empirical distribution $\hat{\mathbb{Q}}_m$, but when $p \in (1,2]$ it can be directly extended to a ball center with continuous distribution.

Corollary 3. *Suppose Assumption 1 in this main text and Assumption 1 and 2 in Gao et al. (2022) hold. The size of the ambiguity set $\varepsilon \geq \Delta(\delta,\Theta)$, when d is p-Wasserstein distance with $p \in (1,2]$, if the Monte Carlo size satisfies:*

$$m \geq \max\left\{\left(\frac{C_1 + C_2\tilde{M}\sqrt{Comp_m(\partial(\mathcal{H}))}}{\mathcal{V}_{\mathbb{P}^*,q}(h^*)}\right)^q, C_0\left(\frac{M}{\varepsilon\mathcal{V}_{\mathbb{P}^*,q}(h^*)}\right)^2 Comp_m(\mathcal{H})\right\},$$

where $\tilde{M} := \sup_{x\in\mathcal{X},\xi\in\mathcal{X}}\left\|\frac{\partial h(x;\xi)}{\partial\xi}\right\|_2$ and $\partial(\mathcal{H}) = \left\{\left\|\frac{\partial h(x;\xi)}{\partial\xi}\right\|_2 : x \in \mathcal{X}\right\}$ for some constant C_0, C_1, C_2, Then with probability at least $1-\delta$, we have $\mathcal{E}(x^{P-DRO_m}) \leq 4\varepsilon\mathcal{V}_{\mathbb{P}^*,q}(h(x^*;\cdot)) + C\varepsilon^p$.

Proof. For simplicity, we abbreviate $h := h(x;\cdot), h^* := h(x^*;\cdot)$ in the following. Similarly in (D.26), by the result in Lemma 9, we have the Monte Carlo error bounded by:

$$\mathbb{E}_{\mathbb{P}^*}[h] - \sup_{W_p(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h] \leq \sup_{W_p(\mathbb{P},\hat{\mathbb{Q}})\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h] - \sup_{W_p(\mathbb{P},\hat{\mathbb{Q}}_m)\leq\varepsilon} \mathbb{E}_{\mathbb{P}}[h] \leq (\mathbb{E}_{\hat{\mathbb{Q}}}[h] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h]) + \varepsilon(\mathcal{V}_{\hat{\mathbb{Q}},q}(h) - \mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)) + 2C\varepsilon^p.$$

Therefore, we would obtain:

$$\begin{aligned}\mathcal{E}(x^{P-DRO_m}) &\le 2\varepsilon \mathcal{V}_{\mathbb{P}^*,q}(h^*) + C\varepsilon^p + \sup_{h\in\mathcal{H}} \left|\mathbb{E}_{\mathbb{P}}[h] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h]\right| + \varepsilon \sup_{h\in\mathcal{H}} \left|\mathcal{V}_{\hat{\mathbb{Q}},q}(h) - \mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)\right| \\ &\le 2\varepsilon \mathcal{V}_{\mathbb{P}^*,q}(h^*) + C\varepsilon^p + C_0 M \sqrt{\frac{\text{Comp}_m(\mathcal{H})/\delta)}{m}} \\ &+ \varepsilon \left[C_1 + C_2 \tilde{M}\sqrt{\text{Comp}_m(\partial(\mathcal{H}))}\right] m^{\frac{1}{p}-1},\end{aligned}$$

And the last term of the last inequality is obtained from the uniform concentration inequality of L_p-norm for $\sup_{h\in\mathcal{H}} |\mathcal{V}_{\hat{\mathbb{Q}},q}(h) - \mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)|$. Concretely to say since $q > 2$, by Theorem 6.10 in Boucheron et al. (2013) and Lemma 7 in Duchi and Namkoong (2021), we first have:

$$\mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h) - \mathbb{E}[\mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)] \le \tilde{M} m^{-\frac{1}{q}} \sqrt{\log(1/\delta)}.$$

Then uniformly bounded in $\partial(\mathcal{H})$ of the covering number argument, we obtain $\forall h \in \mathcal{H}$, with probability at least $1-\delta$:

$$|\mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h) - \mathbb{E}[\mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)]| \le C_2 \tilde{M} m^{-\frac{1}{q}} \sqrt{\text{Comp}_m(\partial(\mathcal{H}))}. \tag{D.27}$$

And by Lemma 9 in Duchi and Namkoong (2021), under some situation conditions, by definition of $\mathcal{V}_{\mathbb{Q},q}(h)$, we have:

$$\mathcal{V}_{\hat{\mathbb{Q}},q}(h) - \frac{2}{p}\sqrt{C} n^{-\frac{1}{q}} \le \mathbb{E}[\mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)] \le \mathcal{V}_{\hat{\mathbb{Q}},q}(h). \tag{D.28}$$

Combining (D.27) and (D.28) would obtain the bound for $|\mathcal{V}_{\hat{\mathbb{Q}},q}(h) - \mathcal{V}_{\hat{\mathbb{Q}}_m,q}(h)|$. □

In general, as long as the ambiguity metric d satisfies the following **"almost exact" regularization effect** with some variability measure $\mathcal{V}_d(x)$, i.e.:

$$\left| \sup_{\mathbb{P}\in\mathcal{A}(\hat{\mathbb{Q}};d,\varepsilon)} \mathbb{E}_{\mathbb{P}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \varepsilon^\alpha \mathcal{V}_d(x) \right| = o(\varepsilon^\alpha),$$

then if the Monte Carlo size exceeds some term w.r.t. the overall complexity class $m \ge C(\text{Comp}_m(\mathcal{H}))^k$ for some constant k, we would obtain $\mathcal{E}(x^{P-DRO_m}) \approx \mathcal{E}(x^{P-DRO})$.

D.4 Required Monte Carlo Size for `P-ERM`

We use similar notations with $x^{P-ERM_m} \in \arg\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]$ and investigate the required Monte Carlo size.

Theorem 9 (Generalization bounds for `P-ERM` with Monte Carlo errors)**.** *Suppose Assumption 1 holds with metric d, denote the corresponding generalization error upper bound in Theorem 2 as $\mathcal{E}_P$. If the Monte Carlo size m satisfies:*

$$\frac{m}{M\text{Comp}_m(\mathcal{H})} \ge \max\left\{\frac{1}{Z(x^*)+\mathcal{E}_P}, \frac{Z(x^*)+\mathcal{E}_P}{\mathcal{E}_P^2}\right\},$$

then with probability at least $1-\delta$, $\mathcal{E}(x^{P-ERM_m}) \le 2\mathcal{E}_P$.

Note that no matter whether we ignore distribution misspecification error in $\mathcal{E}_P$, then the Monte Carlo size required here scales with the function complexity $M\text{Comp}_m(\mathcal{H})$ as well as n^α for some α independent with D_ξ.

Proof. The result is directly from Lemma 4 and Theorem 2. Denote $\mathcal{E}_P$ is the upper bound of $Z(x^{P-ERM}) - Z(x)$ for the generalization error result for the given continuous version, i.e. in RHS for each case in Theorem 2. Then we have:

$$\begin{aligned}Z(x^{P-ERM-S}) - Z(x^*) &\le \mathcal{E}_P + 2\sup_{x\in\mathcal{X}} \left|\mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)] - \mathbb{E}_{\hat{\mathbb{Q}}_m}[h(x;\xi)]\right| \\ &\le \mathcal{E}_P + 2\left(\sqrt{\frac{Z^{P-ERM}(x^{P-ERM})\text{Comp}_m(\mathcal{H})M}{m}} + \frac{\text{Comp}_m(\mathcal{H})M}{m}\right) \\ &\le \mathcal{E}_P + 3\left(\sqrt{\frac{(Z(x^*)+\mathcal{E}_P)\text{Comp}_m(\mathcal{H})M}{m}}\right) \le 2\mathcal{E}_P,\end{aligned}$$

where the second inequality follows from Lemma 4 since $x^{P-ERM} \in \arg\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{Q}}}[h(x;\xi)]$. And the third inequality is a result of the following chain inequality:

$$\begin{aligned} Z^{P-ERM}(x^{P-ERM}) &\le Z^{P-ERM}(x^*) \\ &= Z(x^*) + (Z^{P-ERM}(x^*) - Z(x^*)) \\ &\le Z(x^*) + \mathcal{E}_P. \end{aligned}$$

as in Theorem 2 and the Monte Carlo size $m \ge \frac{M\text{Comp}_m(\mathcal{H})}{Z(x^*)+\mathcal{E}_P}$. The last inequality holds as long as $m \ge \frac{(Z(x^*)+\mathcal{E}_P)M\text{Comp}_m(\mathcal{H})}{\mathcal{E}_P^2}$ further. □

D.5 A Short Discussion of Stochastic Approximation Methods

Besides the Monte Carlo approach mentioned in the main text, another approach investigated in the literature to directly tackle stochastic optimization with underlying continuous distribution $\hat{\mathbb{Q}}$ is through stochastic approximation (SA).

In SA, we apply stochastic gradient descent (SGD) to obtain a batch of samples from $\hat{\mathbb{Q}}$ in each step in each iteration. For example in `Param-ERM` case, after a number of iterations, we can obtain a solution $\hat{x}$ with the **expected** generalization error (since SGD introduces another type of uncertainty due to random sampling to compute gradients) with a polynomial number of iterations w.r.t. $\frac{1}{\gamma}(\gamma > 0)$ (such as Nemirovski et al. (2009)):

$$\mathbb{E}_{\hat{x}}[Z(\hat{x}) - Z(x^*)] \le \mathcal{E}(x^{P-ERM}) + \gamma. \tag{D.29}$$

For some DRO approaches, we can still express our optimization objective as $\min_{x,y\in\mathcal{X}\times\mathcal{Y}} \mathbb{E}_{\hat{\mathbb{Q}}}[G(x,y)]$ for some auxiliary variable y to apply this method. For example, by duality under general f-divergence, shown in Theorem 5.3 of Rahimian and Mehrotra (2019), our optimization problem can be reformulated as:

$$\inf_{x\in\mathcal{X}} \inf_{\lambda\ge 0,\mu\in\mathbb{R}} \{\mu + \lambda\varepsilon + \mathbb{E}_{\xi\sim\hat{\mathbb{Q}}}[(\lambda f)^*(h(x;\xi)-\mu)]\}.$$

Then we can solve the DRO problem under general f-divergence by SA. We will also investigate the properties of them in our future work.

E Complete Experiment Setups and Results

The optimization problems throughout this paper are all convex and solved by CVX and Gurobi implemented by Python 3.8.5. The computational environment is an Intel(R) Core(TM) i7-8650U CPU @1.90GHz personal computer.

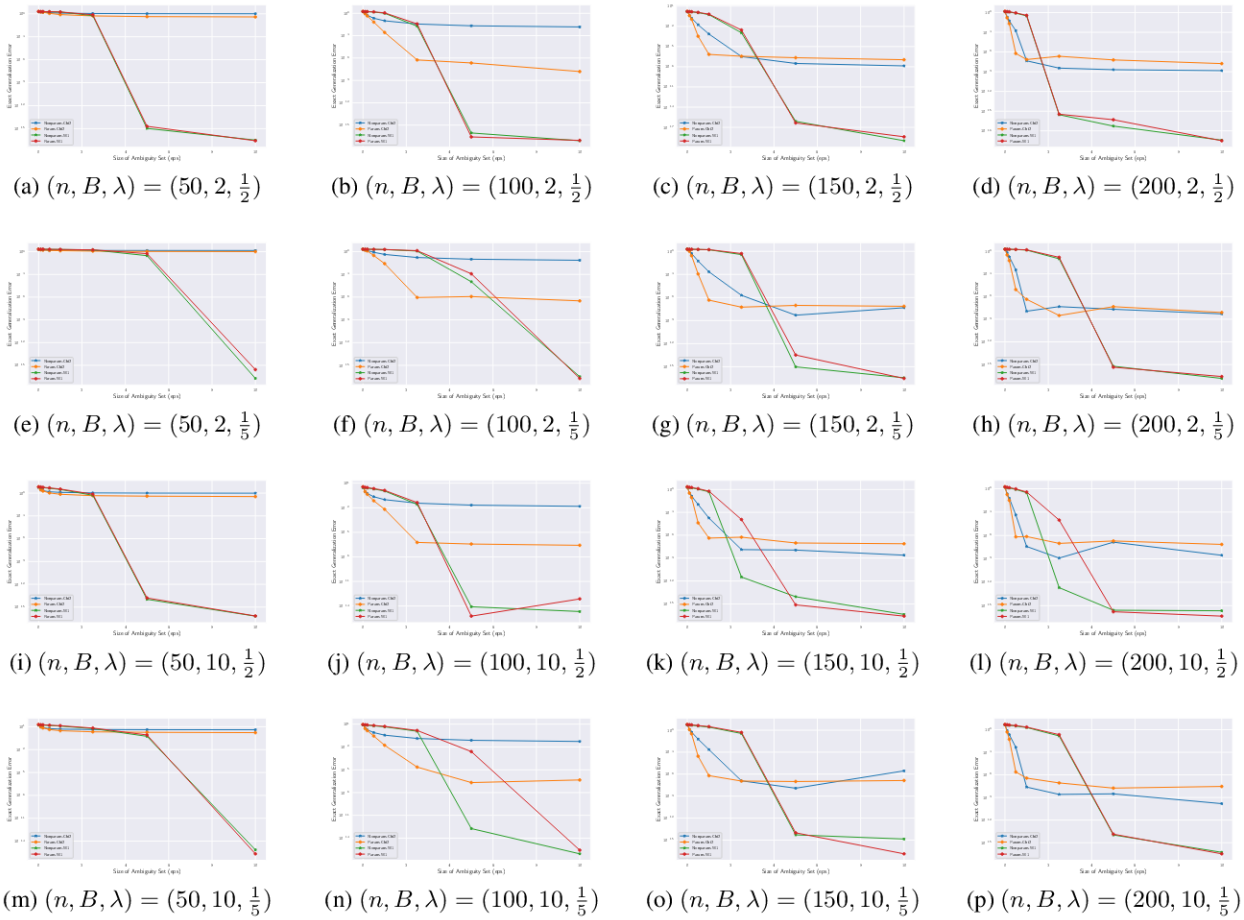
E.1 Detailed Setups and Results for another Synthetic Example

We conduct a small synthetic-data experiment with $\mathcal{V}_d(x^*) \approx 0$ similar to Section 5.2 in Duchi and Namkoong (2019).

To illustrate the model performance under this subcase, we follow an example of the quadratic cost function with linear perturbation in Section 5.2 in Duchi and Namkoong (2019): $h(x;\xi) = \frac{1}{2}\|x-v\|^2 + \xi^\top(x-v)$. We let $D_\xi = 50$ and the decision space $\mathcal{X} = \{x \in \mathbb{R}^{D_\xi} : \|x\|_2 \le B\}$ and set $v = \frac{B}{2\sqrt{D_\xi}}\mathbf{1}$ known beforehand. We use some misspecified distributions with $\mathcal{P}_\Theta$ but keep $\mathbb{E}_{\mathbb{P}^*}[(\xi)_i] = 0$ such that we have: $x^* = v, \forall\lambda$. Then we have $\mathcal{V}_d(x^*) = 0$. We illustrate through experiments and show that here our `P-DRO` model (fit with normal distribution) can also achieve zero error under large ambiguity size ε (like `NP-DRO` in Duchi and Namkoong (2019); Zeng and Lam (2022)), which outperforms the ERM loss no matter whether we use parametric or nonparametric models.

We take the marginal distribution of random variable $(\xi)_i = (\xi_\theta)_i + (\tilde{\xi}_i), \forall i$, where $\xi_\theta \sim \mathcal{N}(0,\Sigma)$, $(\tilde{\xi})_i \overset{d}{\sim} \text{Exp}(\lambda) - \frac{1}{\lambda}, \forall i \in [D_\xi]$ for each marginal, where $\lambda > 0$. Since the pdf of $(\tilde{\xi})_i$ is $f(x) = \lambda e^{-\lambda x}$, the smaller λ is, the larger difference it is compared to the normal ξ_θ. Under this case $\mathbb{E}_{\mathbb{P}^*}[(\xi)_i] = 0, \forall i \in [D_\xi]$.

We vary the decision boundary B from $\{2, 10\}$, noise ratio λ from $\{\frac{1}{5}, \frac{1}{2}\}$. Before each run, we independently generate Σ in the setup. For DRO methods, we apply χ^2-divergence and 1-Wasserstein distance. We choose the parametric class to be $\mathcal{P}_\Theta = \left\{\mathcal{N}(\mu,\Sigma) : \mu \in \mathbb{R}^{D_\xi}, \Sigma \in \mathbb{S}_{++}^{D_\xi}\right\}$ with unknown μ and Σ. Then, we have $\mathcal{E}_{apx}(\mathbb{P}^*, \Theta) > 0$. The misspecification



(a) $(n, B, \lambda) = (50, 2, \frac{1}{2})$ (b) $(n, B, \lambda) = (100, 2, \frac{1}{2})$ (c) $(n, B, \lambda) = (150, 2, \frac{1}{2})$ (d) $(n, B, \lambda) = (200, 2, \frac{1}{2})$

(e) $(n, B, \lambda) = (50, 2, \frac{1}{5})$ (f) $(n, B, \lambda) = (100, 2, \frac{1}{5})$ (g) $(n, B, \lambda) = (150, 2, \frac{1}{5})$ (h) $(n, B, \lambda) = (200, 2, \frac{1}{5})$

(i) $(n, B, \lambda) = (50, 10, \frac{1}{2})$ (j) $(n, B, \lambda) = (100, 10, \frac{1}{2})$ (k) $(n, B, \lambda) = (150, 10, \frac{1}{2})$ (l) $(n, B, \lambda) = (200, 10, \frac{1}{2})$

(m) $(n, B, \lambda) = (50, 10, \frac{1}{5})$ (n) $(n, B, \lambda) = (100, 10, \frac{1}{5})$ (o) $(n, B, \lambda) = (150, 10, \frac{1}{5})$ (p) $(n, B, \lambda) = (200, 10, \frac{1}{5})$

Figure E.1: Exact Generalization Error of DRO models varying sample size n, decision boundary B and noise ratio λ

effect reduces when λ grows larger. We fit the distribution with $\hat{\mathbb{Q}} \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \xi_i, \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\top$. We set each DRO model with ambiguity set size ranging in $\{0.1, 0.2, 0.5, 1, 2.5, 5, 10\}$ to show the trend.

Figure E.1 show different subcases varying (n, B, λ) across 50 independent runs . Across all these setups, since the optimal solution (i.e. $x^* = v$) does not depend on the decision boundary chosen here, increasing this decision boundary as well as the order dimension may not affect the decision quality too much. We have the following observations:

- Compared with the nonparametric version, `P-DRO` performs competitively against `NP-DRO` in all setups under the same ambiguity size ε. Both Wasserstein and χ^2-DRO can beat ERM models by a great extent under large ambiguity size, which can be explained from the error bound that replaces the term $\sup_{x\in\mathcal{X}} \mathcal{V}_d(x)$ with $\mathcal{V}_d(x^*)$;
- `P-ERM` and `NP-ERM` do not differ a lot. It can be directly computed that $x^{N-ERM} = v + \mathbb{E}_{\hat{\mathbb{P}}_n}[\xi]$ and $x^{P-ERM_m} = v + \mathbb{E}_{\hat{\mathbb{Q}}_m}[\xi]$ without restricting the decision space $\mathcal{X}$. Under finite Monte Carlo size m, `P-ERM` would introduce another Monte Carlo error. Therefore, `P-ERM` does not gain since `NP-ERM` only uses the mean of the random variable to predict the decision.

Therefore in this case, we illustrate that the `P-DRO` still inherits the property of `NP-DRO` in eliminating the dependency of the complexity term only to $\mathcal{V}_d(x^*)$.

Note that we can also consider the problem with distribution shift. However, as long as $\mathbb{E}^{te}[\xi] = 0$ such that $\mathcal{V}_d(x^*) = 0$, both DRO models would still incur almost zero generalization error due to the additional error $\mathcal{V}_d(x^*)d(\mathbb{P}^{te}, \mathbb{P}^{tr}) = 0$ under large ambiguity ε. For example, results in (a) - (d) and (e) - (h) of Figure E.1 can be regarded as two distributions but $\mathbb{E}^{tr}[h(x^{tr}; \xi)] = \mathbb{E}^{te}[h(x^*; \xi)]$. Therefore, if noise ratio λ in $\mathbb{P}^{tr}$ is $\frac{1}{5}$ but in $\mathbb{P}^{te}$ is $\frac{1}{2}$, the previous results and observations still hold.

In general, our results show that P−DRO outperforms P−ERM significantly and achieves almost zero generalization error under large fixed ambiguity size ε across 1-Wasserstein distance and χ^2-divergence, which is indicated by Theorem 1.

E.2 Detailed Setups and Analysis for Synthetic Example in the main text

The problem is to minimize the objective (10) in the main text. That is,

$$h(x;\xi) = \left|\min\{0, \xi^\top x - \mu\}\right|^\alpha = \left(\mu - \xi^\top x\right)_+^\alpha.$$

Regarding each marginal $(\xi)_i$, the base case is fully parametrized such that $(\xi)_i \overset{d}{=} 2r \times Beta(\alpha_i, 2) - r$ with $\{\alpha_i\}_{\{i\in[D_\xi]\}}$ i.i.d. drawn from $[1.5, 3]$.

E.2.1 Comparison between Comp$(\mathcal{H})$ and Comp(Θ)

We give concrete representations of Comp$(\mathcal{H})$ and Comp(Θ) in each method here. First, we present an upper bound below for the covering number of $\mathcal{H}$.

Lemma 10 (Theorem 5.4 in Matousek (1999)). *If $\mathcal{H}$ consists of polynomials up to degree D with d variables (e.g. each $h(\xi) \in \mathcal{H}, \xi = (\xi_1, \ldots, \xi_d)^\top \in \mathbb{R}^d$ can be represented as $h(\xi) = \sum_{i_1+\ldots+i_d\le D} a_i \xi_1^{i_1} \ldots \xi_d^{i_d}$, then we have:*

$$VC(\mathcal{H}) \le \binom{d+D}{d} \sim (d+D)^{\min\{d,D\}}.$$

And borrowing Theorem 2.6.7 in van der Vaart et al. (1996), we have the following results:

$$N(\mathcal{H}(\xi), \varepsilon, \|\cdot\|_\infty) \le \sup_{\mathbb{Q}} N(\mathcal{H}, \frac{\varepsilon}{2n}, \|\cdot\|_{L^1(\mathbb{Q})}) \le c\mathrm{VC}(\mathcal{H}) \left(\frac{16Mne}{\varepsilon}\right)^{\mathrm{VC}(\mathcal{H})-1},$$

for some numerical constants c.

Then combining it with Lemma 10, we have the following upper bound for the Covering number of $\mathcal{H}$ in (10):

$$N(\mathcal{H}(\xi), \varepsilon, \|\cdot\|_\infty) \le C(D_\xi + \alpha)^{\min\{D_\xi, \alpha\}} \left(\frac{nM}{\varepsilon}\right)^{(D_\xi+\alpha)^{\min\{D_\xi,\alpha\}}}, \tag{E.1}$$

where we denote $M := \sup_{x\in\mathcal{X}} \|h(x;\cdot)\|_\infty \le (D_\xi \tau r + \mu)^\alpha$ and denote $M^* = \sqrt{\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\cdot)]} \le \|h(x^*;\cdot)\|_\infty \le (r\|x^*\|_1 + \tau)^\alpha$. On the other hand, here we fix the true underlying distribution to be a variant of Beta distributions for simplicity, i.e. $\mathcal{E}_{apx} = 0$ if we just use the Beta distribution to fit our model, i.e. $\mathcal{P}_\Theta = \{\mathbb{P} : \xi = (\xi_1, \ldots, \xi_{D_\xi})^\top \sim \mathbb{P}, \xi_i \overset{d}{=} 2r \times Beta(\alpha_i, 2) - r\}$ and independent with other marginals. Therefore, we construct a set of bounded parametric distributions to allow for explicit bounds for different models. In these error bounds, we approximate the variance term appearing in χ^2-divergence by $\mathrm{Var}_{\mathbb{P}^*}[h(x^*;\cdot)] \le \|h(x^*;\cdot)\|_\infty^2$. Therefore, setting $\varepsilon = \frac{1}{n}$ in (E.1) and (D.7), and ignoring the term $\log(1/\delta)$ and $\log n$, the major dominating terms of the generalization errors in the four methods from existing results and our theorems (under χ^2-divergence for DRO methods) are:

Method	NP−ERM	P−ERM	NP−DRO	**P−DRO**
$\mathcal{E}(\hat{x})$	$M\sqrt{\frac{(D_\xi+\alpha)^\alpha \log M}{n}}$	$M\sqrt{\frac{D_\xi}{n}}$	$M^*\sqrt{\frac{(D_\xi+\alpha)^\alpha \log M}{n}}$	$M^*\sqrt{\frac{D_\xi}{n}}$

In our theorem, we can obtain the following results for Comp(Θ):

Proposition 1. *When d is χ^2-divergence, $\mathcal{P}_\Theta = \{\mathbb{P} : \xi = (\xi_1, \ldots, \xi_{D_\xi})^\top \sim \mathbb{P}, (\xi)_i \overset{d}{=} 2r \times Beta(\alpha_i, 2) - r, \alpha_i \in [k, 2k], \forall i \in [D_\xi]\}$ for some constant k and therefore $\mathbb{P}^* \in \mathcal{P}_\Theta$. Then Assumption 1 holds for $\hat{\mathbb{Q}}$ with $\hat{\alpha}_i$ computed from Maximum Likelihood estimation or Methods of Moments (will define the estimator formula next) and $\mathcal{E}_{apx} = 0$, Comp$(\Theta) = CD_\xi, \alpha = 1$ for some constant C when n is large.*

Then we show briefly the proof sketch under χ^2-divergence related to Assumption 1. We first give the formula of χ^2-divergence under Beta distribution below:

Example 7. *Generally, for* $\mathbb{P}_1 \sim Beta(\alpha_1, \beta_1), \mathbb{P}_2 \sim Beta(\alpha_2, \beta_2)$, *and* $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$, *we have:*

$$\chi^2(\mathbb{P}_1||\mathbb{P}_2) = \frac{B(\alpha_1, \beta_1)B(2\alpha_2 - \alpha_1, 2\beta_2 - \beta_1)}{B(\alpha_2, \beta_2)} - 1,$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$. *In order for the value to be meaningful, we can restrict the support of* $\alpha \in [k_1, 2k_1], \beta \in [k_2, 2k_2], k_1, k_2 > 0$. *Therefore, if true distribution for* $\xi \sim \prod_{i=1}^d \mathbb{P}_i^* (\overset{d}{=} Beta(\alpha_i, \beta_i))$ *and the estimated distribution* $\hat{\xi} \sim \prod_{i=1}^d \mathbb{Q}_i^* (\overset{d}{=} Beta(\hat{\alpha}_i, \hat{\beta}_i))$, *then by the product rule:*

$$\chi^2(\prod_{i=1}^d \mathbb{P}_i^*, \prod_{i=1}^d \hat{\mathbb{Q}}_i) = \prod_{i=1}^d \frac{B(\alpha_i, \beta_i)B(2\hat{\alpha}_i - \alpha_i, 2\hat{\beta}_i - \beta_i)}{(B(\hat{\alpha}_i, \hat{\beta}_i))^2} - 1.$$

Rescaling Beta distribution from $[0, 1]$ to the region $[-r, r]$ does not change the value of the f-divergence.

We then show briefly why MLE / Methods of Moments can help establish the parametric convergence rate. We give a sketch of proof to indicate that these distribution parametric estimators associated with model class $\mathcal{P}_\Theta$ under χ^2-divergence concretely satisfy Assumption 1. For simplicity, we fix $\beta_i = \hat{\beta}_i = 2$ in our problem. Then the divergence reduces to:

$$\chi^2(\prod_{i=1}^d \mathbb{P}_i^*, \prod_{i=1}^d \hat{\mathbb{Q}}_i) = \prod_{i=1}^d \frac{\hat{\alpha}_i}{\alpha_i} \cdot \frac{\hat{\alpha}_i + 1}{\alpha_i + 1} \cdot \frac{\hat{\alpha}_i}{2\hat{\alpha}_i - \alpha_i} \cdot \frac{\hat{\alpha}_i + 1}{2\hat{\alpha}_i - \alpha_i + 1} - 1. \tag{E.2}$$

The formula in (E.2) implies an estimation error such that $\forall i \in [d]$, with probability at least $1 - \delta$, we have:

$$1 - u\sqrt{\Delta} \le \frac{\hat{\alpha}_i}{\alpha_i} \le 1 + u\sqrt{\Delta}, \tag{E.3}$$

$$1 - v\sqrt{\Delta} \le \frac{\hat{\alpha}_i + 1}{\alpha_i + 1} \le 1 + v\sqrt{\Delta}, \tag{E.4}$$

where $\Delta := \frac{\log(1/\delta)}{n}$ in (E.3) and (E.4) and u, v are independent with the sample size n. If (E.3) and (E.4) holds, then:

$$\begin{aligned}
\chi^2(\prod_{i=1}^d \mathbb{P}_i^*, \prod_{i=1}^d \hat{\mathbb{Q}}_i) &\le \left((1 + u^2\Delta)(1 + v^2\Delta)\right)^d - 1 \\
&= \left[1 + 2(u^2 + v^2)\Delta + o(\Delta)\right]^d - 1 \\
&\le 4d(u^2 + v^2)\Delta + o(\Delta),
\end{aligned}$$

where the first inequality holds by $\frac{\hat{\alpha}_i}{\alpha_i} \cdot \frac{\hat{\alpha}_i}{2\hat{\alpha}_i - \alpha_i} = \frac{(\hat{\alpha}_i/\alpha_i)^2}{2(\hat{\alpha}_i/\alpha_i) - 1} \le 1 + \frac{(u\sqrt{\Delta})^2}{1 + 2u\sqrt{\Delta}} \le 1 + (u\sqrt{\Delta})^2$ (as long as $u\sqrt{\Delta} \le \frac{1}{2}$). And the second inequality holds when n is large. Thus we show that $\chi^2(\prod_{i=1}^d \mathbb{P}_i^*, \prod_{i=1}^d \hat{\mathbb{Q}}_i) = O(\frac{D_\xi}{n})$.

For Methods of Moments, we consider the first-order estimation $\mathbb{E}[\frac{\tilde{\xi}_i}{2r} + \frac{1}{2}] = \frac{\alpha}{\alpha+2}$ to obtain $\hat{\alpha}_i = \frac{2}{\frac{1}{2} - \frac{\sum_i \xi_i}{2nr}} - 2 := \frac{2\hat{\mathbb{E}}[\gamma_i]}{1 - \hat{\mathbb{E}}[\gamma_i]}$ if we define $\hat{\gamma}_i = \frac{\hat{\xi}_i + r}{2r}$. Then we have:

$$\frac{\hat{\alpha}_i}{\alpha_i} = \frac{\hat{\mathbb{E}}[\gamma_i]}{\mathbb{E}[\gamma_i]} \cdot \frac{1 - \mathbb{E}[\gamma_i]}{1 - \hat{\mathbb{E}}[\gamma_i]},$$

which can be directly bounded by a Hoeffding-type concentration inequality to bound $|\hat{\mathbb{E}}[\gamma_i] - \mathbb{E}[\gamma_i]|$, which is the same case as for $\frac{\hat{\alpha}_i + 1}{\alpha_i + 1}$.

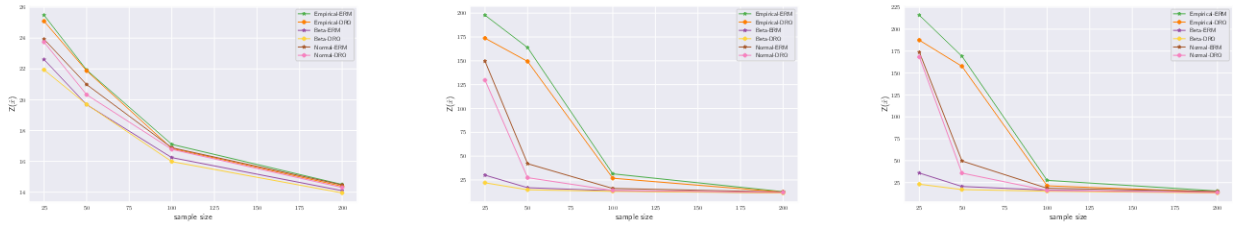
For Maximum Likelihood Estimators, we estimate the parameter of each marginal with:

$$\hat{\alpha}_i = \max\left\{\min\left\{\frac{-a_n - 2 - \sqrt{a_n^2 + 4}}{2a_n}, 3\right\}, 1.5\right\}, \text{ where } a_n = \frac{\sum_{i=1}^n \ln \hat{\gamma}_i}{n}.$$

Focusing on the part $\frac{-a_n - 2 - \sqrt{a_n^2 + 4}}{2a_n}$ and replacing this with α, we would obtain:

$$\frac{\hat{\alpha}}{\alpha} = \frac{a_n}{\hat{a}_n}\left(1 + \frac{\hat{a}_n - a_n + \sqrt{\hat{a}_n^2 + 4} - \sqrt{a_n^2 + 4}}{a_n + 2 + \sqrt{a_n^2 + 4}}\right).$$

Then we can apply a concentration inequality to bound $|a_n - \hat{a}_n|$ since log Beta distribution $\ln \gamma$, $\gamma \in [0, 1]$ is subexponential. □



(a) Fully Parametrized (b) Distribution Mis-specification (c) Distribution Shift

Figure E.2: Value of Cost function across different ERM-DRO models varying sample size n with $(\tau, \alpha) = (2, 2)$.

E.2.2 Detailed Experimental Results

We also fit the model with normal class $\mathcal{P}_\Theta$, i.e. the parametric model in Section E.1 in this case. Intuitively, for χ^2-divergence, $\text{Comp}(\Theta) = C(D_\xi)^2$ with $\mathcal{E}_{apx} > 0$. Even incurring distribution misspecification, Normal-DRO models can still outperform the empirical version.

In general, we observe that the generalization error trend in terms of n is consistent with under different setups like Figure 1 in the main text in this simulation under $(\alpha, \tau) = (2, 2)$, In Figure E.2, under each DRO method, we tune the best hyperparameter $\varepsilon \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$ and results are averaged over 50 independent runs. For the base case (a) without distribution misspecification, Beta-DRO performs significantly best especially under small size. Although the performance gap between `P-ERM` and `P-DRO` are small under large sample size, the difference in values of the cost function is still statistical significant with $p < 0.001$, which is also the case in $(b)(c)$ of Figure E.2 corresponding to the distribution misspecification and distribution shift case.

Detailedly, we demonstrate the results in the three different cases to further illustrate the effects of the ambiguity size and complexity term under fix ambiguity size ε as follows:

(1) Fully Parametrized Case, shown in Figure E.3. We have two major observations: (1) When (α, τ) is large, since parametric approaches does not depend on the overall complexity term, and then the distribution complexity is dominated by the function complexity. `P-DRO` can enjoy relative better performance, especially under small samples. When (α, τ) is small, the gaps would be small. (2) When we restrict the models to DRO models, when the ambiguity size increases from 0 to ∞, the generalization error of `P-DRO` would first decrease and then increase. When the sample size increases, the turning point decreases to 0, which consistent with the trend of "best" ambiguity size (cover $\mathbb{P}^*$ w.h.p. for $\varepsilon_n = \Delta(\delta, \Theta)$) $\varepsilon_n \to 0$ as $n \to \infty$.

(2) Distribution Mis-specification Based on the previous model, we perturb each marginal of the random variable ξ_i to $\xi_i + \zeta_i$, where each $\zeta_i \sim U(-2, 2)$ and independent with ξ_i. The results are shown in Figure E.4 with more noticable performance advantages for `P-DRO` models.

(3) Distribution Shift Here, we randomly generate one shift parameter $C \in [-1, 1]$ and the train distribution satisfies: $\xi_i^{tr} \overset{d}{=} 2r \times Beta(\alpha_{i,1}, \beta_i) - r$. And the test distribution satisfies: $\xi_i^{te} \overset{d}{=} 2r \times Beta(\alpha_{i,2}, \beta_i) - r$, where

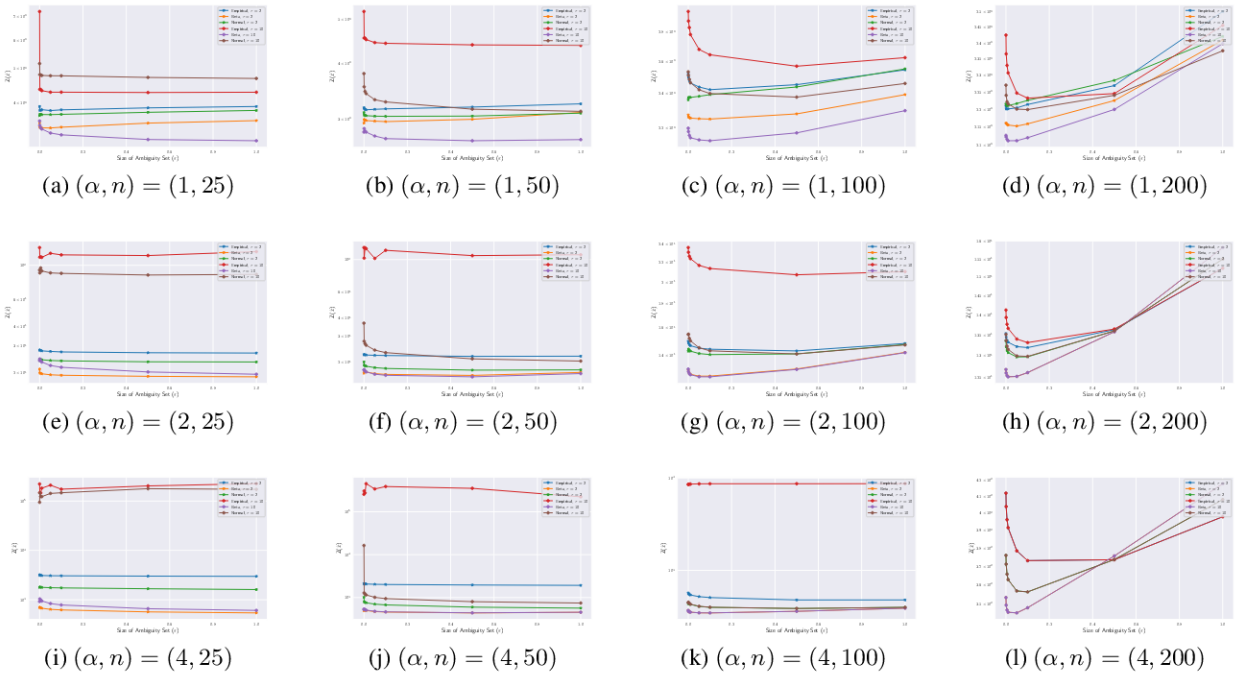
$$\alpha_{i,2} = \alpha_{i,1} + C \min\{\alpha_U - \alpha_{i,1}, \alpha_{i,1} - \alpha_L\}.$$

We also perturb an uniform noise $\zeta_i \sim U(-2, 2)$ to the test distribution as before. On the distribution shift case, if we have distribution shifts, then `P-DRO` models would have better performance than the ERM counterpart. The results are shown in Figure E.5.

In general, the simulation results in Section E.1 and E.2 are almost consistant with our model theoretical analysis (e.g. Table 1 in the main text).

E.3 Detailed Setup for Real-world Portfolio Experiments

To obtain and evaluate the out-of-sample performance, we apply the **"rolling-sample"** approach following DeMiguel et al. (2007) on the monthly data from July 1963 to December 2018 ($T = 666$) with the estimation window size to be $M = 60$



(a) $(\alpha, n) = (1, 25)$ (b) $(\alpha, n) = (1, 50)$ (c) $(\alpha, n) = (1, 100)$ (d) $(\alpha, n) = (1, 200)$

(e) $(\alpha, n) = (2, 25)$ (f) $(\alpha, n) = (2, 50)$ (g) $(\alpha, n) = (2, 100)$ (h) $(\alpha, n) = (2, 200)$

(i) $(\alpha, n) = (4, 25)$ (j) $(\alpha, n) = (4, 50)$ (k) $(\alpha, n) = (4, 100)$ (l) $(\alpha, n) = (4, 200)$

Figure E.3: Value of Cost function across different ERM-DRO models varying sample size n and α.

months[5]. We report the experimental results of different models with the empirical performance $\hat{h} := \frac{1}{T-M}\sum_{i=1}^{T-M}(\mu - \hat{r}_i)_+^2$ given $T - M$ out-of-sample returns.

We use χ^2-divergence in our DRO models with cross validation $\varepsilon = \{0.2, 0.4, \ldots, 1.6, 1.8\}$ in each period. And we fit the observed samples with (1) normal families (the same as in Section E.1 and E.2); (2) variants of beta distributions, where we still fix $\beta_j = 2$, α_j using formula in Section E.2 and choose the boundary parameter $r_j = \max|(\xi)_j|$ for each asset j.

E.4 Detailed Setup for Regression Tasks

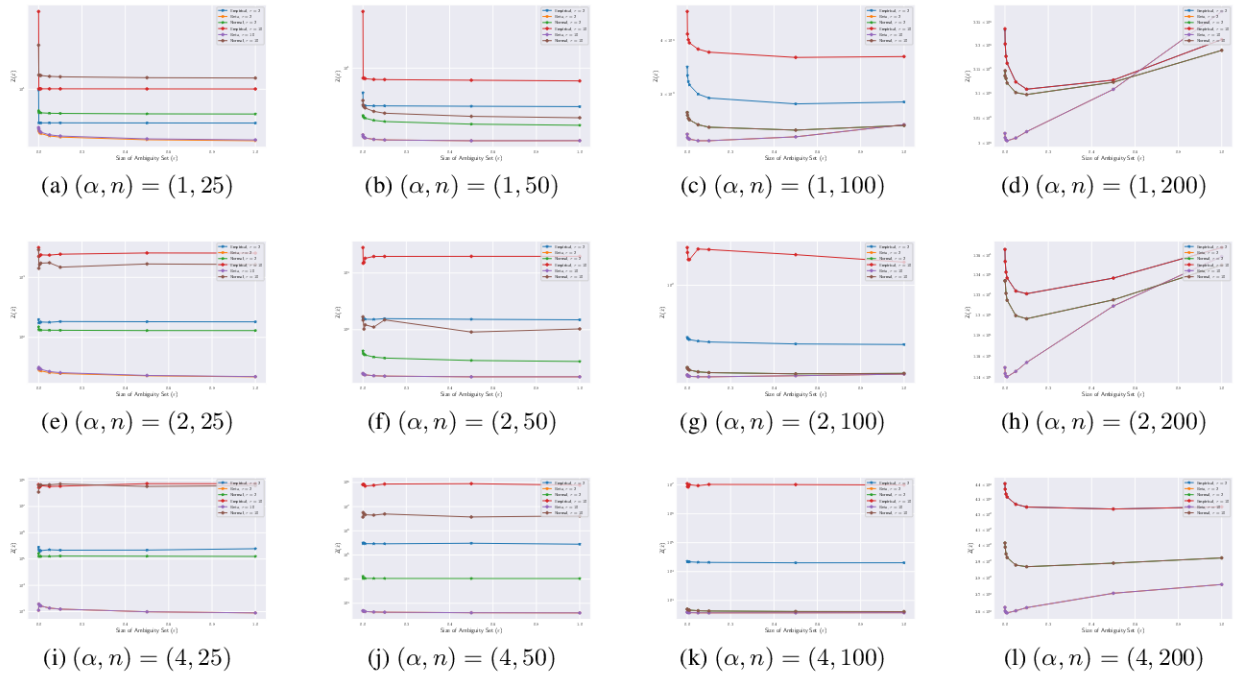
We vary the ambiguity size $\varepsilon \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$ as the hyperparameter candidate set and tune them through cross validation for DRO models. We set the Monte Carlo size $M = 10n$ and $\hat{\mathbb{Q}}$ is constructed as follows:

Mixture Gaussian Model Construction: If we denote the features $[x_1, \ldots, x_d] \in \mathbb{R}^d$ and $x_1, \ldots, x_K$ are K binary category features with $x_i \in \{0, 1\}, \forall i \in [K]$, then we consider the following mixture Gaussian Distribution Class with 2^K groups[6]

$$\begin{aligned}\mathcal{P}_\Theta = \Big\{\mathbb{P} : (x_1, \ldots, x_d, y)^\top \sim \mathbb{P} : \mathbb{P}(\overline{x_1 x_2 \ldots x_K} = s - 1) = p_s, \forall s \in [2^K],\\ (x_{K+1}, \ldots, x_d, y)|(x_1, \ldots, x_K) \overset{d}{=} \mathcal{N}(\mu_k, \Sigma_k)\\ \Big|(p_1, \ldots, p_{2^K}) \in \Delta_{2^K}, \mu_k \in \mathbb{R}^{d-K+1}, \Sigma \in \mathbb{S}_{++}^{d-K+1}, \forall k \in [2^K]\Big\}.\end{aligned}$$

[5]Rolling-sample: For data spanning T months, in order to construct portfolios in month $t+M$ (from $t = 1$), we use the data spanning from months t to $t + M - 1$ as observed samples solve the corresponding problem. Then we apply the optimal weight $\hat{x}$ obtained from $\min_{x\in\mathcal{X}} \hat{Z}(x)$ to compute the returns $\hat{r}_t = \xi_{t+M}^\top \hat{x}$ in month $t+M$. We repeat this procedure to construct portfolios in following months by adding the next and dropping the earliest month until $t = T - M$. This gives us $T - M$ monthly out-of-sample returns $\{\hat{r}_i\}_{i=1}^{T-M}$.

[6]$\overline{x_1 x_2 \ldots x_k}$ represents the decimal number of these binary digits x_i, e.g. $\overline{101} = 5$; Δ_n means the n-dimension probability simplex.



(a) $(\alpha, n) = (1, 25)$ (b) $(\alpha, n) = (1, 50)$ (c) $(\alpha, n) = (1, 100)$ (d) $(\alpha, n) = (1, 200)$

(e) $(\alpha, n) = (2, 25)$ (f) $(\alpha, n) = (2, 50)$ (g) $(\alpha, n) = (2, 100)$ (h) $(\alpha, n) = (2, 200)$

(i) $(\alpha, n) = (4, 25)$ (j) $(\alpha, n) = (4, 50)$ (k) $(\alpha, n) = (4, 100)$ (l) $(\alpha, n) = (4, 200)$

Figure E.4: Value of Cost function across different ERM-DRO models varying sample size n and α under misspecification.

For a dataset with $\{(\hat{\boldsymbol{x}}_i, \hat{y}_i)\}_{i\in[n]}$ with $\hat{x}_{i,j}$ denoting the j-th feature of the i-th sample. We output $\hat{\mathbb{Q}}$ parametrized with $\{(\hat{p}_k, \hat{\mu}_k, \hat{\Sigma}_k)\}_{k=1}^{2^K}$:

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\overline{\hat{x}_{i,1}\ldots\hat{x}_{i,K}}=k-1\}}$$

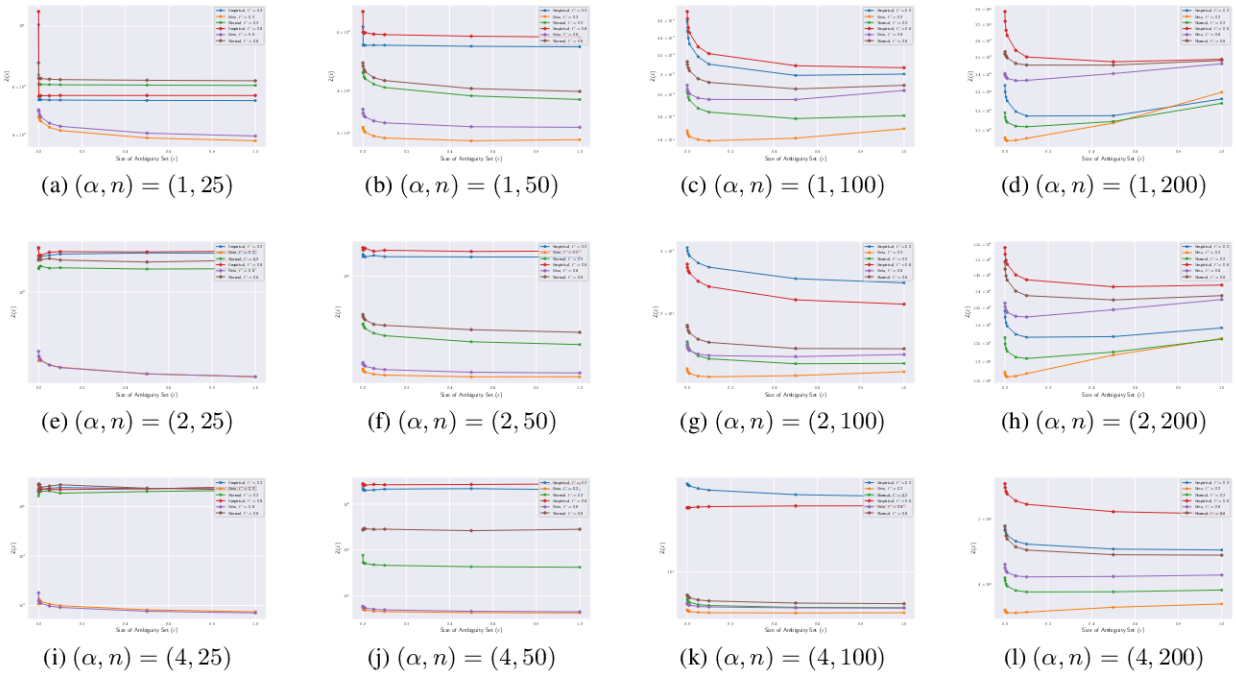
$$\hat{\mu}_k = \frac{1}{n\hat{p}_k}\sum_{i=1}^{n} (\hat{x}_{i,K+1}\ldots\hat{x}_{i,d}\hat{y}_i)\mathbb{I}_{\{\overline{\hat{x}_{i,1}\ldots\hat{x}_{i,K}}=k-1\}}$$

$$\hat{\Sigma}_k = \frac{1}{n\hat{p}_k}\sum_{i=1}^{n} [(\hat{x}_{i,K+1}\ldots\hat{x}_{i,d}\hat{y}_i) - \hat{\mu}_k][(\hat{x}_{i,K+1}\ldots\hat{x}_{i,d}, \hat{y}_i) - \hat{\mu}_k]^\top \mathbb{I}_{\{\overline{x_{i,1}\ldots x_{i,K}}=k-1\}}$$

Conditioned on the subgroups characterized by $K = 4$ binary categorical features `black`, `hispanic`, `married`, `nodegree`, we fit the other continuous variables `age`, `education`, `RE74`, `RE75`, `RE78` with Gaussian models constructed above. After that, for our Monte Carlo sampling, if $\hat{p}_k$ is very small, i.e. not many samples in this group, we would directly use the original data within that group and copy 10 times as the new data. We project the values of some unrealistic features of the Monte Carlo data from $\hat{\mathbb{Q}}$ onto their value boundary if these values violate some common knowledge. For example, if earnings one year is negative, we change it to 0. And we project the value of simulated ages into the interval $[18, 60]$. Then our Monte Carlo empirical distribution $\hat{\mathbb{Q}}_m$ is a real crude approximation only assuming the underlying true distribution has some approximate normal properties, which can be hardly attained in reality. In spite of this, the results of `P-DRO` in the main text performs well and significantly better than others especially when the sample size is not large enough, shown in the main text.

Furthermore, to illustrate that `P-DRO` can eliminate the model misspecification error and show consistenly good performance than `P-ERM`and the empirical models, we replace (A) mixture Gaussian in $\mathcal{P}_\Theta$ to (B) joint Gaussian of all variables; (C) joint Gaussian fixing categorical variables zero correlation following the same setup. All models still have $\mathcal{E}_{apx} > 0$ but the robustness and superior performance indicate the effectiveness of DRO methods.

$n = 200$	`NP-ERM`	`NP-DRO`	`P-ERM`-(A)	`P-DRO`-(A)	`P-ERM`-(B)	`P-DRO`-(B)	`P-ERM`-(C)	`P-DRO`-(C)
Avg-R^2	0.1589	0.4433	< 0	0.5050	< 0	0.5266	0.4926	0.5033



(a) $(\alpha, n) = (1, 25)$ (b) $(\alpha, n) = (1, 50)$ (c) $(\alpha, n) = (1, 100)$ (d) $(\alpha, n) = (1, 200)$

(e) $(\alpha, n) = (2, 25)$ (f) $(\alpha, n) = (2, 50)$ (g) $(\alpha, n) = (2, 100)$ (h) $(\alpha, n) = (2, 200)$

(i) $(\alpha, n) = (4, 25)$ (j) $(\alpha, n) = (4, 50)$ (k) $(\alpha, n) = (4, 100)$ (l) $(\alpha, n) = (4, 200)$

Figure E.5: Value of Cost function across different ERM-DRO models varying sample size n and α under distribution shift.