Rademacher Complexity for Distributionally Robust Learning

Zhengyu Zhou, Weiwei Liu*

School of Computer Science,
Wuhan University, China
{zzysince1999, liuweiwei863}@gmail.com

Abstract

The goal of distributionally robust learning is to learn models capable of performing well against distributional shifts, such as latent heterogeneous subpopulations, unknown covariate shifts, or unmodeled temporal effects. Recently, Duchi and Namkoong (2021) have proven an upper bound for the excess risk of distributionally robust learning through the lens of covering number argument. However, there are situations where the covering argument fails. This motivates us to study the generalization bound through the lens of Rademacher complexity. More specifically, we consider the Cressie-Read divergence (Cressie and Read 1984), $f_k(t) \propto t^k - 1$. Our theoretical results indicate that the excess risk is of the order $O_P(n^{-\frac{1}{2k_*}})$, where $k_* = \frac{k}{k-1}$. The decay rate of the excess risk increases with increasing k. As illustrative examples, we consider three learning settings: 1) linear classifier; 2) Gaussian reproducing kernel Hilbert space; 3) one-hidden-layer networks. The empirical results validate our theoretical findings.

1 Introduction

In safety- and fairness- critical systems (Knight 2002), the goal is to learn machine learning models that achieve uniformly good performance over all input values. Examples include medical diagnosis, autonomous vehicles, criminal justice and credit evaluations, where poor performance on the tails of the inputs leads to high-cost system failures. By contrast, methods that optimize average performance, often produce models that suffer low performance on the "hard" instances of the population. For example, recent works (Blodgett, Green, and O'Connor 2016; Hashimoto et al. 2018) have demonstrated that models with low average error still fail on particular groups of data points. In light of this, various approaches have attempted to reduce the worst-group training loss through Distributionally Robust Learning (DRL) or by simply upweighting the minority groups.

In addition to latent heterogeneity in the population, distributional shifts in covariates (Shimodaira 2000; Ben-David et al. 2006) or unobserved confounding variables (Hand 2006) can contribute to changes in the data generating distribution. The performance of machine learning models degrades significantly on domains that are different from what the model is trained on (Hand 2006; Blitzer, McDonald, and Pereira 2006; Saenko et al. 2010; Torralba and Efros 2011).

To mitigate these challenges, some works have developed a DRL framework that is explicitly robust to local changes in the data-generating distribution. Concretely, let $\mathcal{Z}$ be the *instance space*, P be the data-generating distribution on the instance space $\mathcal{Z}$, Z be a random element of $\mathcal{Z}$, and $\mathcal{H}$ be a class of measurable functions $h : \mathcal{Z} \mapsto \mathbb{R}_+$, each of which quantifies the loss of a certain decision rule applied to instances $z \in \mathcal{Z}$. So, with a slight abuse of terminology, we will refer $\mathcal{H}$ as the *hypothesis class*. Rather than minimizing the average loss $\mathbb{E}_P[h(Z)]$, we study the *distributionally robust* problem:

$$\min_{h\in\mathcal{H}} \left\{ R_f(P,h) \coloneqq \sup_{Q\ll P} \{\mathbb{E}_Q[h(Z)] : D_f(Q\|P) \le \rho\} \right\}, \tag{1}$$

where Q is a distribution over $\mathcal{Z}$, $Q \ll P$ indicates that Q is absolutely continuous with respect to P, and the hyperparameter $\rho > 0$ modulates the distributional shift. Here,

$$D_f(Q\|P) \coloneqq \int f\left(\frac{dQ}{dP}\right) dP$$

is the f-divergence between Q and P, where $f : \mathbb{R} \mapsto \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ is a convex function satisfying $f(1) = 0$ and $f(t) = +\infty$ for any $t < 0$.

The worst-case risk (1) upweights the regions of $\mathcal{Z}$ with high losses $h(z)$; thus (1) optimizes the tail performance, which is measured by the loss of "hard" examples. As long as the alternative distribution Q remains ρ-close to the data-generating distribution P, the hypothesis $h^* \in \mathcal{H}$ that minimizes (1) evidently guarantees that $\mathbb{E}_Q[h^*(Z)] \le R_f(P, h^*)$ and provides the smallest such bound, which is equivalent to controlling the tail-performance under P. Let P_n denote the empirical distribution on $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} P$; accordingly, we minimize the following empirical counterpart of (1):

$$\hat{h} \in \operatorname*{argmin}_{h\in\mathcal{H}} \{R_f(P_n,h) \coloneqq \sup_{Q\ll P_n} \{\mathbb{E}_Q[h(Z)] : D_f(Q\|P_n) \le \rho\}\Big\}. \tag{2}$$

Duchi and Namkoong (2021) have addressed the generalization problem in distributionally robust setting based on the

*Corresponding author.

covering number with respect to L_∞-norm. However, there are situations where the covering number argument fails, e.g., 1) large model classes for whom the covering numbers are prohibitively large; 2) finite VC classes for whom the covering numbers with respect to L_∞-norm are infinite (e.g., class of intervals in $\mathbb{R}$).

A scenario where covering number argument fails Consider the binary classification problem, where the input space is $\mathcal{X} = \mathbb{R}$ and the label space is $\mathcal{Y} = \{0, 1\}$. We choose the hypothesis class as $\mathcal{H} = \{\mathbb{I}_{[a,b]}(\cdot) : a < b\}$. We define the loss function as $\ell(h, (x, y)) \coloneqq \mathbb{I}[h(x) \neq y]$ for $h \in \mathcal{H}$. We also consider the loss function class $\mathcal{L} = \{\mathbb{I}[h(x) \neq y] : h \in \mathcal{H}\}$. Duchi and Namkoong (2021) use covering number of the loss function class with respect to L^∞-norm to provide learning guarantees. We next show the covering number of $\mathcal{L}$ with respect to L^∞-norm is infinite. Given two different hypothese h_1 and h_2, there exists x, such that $h_1(x) \neq h_2(x)$. Therefore, we have $\sup_{x,y} |\mathbb{I}[h_1(x) \neq y] - \mathbb{I}[h_2(x) \neq y]| = 1$. The L^∞-distance between two functions implies the δ-packing number of $\mathcal{L}$ is infinite for any $\delta < 1/2$. The relation between the covering number and the packing number (Wainwright 2019, Lemma 5.5) implies that the $\delta/2$-covering number of $\mathcal{L}$ with respect to L^∞-norm is infinite, for any $\delta < 1$.

In this paper, we try to fill this gap by providing learning guarantees through the lens of Rademacher complexity. Our contributions can be summarized as follows:

- We solve the distributionally robust generalization problem based on f-divergence using the Rademacher complexity and derive a data-dependent upper bound on generalization error.
- We provide a general analysis of the excess risk of locally minimax ERM via the Rademacher complexity procedure under f-divergence, which can provide non-vacuous bound in situations where the covering number argument fails. Our generalization error bounds exhibit dependence on the choice of f_k.
- As illustrative examples, we derive the excess risk bound for linear classifier, Gaussian reproducing kernel Hilbert space and one-hidden-layer neural networks. Furthermore, the experimental results show that l_2 regularization is able to reduce the distributionally robust generalization error, while the generalization gap increases with increasing dimension of the feature space, which validate our theoretical results.

2 Preliminaries

The *uncertainty region* is defined as follows:

$$\mathcal{U}_P \coloneqq \{Q : D_f(Q\|P) \leq \rho\},$$

The likelihood ratio $L(x) \coloneqq dQ(x)/dP(x)$ can be used to reformulate our distributionally robust problem (1) as follows:

$$\begin{aligned} R_f(P, h) &= \sup\{\mathbb{E}_Q[h(Z)] : Q \in \mathcal{U}_P\} \\ &= \sup_{L \geq 0} \{\mathbb{E}_P[L(Z)h(Z)] : \mathbb{E}_P[L(Z)] = 1, \\ &\quad \mathbb{E}_P[f(L(Z))] \leq \rho\}, \end{aligned} \tag{3}$$

where the supremum is over measurable functions. Let $Z_1, \ldots, Z_n$ be an n-tuple of independent and identically distributed (i.i.d.) training examples drawn from P, while P_n is the empirical distribution on $Z_1, \ldots, Z_n$. In the next section, we analyze the performance of the *local minimax ERM*:

$$\hat{h} \coloneqq \operatorname*{argmin}_{h \in \mathcal{H}} R_f(P_n, h). \tag{4}$$

The *generalization error* denoted by $R_f(P, \hat{h})$ and the *excess risk* denoted by $R_f(P, \hat{h}) - \inf_{h' \in \mathcal{H}} R_f(P, h')$ are the key quantities to measure the generalization ability of distributionally robust learning. To study these two quantities, we define the *empirical Rademacher complexity* of $\mathcal{H}$ for a given sample $Z_1, \ldots, Z_n$ as follows:

$$\hat{\mathcal{R}}_n(\mathcal{H}) \coloneqq \mathbb{E}_\varepsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(Z_i) \right],$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of i.i.d. Rademacher variables. The *Rademacher complexity* of $\mathcal{H}$, $\mathcal{R}_n(\mathcal{H})$ is defined as the expectation of this quantity: $\mathcal{R}_n(\mathcal{H}) \coloneqq \mathbb{E}_{(Z_1,\ldots,Z_n)\sim P^n} \left[\hat{\mathcal{R}}_n(\mathcal{H}) \right]$.

Moreover, $f^*(s) \coloneqq \sup_t \{st - f(t)\}$ is denoted as the Fenchel conjugate, and the dual reformulation of (3) is provided by Shapiro (2017).

Lemma 1. *For any probability P on $\mathcal{Z}$, $k \in (1, \infty)$, $k_* = k/(k-1)$, any $\rho > 0$, and $c_k(\rho) \coloneqq (1 + k(k-1)\rho)^{\frac{1}{k}}$, for all $h \in \mathcal{H}$, we have*

$$R_f(P, h) = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_P \left[\lambda f^* \left(\frac{h(Z) - \eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\}. \tag{5}$$

Divergence families: Much of our work centers on two families of divergence. The *Rényi α-divergence* (van Erven and Harremoës 2014) between distribution P and Q is as follows:

$$D_\alpha(P\|Q) \coloneqq \frac{1}{\alpha - 1} log \int \left(\frac{dP}{dQ} \right)^\alpha dQ, \tag{6}$$

where the limit satisfies $D_1(P\|Q) = D_{kl}(P\|Q) \coloneqq \int \log \left(\frac{dP}{dQ} \right) dP$ as α goes to 1. For analytical reasons, we also use the equivalent Cressie-Read family of f-divergence (Cressie and Read 1984):

$$\begin{aligned} f_k(t) &\coloneqq \frac{t^k - kt + k - 1}{k(k-1)}, \\ \text{so } f_k^*(s) &\coloneqq \frac{1}{k} \left[((k-1)s + 1)_+^{k_*} - 1 \right], \end{aligned} \tag{7}$$

where $k \in (1, +\infty)$, $k_* = \frac{k}{k-1}$ and $t_+ = \max\{0, t\}$. Let $f_k(t) = +\infty$ for $t < 0$; we further define f_1 and f_0 as their respective limit as k goes to 0 and 1. The family of divergence (7) includes χ^2-divergence ($k = 2$), empirical likelihood $f_0(t) = -\log t + t - 1$, and KL-divergence $f_1(t) = t \log t - t + 1$, and we frequently employ the following shorthand:

$$R_k(P, h) \coloneqq \sup_{Q \ll P} \{\mathbb{E}_Q[h(Z)] : D_{f_k}(Q\|P) \leq \rho\}. \tag{8}$$

For ease of exposition, we here focus on $k \in (1, \infty)$. By minimizing (5) over $\lambda \geq 0$, we obtain a simplified dual formulation for the Cressie-Read family (7), which is used to protect against worst-case distributional shifts.

Lemma 2. *For any probability P on $\mathcal{Z}$,$k \in (1, \infty)$, $k_* = k/(k-1)$, any $\rho > 0$, and $c_k(\rho) \coloneqq (1 + k(k-1)\rho)^{\frac{1}{k}}$, for all $h \in \mathcal{H}$, we have:*

$$R_k(P, h) = \inf_{\eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(h(Z) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\}. \quad (9)$$

Proof of Lemma 2. Invoking Lemma 25 in the Appendix, we have the Fenchel conjugate of f_k:

$$f_k^*(s) = \frac{1}{k}((k-1)s + 1)_+^{k_*} - \frac{1}{k}$$

Substituting this into the dual formulation (5), we arrive at

$$\begin{aligned}
&\sup_{Q \ll P} \{\mathbb{E}_Q[Z] : D_{f_k}(Q\|P) \leq \rho\} \\
&= \inf_{\lambda \geq 0, \eta} \left\{ \lambda \mathbb{E}_P \left[f_k^* \left(\frac{Z - \eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\} \\
&= \inf_{\lambda \geq 0, \eta} \left\{ \frac{(k-1)^{k_*}}{k} \lambda^{1-k_*} \mathbb{E}_P \left[\left(Z - \eta + \frac{\lambda}{k-1} \right)_+^{k_*} \right] \right. \\
&\left. + \lambda(\rho - \frac{1}{k}) + \eta \right\} \\
&= \inf_{\lambda \geq 0, \tilde{\eta}} \left\{ (k-1)^{k_*} k^{-1} \mathbb{E}_P \left[(Z - \tilde{\eta})_+^{k_*} \right] \lambda^{1-k_*} \right. \\
&\left. + \left(\rho + \frac{1}{k(k-1)} \right) \lambda + \tilde{\eta} \right\},
\end{aligned}$$

where the last line is followed by setting $\tilde{\eta} \coloneqq \eta - \frac{\lambda}{k-1}$. Taking derivative of λ to minimize the preceding expression, we have (noting that $(k_* - 1)/k_* = 1/k$):

$$\lambda = (k-1)(k(k-1)\rho + 1)^{-\frac{1}{k_*}} \left(\mathbb{E}_P \left[(Z - \tilde{\eta})_+^{k_*} \right] \right)^{\frac{1}{k_*}}.$$

By substituting it into the preceding expression, we find that the supremum is

$$\inf_{\tilde{\eta}} (k(k-1)\rho + 1)^{\frac{1}{k}} \left(\mathbb{E}_P \left[(Z - \tilde{\eta})_+^{k_*} \right] \right)^{1/k_*} + \tilde{\eta},$$

which concludes our proof. □

To illustrate that the simplified dual form (9) is equivalent to optimizing the tail-performance of a model, we introduce a risk-averse loss, namely conditional value-at-risk (CVaR) (Rockafellar, Uryas'ev et al. 2000). Krokhmal (2007) shows that the dual form (9) is a higher-order generalization of the classical CVaR (Rockafellar, Uryas'ev et al. 2000), which corresponds to $R_k(P, h)$ with $k = \infty$ (or k_*=1).

Case k $= \infty$.

Example 1. For $0 < \alpha \leq 1$, the conditional value-at-risk is

$$\text{CVaR}_\alpha(P, h) \coloneqq \inf_{\eta \in \mathbb{R}} \left\{ \alpha^{-1} \mathbb{E}_P \left[(h(Z) - \eta)_+ \right] + \eta \right\}. \quad (10)$$

(10) corresponds to an uncertainty set arising out of limiting $f-$divergence or Rényi divergence. Recalling the Rényi divergence (6), we have $D_\infty(Q\|P) \coloneqq \lim_{\alpha \to \infty} D_\alpha(P\|Q) \coloneqq \lim_{\alpha \to \infty} \frac{1}{\alpha - 1} log \int \left(\frac{dP}{dQ} \right)^\alpha dQ = \text{ess sup} \log \frac{dQ}{dP}$. If we define $f_{\infty,c} = 0$ for $0 \leq t \leq c$ and $+\infty$ otherwise, then the uncertainty region can be expressed via the following calculation (Shapiro, Dentcheva, and Ruszczynski 2014, Example 6.19):

$$\begin{aligned}
\mathcal{U}_P &\coloneqq \left\{ Q : D_\infty(Q\|P) \leq \log \frac{1}{\alpha} \right\} \\
&= \left\{ Q : D_{f_{\infty,\alpha^{-1}}}(Q\|P) \leq 1 \right\} \\
&= \{ Q : \text{there exists } Q', \beta \in [\alpha, 1], \\
&\quad \text{s.t. } P = \beta Q + (1 - \beta) Q' \}.
\end{aligned}$$

It can be readily observed that the optimal η of (10) is the α-quantile of $h(Z)$, which is defined as below.

$$q(\alpha) \coloneqq \inf_q \{ P[h(Z) > q] \leq \alpha \}.$$

The dual form (10) shows that CVaR minimizes the expected risk on the worst α portion of the training data.

To study the relation between the distributionally robust risk $R_k(P, h)$ and the expected loss of hypothesis h, we consider the case of $k = 2$.

Case $k = 2$. We derive a bound of the distributionally robust loss in terms of the expected loss and variance of hypothesis h for χ^2-divergence. In the interests of conciseness, we first introduce a shorthand notation for the *loss variance* of a hypothesis $h \in \mathcal{H}$.

$$\sigma(h) \coloneqq \sqrt{\mathbb{E}_P \left[(h(Z) - \mathbb{E}[h(Z)])^2 \right]}.$$

Proposition 3. *For any function $h : \mathcal{H} \to \mathbb{R}$ with finite first and second moment under distribution P,*

$$\mathbb{E}[h(Z)] \leq R_2(P, h) \leq \mathbb{E}[h(Z)] + \sqrt{2} \rho^{1/2} \sigma(h).$$

Remark 4. In light of this proposition, we suggest that in the χ^2-divergence case, introducing the variance of the hypothesis as a regularizer can improve the distributional robustness, which is measured by $R_2(P, h)$.

3 Learning Guarantees for DRL

When no ambiguity exists, we use $\varphi_{\eta,h}(z)$ and c_k to denote $h(z) - \eta$ and $c_k(\rho)$, respectively. This section will be structured as follows: first, §3.1 provides an upper bound of the generalization error via entropy integral. §3.2 derives an upper bound of the excess risk through the lens of Rademacher complexity, and further bounds the Rademacher complexity with entropy integral. Finally, we establish the expected loss guarantee for the empirical DRL minimizer in §3.3. The proofs of the main results in this paper can be found in the Appendix.

3.1 Data-Dependent Bound on Generalization Error

We begin by imposing standard regularity assumptions which enable us to invoke concentration-of-measure results for empirical processes.

Assumption 5. The instance space $\mathcal{Z}$ is bounded as follows:

$$\operatorname{diam}(\mathcal{Z}) \coloneqq \sup_{z,z' \in \mathcal{Z}} \|z' - z\|_2 < \infty.$$

Assumption 6. The functions in $\mathcal{H}$ are upper semicontinuous and uniformly bounded as follows: $0 \leq h(z) \leq M < \infty$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$.

We use the *entropy integral* (Talagrand 2014) to measure the complexity of the hypothesis class $\mathcal{H}$:

$$\mathfrak{C}(\mathcal{H}) = \int_0^\infty \sqrt{\log N(\mathcal{H}, \|\cdot\|_\infty, u)} du,$$

where $N(\mathcal{H}, \|\cdot\|_\infty, u)$ denotes the covering number of $\mathcal{H}$ in the uniform metric $\|f - f'\|_\infty = \sup_{z \in \mathcal{Z}} |f(z) - f'(z)|$ with radius u.

Before introducing our main theoretical results, we first present a useful lemma. For ease of notation, for any fixed $h \in \mathcal{H}$, let

$$g_k(\eta, P) \coloneqq c_k \left(\mathbb{E}_P \left[(h(Z) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} \right) + \eta.$$

We have $R_k(P, h) = \inf_\eta g_k(\eta, P)$ from Lemma 2.

Lemma 7. *If $h(Z) \in [0, M]$, then for any distribution P:*

$$\inf_{\eta \in \mathbb{R}} g_k(\eta, P) = \inf_\eta \left\{ g_k(\eta, P) : \eta \in \left[-\frac{1}{c_k - 1} M, M \right] \right\}.$$

Remark 8. The above lemma restricts the domain of η to a compact set, which is crucial to our uniform concentration result.

Theorem 9. *Under Assumptions 5-6 and for any $t > 0$,*

$$\mathbf{P}\left(\exists h \in \mathcal{H} : R_k(P, h) > \min_{\eta \geq 0} \left\{ \mathbb{E}_{P_n}[(\varphi_{\eta,h})_+^{k_*}]^{\frac{1}{k_*}} + \eta + 1 \right.\right.$$
$$+ c_k \left(\frac{t}{\sqrt{n}} C_1(k, \rho, M) + \sqrt{\frac{\log(\eta + 1)}{n}} C_1(k, \rho, M) \right.$$
$$\left.\left.\left. + \frac{24}{\sqrt{n}} C_2(k, \rho, M) \mathfrak{C}(\mathcal{H}) \right)^{\frac{1}{k_*}} \right\} \right) \leq 2 \exp\left(-2t^2\right)$$

and

$$\mathbf{P}\left(\exists h \in \mathcal{H} : R_k(P_n, h) > \min_{\eta \geq 0} \left\{ \mathbb{E}_P[(\varphi_{\eta,h})_+^{k_*}]^{\frac{1}{k_*}} + \eta + 1 \right.\right.$$
$$+ c_k \left(\frac{t}{\sqrt{n}} C_1(k, \rho, M) + \sqrt{\frac{\log(\eta + 1)}{n}} C_1(k, \rho, M) \right.$$
$$\left.\left.\left. + \frac{24}{\sqrt{n}} C_2(k, \rho, M) \mathfrak{C}(\mathcal{H}) \right)^{\frac{1}{k_*}} \right\} \right) \leq 2 \exp\left(-2t^2\right)$$

where c_k and k_* are defined in Lemma 1 and $C_1(k, \rho, M)$, $C_2(k, \rho, M)$ are constants. More specifically, $C_1(k, \rho, M) = \left(\frac{c_k M}{c_k - 1} \right)^{k_*}$ and $C_2(k, \rho, M) = k_* \left(\frac{c_k M}{c_k - 1} \right)^{k_* - 1}$.

3.2 Excess Risk Bounds

Theorem 10. *Under Assumptions 5-6, for any $h \in \mathcal{H}$, the following holds with probability of at least $1 - \delta$:*

$$R_k(P, \hat{h}) - R_k(P, h) \leq 2 c_k \Big(\mathcal{R}_n(\psi \circ \Phi) \Big)^{\frac{1}{k_*}} + 3 C_3(k, \rho, M) \left(\frac{\log(2/\delta)}{2n} \right)^{\frac{1}{2k_*}}, \tag{11}$$

where

$$\Phi = \left\{ \varphi_{\eta,h} : h \in \mathcal{H},\, \eta \in \left[-\frac{1}{c_k - 1} M, M \right] \right\},$$

$\psi(t) = t_+^{k_*}$, $\psi \circ \Phi \coloneqq \{\psi \circ \varphi : \varphi \in \Phi\}$ and $C_3(k, \rho, M) = \frac{c_k^2 M}{c_k - 1}$. In particular, if we take $h = \operatorname*{argmin}_{h' \in \mathcal{H}} R_k(P, h')$, the left-hand side represents the excess risk.

Remark 11. Note that we bound the excess risk with the Rademacher complexity of the function class $\psi \circ \Phi$, which can be bounded by the Dudley's entropy integral (Talagrand 2014).

We next invoke an important lemma, which is useful for bounding the Rademacher complexity of $\psi \circ \Phi$ and can be found in (Mohri, Rostamizadeh, and Talwalkar 2012, Lemma 4.2).

Lemma 12 (Talagrand's contraction inequality)**.** *Let ψ be a ρ-Lipschitz function. For any function class $\mathcal{H}$, we have*

$$\mathcal{R}_n(\psi \circ \mathcal{H}) \leq \rho \mathcal{R}_n(\mathcal{H}).$$

Remark 13. Note that the function $t \mapsto t_+^{k_*}$ is not Lipschitz on the entire domain; however, the function class is bounded. By a truncation argument (Wainwright 2019, Example 5.29), the invoked lemma can be made applicable to our cases.

Proposition 14.

$$\mathcal{R}_n(\psi \circ \Phi) \leq \frac{1}{\sqrt{n}} \Big(C_4(k, \rho, M) \mathfrak{C}(\mathcal{H}) + C_5(k, \rho, M) \Big),$$

where $C_4(k, \rho, M) = 24 k_* \left(\frac{c_k M}{c_k - 1} \right)^{k_* - 1}$, $C_5(k, \rho, M) = 24 k_* \left(\frac{c_k M}{c_k - 1} \right)^{k_*}$.

Remark 15. The Rademacher complexity of the function class $\psi \circ \Phi$ can be bounded by the Dudley entropy integral of the original function class $\mathcal{H}$.

3.3 Expected Loss Guarantee

To establish expected loss guarantees for the locally minimax ERM, we present the following inequality, which relates the expected loss to the distributionally robust loss.

Proposition 16. *Under Assumption 2, we have*

$$\mathbb{E}_P[h(Z)] \leq R_k(P, h) \leq c_k(\rho) M^{\frac{1}{k}} \mathbb{E}_P[h(Z)]^{1 - \frac{1}{k}}.$$

Remark 17. Note that, for fixed ρ, when k approaches infinity, the right-hand side converges to $\mathbb{E}_P[h(Z)]$, which coincides with our intuition: the uncertainty set $\mathcal{U}_P$ shrinks as k increases.

Theorem 18. *Under Assumptions 1-2, the following holds with probability of at least* $1-\delta$*:*

$$\begin{aligned}\mathbb{E}_P\left[\hat{h}(Z)\right] \leq & c_k M^{\frac{1}{k}} \mathbb{E}_P\left[h^*_{avg}(Z)\right]^{1-\frac{1}{k}} \\ & + 2c_k\left(\mathcal{R}_n(\psi\circ\Phi)\right)^{\frac{1}{k_*}} \\ & + 3C_3(k,\rho,M)\left(\frac{\log(2/\delta)}{2n}\right)^{\frac{1}{2k_*}},\end{aligned}$$

where $\hat{h}$ *denotes the locally minimax ERM, and* $h^*_{avg} \coloneqq argmin_{h\in\mathcal{H}}\mathbb{E}_P[h(Z)]$.

Remark 19. Theorem 18 shows that when k approaches infinity, c_k and $M^{\frac{1}{k}}$ converge to 1 and $\mathbb{E}\left[h^*_{\text{avg}}(Z)\right]^{1-\frac{1}{k}}$ converges to $\mathbb{E}\left[h^*_{\text{avg}}(Z)\right]$. Therefore, given sufficient samples, the expected loss of locally minimax ERM is close to the optimal expected loss, which is in line with Hu et al. (2018, Theorem 2).

Proof sketch. Using the left half of Proposition 16 and Theorem 10,

$$\begin{aligned}\mathbb{E}_P\left[\hat{h}(Z)\right] \leq R_k(P,\hat{h}) \leq & R_k(P,h^*_{\text{avg}}) \\ & + 2c_k\left(\mathcal{R}_n(\psi\circ\Phi)\right)^{\frac{1}{k_*}} \\ & + 3C_3(k,\rho,M)\left(\frac{\log(2/\delta)}{2n}\right)^{\frac{1}{2k_*}}.\end{aligned}$$

We can then bound $R_k(P,h^*_{\text{avg}})$ using the right half of Proposition 16, and thereby obtain the desired result. □

4 Example Bounds

In this section, we illustrate the use of Theorem 10. Let the instance space $\mathcal{X}$ be a subset of the d-dimensional Euclidean space, namely $\mathbb{R}^d$. $\mathcal{Z}$ is equipped with the following Euclidean distance:

$$d_{\mathcal{Z}}(z,z') = \sqrt{\|x-x'\|_2^2 + |y-y'|^2}. \tag{12}$$

We use metric entropy to bound the Rademacher complexity, and accordingly require the following estimate of the covering number of balls in some metric spaces (Wainwright 2019, Lemma 5.7).

The following lemma relates the metric entropy to the so-called volume ratio. It involves the Minkowski sum $A+B \coloneqq \{a+b : a\in A, b\in B\}$, and the volume of the unit ball based on the Lebesgue measure is denoted by $\text{vol}(\mathbb{B}) \coloneqq \int \mathbb{I}\{x\in\mathbb{B}\}dx$.

Lemma 20 (Volume ratios and metric entropy)**.** *Consider a pair of norms* $\|\cdot\|$ *and* $\|\cdot\|'$ *on* $\mathbb{R}^d$*, and let* $\mathbb{B}$ *and* $\mathbb{B}'$ *be their corresponding unit balls (i.e.,* $\mathbb{B} = \left\{\theta\in\mathbb{R}^d \middle| \|\theta\|\leq 1\right\}$*, with* $\mathbb{B}'$ *similarily defined). Then the* δ*-covering number of* $\mathbb{B}$ *in the* $\|\cdot\|'$*-norm therefore obeys the bounds*

$$\left(\frac{1}{\delta}\right)^d \frac{vol(\mathbb{B})}{vol(\mathbb{B})'} \leq N(\delta,\mathbb{B},\|\cdot\|') \overset{(a)}{\leq} \frac{vol(\frac{2}{\delta}\mathbb{B}+\mathbb{B}')}{vol(\mathbb{B}')}.$$

Whenever $\mathbb{B}'\subseteq\mathbb{B}$, the upper bound (a) may be simplified by observing that

$$\text{vol}\left(\frac{2}{\delta}\mathbb{B}+\mathbb{B}'\right) \leq \text{vol}\left(\left(\frac{2}{\delta}+1\right)\mathbb{B}\right) = \left(\frac{2}{\delta}+1\right)^d \text{vol}(\mathbb{B}).$$

4.1 Linear Classifier in Binary Classification

We begin here with binary linear classifiers. In this setting, we define $\mathcal{X} = \{x\in\mathbb{R}^d : \|x\|_2\leq r_0\}$, $\mathcal{Y} = \{-1,+1\}$, and let the hypothesis class $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}$ be a set of linear functions of $x\in\mathcal{X}$. More specifically, we define $f_w(x) = \langle w,x\rangle$ and consider prediction vector w with l_2 norm constraint, i.e.,

$$\mathcal{F} = \{f_w(x) : \|w\|_2 \leq W\}.$$

Here we consider the hinge loss, i.e., $\ell(f_w(x),y) = \phi(y\langle w,x\rangle) = \max\{0,1-y\langle w,x\rangle\}$.
Thanks to Proposition 14, it is sufficient to bound the entropy integral of

$$\mathcal{H}_L \coloneqq \{(x,y)\mapsto\phi(y\langle w,x\rangle) : \|w\|_2\leq W\}.$$

Corollary 21. *For any distribution* P *on* $\mathcal{Z}$*, with probability of at least* $1-\delta$*,*

$$\begin{aligned}& R_k(P,\hat{h}) - R^*_k(P,\mathcal{H}_L) \\ \leq & 2c_k n^{-\frac{1}{2k^*}}\Big(C_4(k,\rho,1+Wr_0)(\sqrt{d\log 3}+3Wr_0\sqrt{d}/2) \\ & + C_5(k,\rho,1+Wr_0)\Big)^{\frac{1}{2k^*}} \\ & + 3C_3(k,\rho,1+Wr_0)\left(\frac{\log(2/\delta)}{2n}\right)^{\frac{1}{2k_*}},\end{aligned} \tag{13}$$

where $R^*_k(P,\mathcal{H}_L) \coloneqq \inf_{h'\in\mathcal{H}_L} R_k(P,h')$.

4.2 Gaussian Reproducing Kernel Hilbert Space

In this case, we consider the instance space $\mathcal{X} = \{x\in\mathcal{X} : \|x\|\leq r_0)\}$ and label space $\mathcal{Y} = [-B,B]$ for some value of $r_0,\ B>0$, and equip $\mathcal{Z}$ with the Euclidean metric (12).

Let $(\mathcal{H}_K,\|\cdot\|_K)$ be the Gaussian reproducing kernel Hilbert space (RKHS) with the kernel $K(x_1,x_2) = \exp(-\|x_1-x_2\|_2^2/\sigma^2)$ for some $\sigma>0$, and let $B_r \coloneqq \{h\in\mathcal{H}_K : \|h\|\leq r\}$ be the radius-r ball in $\mathcal{H}_K$. Let $\mathcal{H}$ be the class of all functions of the form $h(z) = (y-f_0(x))^2$, where the predictor $f_0 : \mathcal{X}\mapsto\mathbb{R}$ belongs to $I_K(B_r)$. Here, $I_K(B_r)$ denotes an embedding of B_r into the space $C(\mathcal{X})$ of continuous real-valued functions on $\mathcal{X}$ equipped with the sup norm $\|f\|_{\mathcal{X}} \coloneqq \sup_{x\in\mathcal{X}}|f(x)|$.

Using the covering number estimates drawn from the work of Cucker and Ding Xuan (2007), we can prove the generalization bound for Gaussian RKHS.

Proposition 22. *For compact* $\mathcal{X}\subset\mathbb{R}^d$*, the following holds for all* $u\in(0,r/2]$*:*

$$\begin{aligned}& \log N(I_K(B_r),\|\cdot\|_{\mathcal{X}},u) \\ \leq & d\left(32+\frac{640d(diam(\mathcal{X}))2}{\sigma^2}\right)^{d+1}\left(\log\frac{r}{u}\right)^{d+1}.\end{aligned}$$

Corollary 23. *For any distribution P on $\mathcal{Z}$, with probability of at least $1-\delta$,*

$$\begin{aligned} & R_k(P,\hat{h}) - R_k^*(P,\mathcal{H}) \\ \leq & 2c_k n^{-\frac{1}{2k^*}} \Big(C_4(k,\rho,2B^2+2r^2)K_1 \\ & + C_5(k,\rho,2(B^2+r^2)) \Big)^{\frac{1}{2k^*}} \\ & + 3C_3(k,\rho,2(B^2+r^2)) \left(\frac{\log(2/\delta)}{2n} \right)^{\frac{1}{2k_*}}, \end{aligned} \tag{14}$$

where $R_k^*(P,\mathcal{H}) := \inf_{h' \in \mathcal{H}} R_k(P,h')$ and

$$K_1 = \sqrt{d} \left(2\Gamma\left(\frac{d+3}{2}, \log 2 \right) + (\log 2)^{\frac{d+1}{2}} \right) \left(32 + \frac{2560 d r_0^2}{\sigma^2} \right)^{\frac{d+1}{2}} \left(r^2 + Br \right).$$

4.3 One-Hidden-Layer Neural Network

Here, we consider the loss class $\mathcal{H} = \{(x,y) \mapsto \max\{0, 1 - y\sum_{j=1}^m u_j s(\alpha_j^T x)\} : \|u\|_1 \leq \Lambda, \|\alpha_j\|_2 \leq \Omega, \text{for } j = 1,\ldots,m\}$, where $u = (u_1,\ldots,u_m)$, $\alpha_j \in \mathbb{R}^m$ for $j = 1,\ldots,m$ and $s(\cdot)$ is the ReLU activation.

Corollary 24. *For any distribution P on $\mathcal{Z}$, with probability of at least $1-\delta$,*

$$\begin{aligned} & R_k(P,\hat{h}) - R_k^*(P,\mathcal{H}) \\ \leq & 2c_k n^{-\frac{1}{2k^*}} \Big(C_4(k,\rho,1+\Omega\Lambda r_0)K_2 \\ & + C_5(k,\rho,1+\Omega\Lambda r_0) \Big)^{\frac{1}{2k^*}} \\ & + 3C_3(k,\rho,1+\Omega\Lambda r_0) \left(\frac{\log(2/\delta)}{2n} \right)^{\frac{1}{2k_*}}, \end{aligned} \tag{15}$$

where $R_k^*(P,\mathcal{H}) := \inf_{h' \in \mathcal{H}} R_k(P,h')$ and $K_2 = 3d^{\frac{1}{2}} m^{\frac{3}{2}} \Lambda\Omega r_0 + 3\Lambda d^{\frac{1}{2}}$.

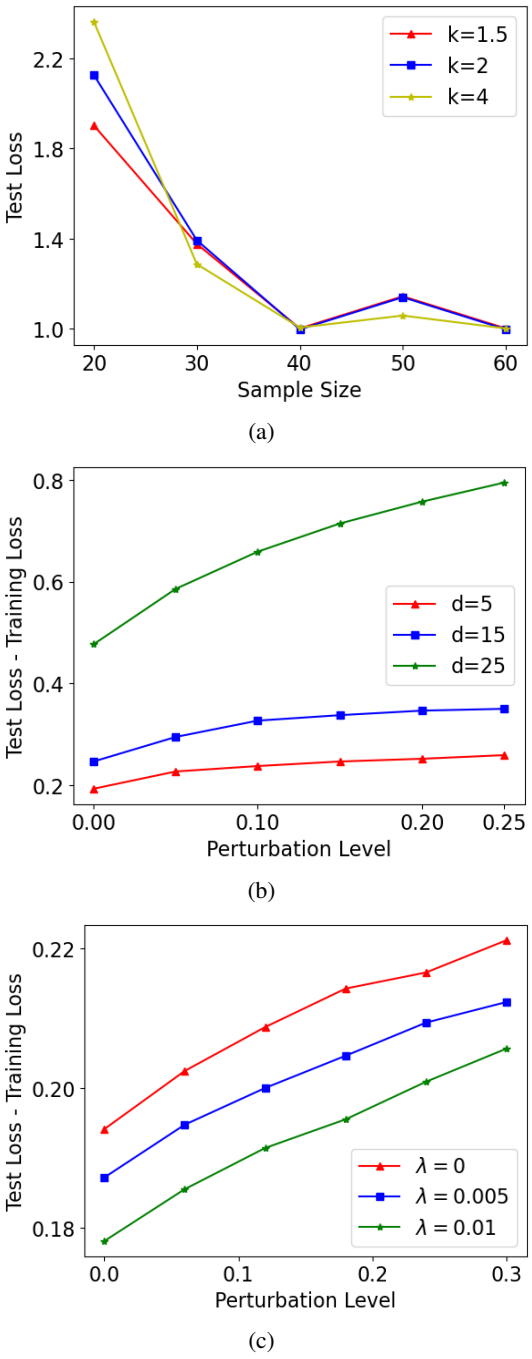
5 Numerical Study

In this section, we validate three theoretical findings for linear classifiers, as follows: (i) verifying that the decay rate of distributionally robust generalization is dependent on k; (ii) establishing that there is a dimension dependence in distributionally robust generalization, i.e., that distributionally robust generalization is more difficult when the dimension of the feature space is higher; (iii) verifying that controlling the l_2 norm of the model parameter can reduce distributionally robust generalization.

We investigate distributionally robust generalization via a binary classification experiment using hinge loss $l(w,(x,y)) = (1 - yx^T w)_+$, where $y \in \{+1,-1\}$ and $x \in \mathbb{R}^d$. We select a vector $w_0^* \in \mathbb{R}^d$ uniformly on the unit sphere and generate data as follows:

$$X \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, I_d) \quad \text{and}$$

$$Y|X = \begin{cases} \text{sign}(X^T w_0^*) & \text{w.p. } 0.9 \\ -\text{sign}(X^T w_0^*) & \text{w.p. } 0.1 \end{cases}$$



(a)

(b)

(c)

Figure 1: (a) Linear classifier. Test loss under different sample sizes n and parameters k; (b) Linear classifier. Test loss under different perturbations ρ and feature space dimensions d; (c) Linear classifier. Excess risk under different perturbations ρ and regularization coefficients λ.

where $N(\mathbf{0}, I_d)$ denotes the normal distribution with mean

$\mathbf{0}$ and covariance matrix I_d, and

$$\text{sign}(x) = \begin{cases} 1, & \text{if} \quad x > 0 \\ 0, & \text{if} \quad x = 0 \\ -1, & \text{if} \quad x < 0 \end{cases}$$

We train the binary linear classifier using the following objective function:

$$\min_{w} \sup_{Q \ll P_n} \{\mathbb{E}_Q[l(f_w(x), y)] : D_{f_k}(Q \| P_n) \leq \rho\} + \lambda \|w\|_2^2, \tag{16}$$

where $l(\cdot)$ is the hinge loss and $f_w(x) \coloneqq \langle w, x \rangle$. With the aid of the duality formulation, we can transfer the mini-max problem (16) to a min-min problem:

$$\min_{w,\eta} \left\{ c_k(\rho) \mathbb{E}_{P_n} \left[(l(f_w(x), y) - \eta)_+^{k*} \right]^{\frac{1}{k_*}} + \eta + \lambda \|w\|_2^2 \right\}.$$

Thus, we can jointly minimize the distributionally robust risk via gradient descent.

In our first experiment, we set $d = 5$, $\lambda = 0$ and vary the values of k and sample size n. For each (n, k) pair, we run the training algorithm; in each run, we sample n training data independently. In Figure 1(a), we plot the test risk as a function of n and k. As we can see from Figure 1(a), the decay rate of the test risk increases with the increasing k, which is in line with our Theorem 10: the test risk has an $O_p(n^{-\frac{1}{2k_*}})$ rate, where $k_* = \frac{k}{k-1}$. However, we can also observe that at a lower number of training samples, the generalization error increases with the increasing k. This phenomenon is explicable; in our theoretical result Corollary 21, when k increases, the constants $C_3(k, \rho, 1 + Wr_0)$ and $C_4(k, \rho, 1 + Wr_0)$ also increase.

In our second experiment, we choose $\lambda = 0$, $k = 2$, and study the dependence of distributional generalization error on the dimension of the feature space. We construct two additional synthetic datasets with $d = 15$ and $d = 20$, respectively. To facilitate fair comparison, we generate the data in a similar way to the first experiment,

$$X \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, I_d/d).$$

Moreover, we keep the distribution of Y conditioned on X, in the same way as in the first experiment. This construction ensures that the l_2 norm of every single data point is kept in the same order across the three datasets. The distributional generalization error gap is plotted in Figure 1(b). From Figure 1(b), we can observe that the generalization gap increases as the dimension d increases.

In our third experiment, we fix $k = 2$ and vary the values of ρ and λ. We run the training algorithm for each (ρ, λ) pair; in each run we sample 100 points of training data independently. In Figure 1(c), we plot the distributional generalization error gap as a function of ρ and λ. From Figure 1(c), we can observe that the generalization gap decreases with increasing λ. Accordingly, we conclude that l_2 regularization is helpful for reducing the distributionally generalization error.

6 Related Work

Distributional shifts arise in many guises across the fields of statistics, machine learning, applied probability, simulation and optimization. We here provide a necessarily abridged survey of the strands of work in this area, along with their respective foci. Domain adaptation seeks models that are trained on data from one domain, and performs well on a specified target domain. A typical approach of this kind aims to reweight the distribution P to make it "closer" to the known target distribution P_{target} (Shimodaira 2000; Huang et al. 2006; Bickel, Brückner, and Scheffer 2007; Sugiyama, Krauledat, and Müller 2007; Sugiyama et al. 2007).

In the optimization literature, a substantial body of work focuses on distributionally robust optimization problems. Several authors investigate worst-case regions arising out of moment conditions on the data vector X (Delage and Ye 2010; Jiang and Guan 2016). Other works (Ben-Tal et al. 2013; Duchi, Glynn, and Namkoong 2021; Namkoong and Duchi 2017; Lam 2016; Lam and Zhou 2017; Zhou and Liu 2023) studies a scenario similar to our f-divergence formulation (1).

An alternative to our f-divergence based sets $\{Q : D_f(Q\|P) \leq \rho\}$ is Wasserstein balls (Wozabal 2012; Shafieezadeh-Abadeh, Esfahani, and Kuhn 2015; Blanchet and Murthy 2019; Blanchet, Kang, and M. 2019; Rui and Kleywegt 2016; Esfahani and Kuhn 2018; Sinha, Namkoong, and Duchi 2017). Wasserstein ball allows worst-case distributions with different support from the data-generating distribution P. This property, however, means that tractable reformulations are only available under restrictive scenarios (Shafieezadeh-Abadeh, Esfahani, and Kuhn 2015; Esfahani and Kuhn 2018; Sinha, Namkoong, and Duchi 2017) and that they remain computationally challenging. In comparison, our f-divergence formulation is computationally efficient to solve.

7 Conclusion

In this work, we present the theoretical guarantees for distributionally robust learning (Theorems 9 and 10). Our results are derived through the lens of Rademacher complexity, which has not previously been applied before. We prove an $O_P(n^{-\frac{1}{2k_*}} (\log(2/\delta))^{\frac{1}{2k_*}})$ convergence rate with probability of at least $1 - \delta$. The empirical results verify our theoretical findings.

Acknowledgments

This work is supported by the Key R&D Program of Hubei Province under Grant 2024BAB038, the National Key R&D Program of China under Grant 2023YFC3604702, the Fundamental Research Funds for the Central Universities under Grant 2042025kf0045.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of Representations for Domain Adaptation. In *NeurIPS*, 137–144.

Ben-Tal, A.; den Hertog, D.; Waegenaere, A. D.; Melenberg, B.; and Rennen, G. 2013. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2): 341–357.

Bickel, S.; Brückner, M.; and Scheffer, T. 2007. Discriminative learning for differing training and test distributions. In *ICML*, volume 227, 81–88.

Blanchet, J. H.; Kang, Y.; and M., K. R. A. 2019. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3): 830–857.

Blanchet, J. H.; and Murthy, K. R. A. 2019. Quantifying Distributional Model Risk via Optimal Transport. *Mathematical Operation Research*, 44(2): 565–600.

Blitzer, J.; McDonald, R. T.; and Pereira, F. 2006. Domain Adaptation with Structural Correspondence Learning. In *EMNLP*, 120–128.

Blodgett, S. L.; Green, L.; and O'Connor, B. T. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *EMNLP*, 1119–1130.

Cressie, N.; and Read, T. R. C. 1984. Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society*, 46(3): 440–464.

Cucker, F.; and Ding Xuan, Z. 2007. *Learning theory: An approximation theory viewpoint*, volume 24.

Delage, E.; and Ye, Y. 2010. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operation Research*, 58(3): 595–612.

Duchi, J. C.; Glynn, P. W.; and Namkoong, H. 2021. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 46(3): 946–969.

Duchi, J. C.; and Namkoong, H. 2021. Learning Models with Uniform Performance via Distributionally Robust Optimization. *The Annals of Statistics*, 49(3): 1378–1406.

Esfahani, P. M.; and Kuhn, D. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2): 115–166.

Gong, X.; Yuan, D.; and Bao, W. 2021a. Discriminative metric learning for partial label learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4428–4439.

Gong, X.; Yuan, D.; and Bao, W. 2021b. Understanding partial multi-label learning via mutual information. In *NeurIPS*.

Gong, X.; Yuan, D.; and Bao, W. 2022. Partial label learning via label influence function. In *ICML*.

Gong, X.; Yuan, D.; Bao, W.; and Luo, F. 2022. A unifying probabilistic framework for partially labeled data learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8036–8048.

Hand, D. J. 2006. Classifier technology and the illusion of progress. *Statistical Science*, 21(1): 1–14.

Hashimoto, T. B.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *ICML*, volume 80, 1934–1943.

Hu, W.; Niu, G.; Sato, I.; and Sugiyama, M. 2018. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In *ICML*, volume 80, 2034–2042.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2006. Correcting Sample Selection Bias by Unlabeled Data. In *In NeurIPS*, 601–608.

Jiang, R.; and Guan, Y. 2016. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2): 291–327.

Knight, J. C. 2002. Safety critical systems: Challenges and directions. In *ICSE*, 547–550.

Krokhmal, P. A. 2007. Higher moment coherent risk measures.

Lam, H. 2016. Robust Sensitivity Analysis for Stochastic Systems. *Mathematics of Operations Research*, 41(4): 1248–1275.

Lam, H.; and Zhou, E. 2017. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4): 301–307.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*.

Namkoong, H.; and Duchi, J. C. 2017. Variance-based Regularization with Convex Objectives. In *NeurIPS*, 2971–2980.

Rockafellar, R. T.; Uryas'ev, S.; et al. 2000. Optimization of conditional value-at-risk. *Journal of Risk*, 2: 21–42.

Rui, G.; and Kleywegt, A. J. 2016. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting Visual Category Models to New Domains. In *ECCV*, volume 6314, 213–226.

Shafieezadeh-Abadeh, S.; Esfahani, P. M.; and Kuhn, D. 2015. Distributionally Robust Logistic Regression. In *NeurIPS*, 1576–1584.

Shapiro, A. 2017. Distributionally Robust Stochastic Programming. *SIAM Journal on Optimization*, 27(4): 2258–2275.

Shapiro, A.; Dentcheva, D.; and Ruszczynski, A. 2014. *Lectures on stochastic programming: Modeling and theory*.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2): 227–244.

Sinha, A.; Namkoong, H.; and Duchi, J. C. 2017. Certifiable Distributional Robustness with Principled Adversarial Training. *CoRR*, abs/1710.10571.

Sugiyama, M.; Krauledat, M.; and Müller, K. 2007. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, 8: 985–1005.

Sugiyama, M.; Nakajima, S.; Kashima, H.; von Bünau, P.; and Kawanabe, M. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *NeurIPS*, 1433–1440.

Talagrand, M. 2014. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*, 1521–1528.

van Erven, T.; and Harremoës, P. 2014. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7): 3797–3820.

Wainwright, M. J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48.

Wozabal, D. 2012. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1): 21–47.

Zheng, C.; Shi, Z.; Miao, R.; Liu, W.; Yang, T.; Cui, B.; and Uhlig, S. 2025. Answering Subset Query Over Multi-Attribute Data Streams Using Hyper-USS. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, Z.; and Liu, W. 2023. Sample complexity for distributionally robust learning under chi-square divergence. *Journal of Machine Learning Research*.