Coverage-Guaranteed Prediction Sets for Out-of-Distribution Data

Xin Zou, Weiwei Liu*

School of Computer Science,
Institute of Artificial Intelligence,
National Engineering Research Center for Multimedia Software,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
Wuhan University, China
{zouxin2021, liuweiwei863}@gmail.com

Abstract

Out-of-distribution (OOD) generalization has attracted increasing research attention in recent years, due to its promising experimental results in real-world applications. In this paper, we study the confidence set prediction problem in the OOD generalization setting. Split conformal prediction (SCP) is an efficient framework for handling the confidence set prediction problem. However, the validity of SCP requires the examples to be exchangeable, which is violated in the OOD setting. Empirically, we show that trivially applying SCP results in a failure to maintain the marginal coverage when the unseen target domain is different from the source domain. To address this issue, we develop a method for forming confident prediction sets in the OOD setting and theoretically prove the validity of our method. Finally, we conduct experiments on simulated data to empirically verify the correctness of our theory and the validity of our proposed method.

1 Introduction

Recent years have witnessed the remarkable success of modern machine learning techniques in many applications. A fundamental assumption of most machine learning algorithms is that the training and test data are drawn from the same underlying distribution. However, this assumption is consistently violated in many practical applications. In reality, the test environment is influenced by a range of factors, such as the distributional shifts across photos caused by the use of different cameras in image classification tasks, the voices of different persons in voice recognition tasks, and the variations between scenes in self-driving tasks (Nagarajan, Andreassen, and Neyshabur 2021). As a result, there is now a rapidly growing body of research with a focus on generalizing to unseen target domains with the help of the source domains, namely OOD generalization (Shen et al. 2021).

Existing OOD generalization methods focus on improving worst-case performance on the target domains, i.e., improving the average test accuracy of the model on the worst target domain. However, in some systems that require high security (such as medical diagnosis), even a single mistake may have disastrous consequences. In these cases, it is important to quantify the uncertainty of the predictions. One way to perform uncertainty estimation (Amodei et al. 2016; Jiang et al. 2012, 2018; Angelopoulos et al. 2021) is to create confident prediction sets that provably contain the correct answer with high probability. Let $X_{n+1} \in \mathcal{X}$ be a new test example for which we would like to predict the corresponding label $Y_{n+1} \in \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. For any given $\alpha \in (0, 1)$, the aim of confidence set prediction is to construct a set-valued output, $\mathcal{C}(X_{n+1})$, which contains the label Y_{n+1} with distribution-free marginal coverage at a significance level α, i.e., $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1})\right) \geq 1 - \alpha$. A confidence set predictor $\mathcal{C}^\alpha$ is said to be **valid** if $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^\alpha(X_{n+1})\right) \geq 1 - \alpha$ for any $\alpha \in (0, 1)$, where α is a hyper-parameter of the predictor. To simplify the notation, we omit the superscript α in the remainder of this paper.

Conformal prediction (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008, **CP**) is a model-agnostic, non-parametric and distribution-free (the coverage guarantee holds for any distribution) framework for creating confident prediction sets. Split conformal prediction (Vovk 2013; Vovk, Gammerman, and Shafer 2005, **SCP**), a special type of CP, has been shown to be computationally efficient. SCP reserves a set of data as the calibration set, and then uses the relative value of scores of the calibration set and that of a new test example to construct the prediction set. The validity of SCP relies on the assumption that the examples are exchangeable. However, in the OOD setting, the distributional shift between the training and test distributions leads to the violation of the exchangeability assumption. We empirically evaluate the performance of SCP in the OOD setting in Section 4. Unfortunately, we find that trivially applying SCP results in a failure to maintain marginal coverage in the OOD setting.

To address this issue, we construct a set predictor based on the f-divergence (Alfréd 1961) between the test distribution (target domain) and the convex hull of the training distributions (source domains). We theoretically show that our set predictor is guaranteed to maintain the marginal coverage (Corollary 9). We then conduct simulation experiments to verify our theory.

The remainder of this article is structured as follows: §2 introduces some related works; §3 presents the notation definitions and preliminaries; §4 conducts experiments that show the failure of SCP in the OOD generalization setting; §5 creates corrected confidence set predictor in the OOD gen-

*corresponding author

eralization setting; §6 provides our experimental results. §7 make discussions with the most related work. Finally, the conclusions are presented in §8. All of our proofs are attached in Appendix A.

2 Related Works

OOD generalization. OOD generalization aims to train a model with data from the source domains so that it is capable of generalizing to an unseen target domain. A large number of algorithms have been developed to improve OOD generalization. One series of works focuses on minimizing the discrepancies between the source domains (Li et al. 2018b; Ganin et al. 2016; Li et al. 2018c; Sun and Saenko 2016). Meta-learning domain generalization (Li et al. 2018a, MLDG) leverages the meta-learning approach and simulates train/test distributional shift during training by synthesizing virtual testing domains within each mini-batch. Another line of works (Xin et al. 2022; Wang et al. 2022) conducts adversarial training (Madry et al. 2018) to improve the OOD generalization performance. (Zou and Liu 2023) considers improving the adversarial robustness of the unseen target domain. Notably, the above works all focus on improving the average performance on the target domain; in contrast, we focus on designing valid confidence set predictors for data from the unseen target domains, as this is a crucial element of making high-stakes decisions in systems that require high security.

Conformal prediction. As introduced in §1, conformal prediction is a model-agnostic, non-parametric, and distribution-free framework that provides valid confidence set predictors. Generally speaking, examples are assumed to be exchangeable in a CP context. Most pertinent to our work, (Gendler et al. 2022; Tibshirani et al. 2019; Fisch et al. 2021; Cauchois et al. 2020; Gibbs and Candès 2021; Oliveira et al. 2022) all consider various situations in which the exchangeability of the examples is violated to some extent. (Gendler et al. 2022) considers the case in which the test examples may be adversarially attacked (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018); (Tibshirani et al. 2019) investigates the situation in which the density ratio between the target domain and the source domain is known; (Fisch et al. 2021) studies the few-shot learning setting and assumes that the source domains and the target domain are independent and identically distributed (i.i.d.) from some distribution on the domains; (Gibbs and Candès 2021) considers an online learning setting and (Oliveira et al. 2022) provides results when the examples are mixing (Achim 2013; Xiaohong, Lars Peter, and Marine 2010; Bin 1994). Different from all the works discussed above, we consider the OOD generalization setting in which the f-divergence between the target domain and the convex hull of the source domains is constrained. The most related work among them is (Cauchois et al. 2020), which studies the worst-case coverage guarantee of a f-divergence ball centered at the single source domain. For the discussions about similarities and differences with (Cauchois et al. 2020), please refer to Section 7.

3 Preliminaries

We begin with the OOD setups and a review of conformal prediction.

Notations. We denote $\{1, 2, \ldots, n\}$ by $[n]$ for $n \in \mathbb{N}_+$. For a distribution P on $\mathbb{R}$, we define the quantile function of P as $\mathcal{Q}(\beta; P) := \inf\{s \in \mathbb{R} | P(S \leq s) \geq \beta\}$. Similarly, for a cumulative distribution function (c.d.f.) F on $\mathbb{R}$, we define $\mathcal{Q}(\beta; F) := \inf\{s \in \mathbb{R} | F(s) \geq \beta\}$. For n distributions $P_1, \ldots, P_n$, we define $\mathcal{CH}(P_1, \ldots, P_n) := \{\sum_{i=1}^n \lambda_i P_i | \lambda_1, \ldots, \lambda_n \geq 0; \sum_{i=1}^n \lambda_i = 1\}$ as the convex hull of the distributions $P_1, \ldots, P_n$. We further define $\mathcal{N}(\mu, \Sigma)$ as the multi-variable Gaussian distribution with mean vector μ and covariance matrix Σ. For a set A, we define the indicator function as $\mathbb{I}_A(\cdot)$, where $\mathbb{I}_A(x) = 1$ if $x \in A$ and $\mathbb{I}_A(x) = 0$ otherwise.

3.1 Out-of-Distribution Generalization

We define the input space as $\mathcal{X}$ and the label space as $\mathcal{Y}$. We set $\mathcal{Y} = \{\pm 1\}$, $\mathcal{Y} = \{1, 2, \ldots, K\}$ (where K is the number of classes), and $\mathcal{Y} = \mathbb{R}$ for the binary classification problem, the multi-class classification problem, and the regression problem, respectively. Let $\mathcal{S} := \{S_1, \ldots, S_d\}$ be the set of source domains, where d is the number of source domains. $S_1, \ldots, S_d$ are distributions on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and we use the terminologies "domain" and "distribution" interchangeably in this paper. Let T denote the target domain. The goal of OOD generalization is to obtain good performance on all $T \in \mathcal{T}$, where $\mathcal{T}$ is the set of all possible target domains; we usually assume $\mathcal{S} \subseteq \mathcal{T}$.

In a standard OOD generalization setting, we learn a predictor $h \in \mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ from the source domains $\mathcal{S}$ and define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_*$ where $\mathbb{R}_* = [0, +\infty)$. We aim to minimize the worst-case population risk of the predictor h on the unseen target domain as follows:

$$\mathcal{R}_{\mathcal{T}}(h) = \max_{T \in \mathcal{T}} \mathop{\mathbb{E}}_{(X,Y) \sim T} \left[\ell\left(h(X), Y\right)\right].$$

However, in some systems that require high security, a mistake may lead to serious disasters. In these cases, a good solution is to output a prediction set with a marginal coverage guarantee. For a predefined confidence level $1 - \alpha \in (0, 1)$, we wish to output a prediction set $\mathcal{C}(x) \subseteq \mathcal{Y}$ such that, for any $T \in \mathcal{T}$:

$$\mathop{\mathbb{P}}_{(X,Y) \sim T, \mathcal{C}} [Y \in \mathcal{C}(X)] \geq 1 - \alpha, \tag{1}$$

where the probability is over the randomness of test examples $(X, Y) \sim T$ and the randomness of the prediction set $\mathcal{C}$. To achieve (1), we follow the idea of SCP (Vovk 2013; Vovk, Gammerman, and Shafer 2005) to construct $\mathcal{C}(x)$. The next section introduces the main idea of SCP.

3.2 Split Conformal Prediction

Nonconformity score. In SCP, we consider a supervised learning problem that involves predicting the label $y \in \mathcal{Y}$ of the input $x \in \mathcal{X}$. We assume that we have a predictive model $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which outputs the nonconformity score $s(x, y)$. The nonconformity score function $s(\cdot, \cdot)$ is

usually trained with a set of training data. $s(x, y) < s(x, y')$ means that for the input x, y is more likely than y' to be the label. Some examples of nonconformity scores are as follows: for a probabilistic model $p(y|x)$, we can take the negative log-likelihood as the score, $s(x, y) = -\log(p(y|x))$; for a regression model $h : \mathcal{X} \to \mathcal{Y}$, a typical choice is $s(x, y) = |h(x) - y|$; for a multi-class classifier $h : \mathcal{X} \to \Delta^{K-1}$, where Δ^{K-1} is the $K-1$ dimensional simplex in $\mathbb{R}^K$, we can take $s(x, y) = 1 - h(x)_y$.

In SCP, we assume that the examples $\{(X_i, Y_i)\}_{i=1}^{n+1} \subseteq \mathcal{X} \times \mathcal{Y}$ are exchangeable (Definition 1). For a predefined significance level $\alpha \in (0, 1)$, the goal is to provide a valid confidence set $\widehat{\mathcal{C}}(X_{n+1})$. CP methods (Shafer and Vovk 2008) take advantage of the exchangeability of the data and the properties of the quantile function to make such a construction possible.

Definition 1 ((Shafer and Vovk 2008, Exchangeability))**.** The random variables $Z_1, \ldots, Z_n$ are exchangeable if for every permutation τ for integers $1, \ldots, n$, the variables $W_1, \ldots, W_n$, where $W_i = Z_{\tau(i)}$, have the same joint probability distribution as $Z_1, \ldots, Z_n$.

Let $V_i = s(X_i, Y_i)$ for $i \in [n + 1]$ be the nonconformity scores corresponding to the examples $\{(X_i, Y_i)\}_{i=1}^{n+1}$, where $s(\cdot, \cdot)$ is independent of $\{(X_i, Y_i)\}_{i=1}^{n+1}$. The independence between $s(\cdot, \cdot)$ and $\{(X_i, Y_i)\}_{i=1}^{n+1}$ is useful since in this case we can prove that the scores $\{V_i\}_{i=1}^{n+1}$ are exchangeable. The exchangeability of $\{V_i\}_{i=1}^{n+1}$ comes from the exchangeability of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ and the independence between the $s(\cdot, \cdot)$ and $\{(X_i, Y_i)\}_{i=1}^{n+1}$. Next, define $\text{rank}(V_i)$ as the rank of V_i among $\{V_i\}_{i=1}^{n+1}$ for any $i \in [n + 1]$ (in ascending order; we assume that ties are broken randomly). By the exchangeability of $\{V_i\}_{i=1}^{n+1}$, $\text{rank}(V_i)$ is uniform on $[n + 1]$, which is used to prove the validity of SCP in Lemma 2. We use $\widehat{P}(\{V_i\}_{i=1}^{n})$ to denote the empirical distribution determined by the examples $V_1, \ldots, V_n$. Let

$$\widehat{\mathcal{C}}_n(x) := \left\{ y \in \mathcal{Y} \middle| s(x, y) \leq \mathcal{Q}\left(\frac{n+1}{n}(1-\alpha); \widehat{P}(\{V_i\}_{i=1}^{n}) \right) \right\}, \tag{2}$$

we then have the following marginal coverage guarantee.

Lemma 2 (The validity of SCP)**.** *Assume that examples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable. For any nonconformity score $s(\cdot, \cdot)$ and any $\alpha \in (0, 1)$, the prediction set defined in Equation* (2) *satisfies:*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1}) \right) \geq 1 - \alpha, \tag{3}$$

where the probability is over the randomness of $\{(X_i, Y_i)\}_{i=1}^{n+1}$.

In the OOD generalization setting, we also want to obtain a valid set predictor that is valid for any $T \in \mathcal{T}$. In light of this, some natural questions arise:

Does the set predictor defined in Equation (2) *remain valid when the unseen target domain is different from the source domains? If not, can we construct a new set predictor that is valid in the OOD generalization setting?*

Unfortunately, the answer to the first question is **negative**. Theoretically, as shown in Appendix A.1, the proof of Lemma 2 is highly dependent on the exchangeability of the examples $\{(X_i, Y_i)\}_{i=1}^{n+1}$, which is easily violated if there is any distributional shift between the distribution of $\{(X_i, Y_i)\}_{i=1}^{n}$ and the distribution of (X_{n+1}, Y_{n+1}). This means that in the OOD setting, the proof technique of Lemma 2 cannot be applied. Empirically, in Section 4, we provide a toy example to show that the set predictor $\widehat{\mathcal{C}}_n(x)$ is no longer valid in the OOD setting.

In Section 5, we give an **affirmative** answer to the second question. We first construct a new set predictor based on the f-divergence between the target domain and the convex hull of the source domains, then provide marginal coverage guarantees for the constructed predictor.

4 SCP Fails in the OOD Setting

In this section, we construct a toy example to show that for the OOD confidence set prediction problem, SCP is no longer valid, even under a slight distributional shift.

For simplicity, we consider a single-domain case. Specifically, we consider the regression problem and set $\mathcal{X} = \mathbb{R}^l$, $\mathcal{Y} = \mathbb{R}$. We define the source domain S as follows: given a linear predictor $L(x) = \langle w^\star, x \rangle + b^\star$ where $w^\star \in \mathbb{R}^l$ and $b^\star \in \mathbb{R}$. The marginal distribution of X and the conditional distribution of Y given X are defined as:

$$X \sim \mathcal{N}(\mu_s, \sigma_{s,x}^2 I_l), \quad Y|X = x \sim \mathcal{N}(L(x), \sigma_{s,y}^2),$$

where $\mu_s \in \mathbb{R}^l$ is the mean vector of X, $\sigma_{s,x}, \sigma_{s,y}$ are positive scalars, and $I_l \in \mathbb{R}^{l \times l}$ is an identity matrix. Similarly, for the target domain T, we define:

$$X \sim \mathcal{N}(\mu_t, \sigma_{t,x}^2 I_l), \quad Y|X = x \sim \mathcal{N}(L(x), \sigma_{t,y}^2),$$

where $\mu_t \in \mathbb{R}^l$ is the mean vector of X and $\sigma_{t,x}, \sigma_{t,y}$ are positive scalars. For simplicity, we set $\mu_s = \mu_t, \sigma_{s,x} = \sigma_{t,x}$ and $\sigma_{s,y} \neq \sigma_{t,y}$. We sample m_{train} training examples from S to train a linear predictor $\hat{L}(x) = \langle \hat{w}, x \rangle + \hat{b}$, where $\hat{w} \in \mathbb{R}^l$ and $\hat{b} \in \mathbb{R}$. We then define the nonconformity score as $s(x, y) = |\hat{L}(x) - y|$. We sample n examples from S to construct the prediction set $\widehat{\mathcal{C}}_n(x)$ in Equation (2) and sample m_{test} examples from T to form the test data. We run 1000 times with different random seeds. The results for the coverage (left) and length (right) of the prediction set are presented in box plot form in Figure 1. Here, the coverage is the ratio between the number of test examples such that $y_i \in \widehat{\mathcal{C}}_n(x_i)$ and the size of the test set. The red lines stand for the desired marginal coverages. Since the boxes are below the red coverage lines, we conclude that SCP fails to provide a prediction set with desired coverage when there exists a distributional shift between the source domain and the target domain.

5 Corrected SCP for OOD Data

In this section, we consider correcting SCP for OOD data. We first consider the case in which we have access to the population distributions of the scores from the source domains.

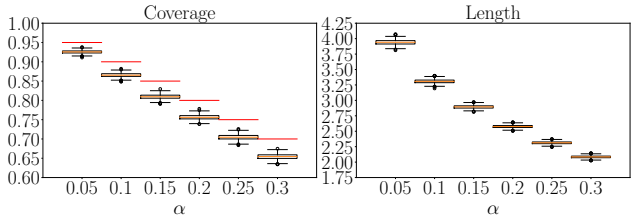

Figure 1: The box plots for the results of the 1000 runs. We show the results for $\alpha = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and the horizontal axis represents the value of α. The left plot shows the results for the coverage of the prediction sets. The red lines are the marginal coverage guarantees that we wish to achieve. The right plot shows the results for the length of the prediction sets.

We then consider the case in which we only have access to the empirical distributions and correct the prediction set to obtain a marginal coverage guarantee in Equation (3).

5.1 Target Distribution Set and Confidence Sets

It is obvious that obtaining a marginal coverage guarantee for an arbitrary target distribution is impossible unless we set $\widehat{\mathcal{C}}(x) = \mathcal{Y}$ for all $x \in \mathcal{X}$, which is a trivial confidence set predictor and does not provide any useful information. In this paper, we consider the case in which the f-divergence (Alfréd 1961) between the target domain and the convex hull of the source domains does not exceed a predefined value. The well-known KL divergence and TV distance are both special cases of f-divergence.

As Section 4 shows, when the target domain differs from the source domain, the marginal coverage does not hold for the predictor (2). Below, we construct new prediction sets where the marginal coverage (3) holds.

We define a set of distributions $\mathcal{T} \subseteq \{Q|Q \text{ is a distribution on } \mathcal{X}\times\mathcal{Y}\}$. For each $T \in \mathcal{T}$, the distribution of the score for the data from T is defined as the push forward distribution $s\#T$, where $(s\#T)(A) = T(s^{-1}(A))$ for any measurable set $A \subseteq \mathbb{R}$. We define the distribution set of the scores as $\mathcal{P} := \{s\#T : T \in \mathcal{T}\}$. For a given $\alpha \in (0,1)$, our goal is to choose a threshold $t \in \mathbb{R}$ such that the confidence set $\widetilde{\mathcal{C}}(x) := \{y \in \mathcal{Y}|s(x,y) \le t\}$ satisfies (3) when (X_{n+1}, Y_{n+1}) is drawn from any target domain $T \in \mathcal{T}$. The following lemma provides a proper choice of t.

Lemma 3. *For any unknown target distribution $T \in \mathcal{T}$, assume that (X_{n+1}, Y_{n+1}) is drawn from T. If we set $t \ge \max_{P\in\mathcal{P}} \mathcal{Q}(1-\alpha; P)$, then:*

$$\mathbb{P}\left(Y_{n+1} \in \widetilde{\mathcal{C}}(X_{n+1})\right) \ge 1-\alpha. \tag{4}$$

For a given set $\mathcal{P}$ of distributions for the score, Lemma 3 reduces the problem of finding a valid confidence set predictor to the following optimization problem:

$$\max\ \mathcal{Q}(1-\alpha; P) \text{ s.t. } P \in \mathcal{P}. \tag{5}$$

Next, we formulate the set $\mathcal{T}$ through the lens of f-divergence.

Definition 4 (f-divergence)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be a closed convex function satisfying $f(1) = 0$ and $f(t) = +\infty$ for $t < 0$. Let P, Q be two probability distributions such that $P \ll Q$ (P is absolutely continuous with respect to Q). The f-divergence between P and Q can then be defined as follows:

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

where $\frac{dP}{dQ}$ is the **Radon-Nikodym derivative** (Patrick 2008).

Remark 1. *For a given function f that satisfies the conditions in Definition 4, define $f_0(t) := f(t) - f'(1)(t-1)$. We then obtain that, for any $P \ll Q$:*

$$D_{f_0}(P\|Q) = \int f_0\left(\frac{dP}{dQ}\right) dQ = D_f(P\|Q).$$

By the convexity of f, it can be easily observed that $f_0(t) \ge 0$ for all $t \in \mathbb{R}$. Moreover, $\inf_t f_0(t) = f_0(1) = 0$ and $f_0'(1) = 0$. Since f_0 produces the same f-divergence as f, without loss of generality, we can assume that $f'(1) = f(1) = 0$ and $f \ge 0$.

Equipped with the f-divergence, we can now define our target distribution set $\mathcal{T}$ for a given threshold $\rho > 0$:

$$\mathcal{T}_{f,\rho}(S_1,\cdots,S_d) := \{T|\exists Q \in \mathcal{CH}(S_1,\cdots,S_d) \text{ s.t. } D_f(T\|Q) \le \rho\}.$$

We omit $S_1, \ldots, S_d$ and use $\mathcal{T}_{f,\rho}$ for simplicity. The corresponding distribution set for the scores is then:

$$\mathcal{P} := \{s\#T | T \in \mathcal{T}_{f,\rho}\}. \tag{6}$$

However, it is hard to obtain the precise relationship between $\mathcal{P}$ and the distributions $s\#S_1, \ldots, s\#S_d$, which makes it difficult to analyze $\mathcal{P}$. We instead consider the following distribution set of scores:

$$\mathcal{P}_{f,\rho} := \{S \text{ is a distribution on } \mathbb{R}| \exists S_0 \in \mathcal{CH}(s\#S_1, \cdots, s\#S_d) \text{ s.t. } D_f(S\|S_0) \le \rho\}. \tag{7}$$

The following lemma reveals the relationship between $\mathcal{P}$ and $\mathcal{P}_{f,\rho}$.

Lemma 5. *Let $\mathcal{P}$, $\mathcal{P}_{f,\rho}$ be defined as in* (6), (7) *respectively. Then, $\mathcal{P} \subseteq \mathcal{P}_{f,\rho}$.*

Remark 2. *According to Lemma 5, $\sup_{P\in\mathcal{P}_{f,\rho}} \mathcal{Q}(1-\alpha; P) \ge \sup_{P\in\mathcal{P}} \mathcal{Q}(1-\alpha; P)$. Lemma 3 accordingly tells us that if we set $t = \sup_{P\in\mathcal{P}_{f,\rho}} \mathcal{Q}(1-\alpha; P)$, then for (X_{n+1}, Y_{n+1}) drawn from any target distribution $T \in \mathcal{T}_{f,\rho}$, we have $\mathbb{P}\left(Y_{n+1} \in \widetilde{\mathcal{C}}(X_{n+1})\right) \ge 1-\alpha$. Our goal is now to solve Problem* (5) *for the set $\mathcal{P}_{f,\rho}$.*

According to Remark 2, we define the worst-case quantile function for the distribution set $\mathcal{P}_{f,\rho}$ as $\widetilde{\mathcal{Q}}(\alpha; \mathcal{P}_{f,\rho}) := \sup_{P\in\mathcal{P}_{f,\rho}} \mathcal{Q}(\alpha; P)$. Remark 2 tells us that taking $t = \widetilde{\mathcal{Q}}(1-\alpha; \mathcal{P}_{f,\rho})$ produces a valid confidence set $\widetilde{\mathcal{C}}$. The next theorem allows us to express the worst-case quantile function in terms of the standard quantile function, which helps us to calculate the worst-case quantile efficiently.

Theorem 6. *Let $F_1, \ldots, F_d$ be the c.d.f.'s of the distributions $s\#S_1, \ldots, s\#S_d$. Define the function $g_{f,\rho} : [0,1] \to [0,1]$ as*

$$g_{f,\rho}(\beta) := \inf\left\{ z \in [0,1] \middle| \beta f\left(\frac{z}{\beta}\right) + (1-\beta) f\left(\frac{1-z}{1-\beta}\right) \leq \rho \right\}$$

and define the inverse of $g_{f,\rho}$ as $g_{f,\rho}^{-1}(\tau) = \sup\{\beta \in [0,1] | g_{f,\rho}(\beta) \leq \tau\}$. Let $F_{\min}(x) := \min_{1\leq i\leq d} F_i(x)$ be a c.d.f., the following holds for all $\alpha \in (0,1)$:

$$\widetilde{\mathcal{Q}}(\alpha; \mathcal{P}_{f,\rho}) = \mathcal{Q}(g_{f,\rho}^{-1}(\alpha); F_{\min}).$$

5.2 Marginal Coverage Guarantee for Empirical Source Distributions

In the previous section, we presented marginal coverage guarantees when we have access to the population distributions of the scores for source domains. However, in practice, it is difficult or even impossible to access these population distributions. In this section, we provide marginal coverage guarantees even when we only have access to the empirical distributions, which is useful in practice.

For any $i \in [d]$, assume we have m_i i.i.d. examples $\{V_{ij} = s(X_{ij}, Y_{ij})\}_{j=1}^{m_i}$ from the source distribution S_i. Further, suppose that $\hat{F}_i$ is the empirical c.d.f. corresponding to F_i, which is defined as $\hat{F}_i(x) = \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbb{I}_{(-\infty,x]}(V_{ij})$. Define $\hat{F}_{\min}(x) = \min_{1\leq i\leq d} \hat{F}_i(x)$. We first provide an error bound when we estimate $F_{\min}$ with $\hat{F}_{\min}$.

Proposition 7. *Let $F_1, \ldots, F_d$ be c.d.f.'s on $\mathbb{R}$, define $F_{\min}(x) = \min_{1\leq i\leq d} F_i(x)$. Suppose $\hat{F}_1, \ldots, \hat{F}_d$ are the empirical c.d.f.'s corresponding to $F_1, \ldots, F_d$, defined with $m_1, \ldots, m_d$ examples, respectively. Define $\hat{F}_{\min}(x) = \min_{1\leq i\leq d} \hat{F}_i(x)$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left| F_{\min}(x) - \hat{F}_{\min}(x) \right| > \epsilon\right) \leq 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2},$$

where the probability is over the randomness of the examples that define the empirical c.d.f.'s.

The above Proposition 7 allows us to quantify the error caused by replacing the population distributions with the empirical distributions, which leads to the following marginal coverage guarantee for the prediction set $\widetilde{\mathcal{C}}$ that we have defined before.

Theorem 8 (Marginal coverage guarantee for the empirical estimations)**.** *Assume $V_{n+1} = s(X_{n+1}, Y_{n+1}) \sim P \in \mathcal{P}_{f,\rho}$ is independent of $\{V_{ij}\}_{i,j=1}^{d,m_i}$ where $\{V_{ij}\}_{j=1}^{m_i} \overset{i.i.d.}{\sim} s\#S_i$ for $i \in [d]$. Suppose $\rho^\star = \inf_{P_0 \in \mathcal{CH}_s} D_f(P\|P_0) \leq \rho$ where $\mathcal{CH}_s = \mathcal{CH}(s\#S_1, \cdots, s\#S_d)$. Let $\hat{F}_{\min}$ be defined as in Proposition 7 and let $\hat{S}_1, \ldots, \hat{S}_d$ be the empirical distributions of $S_1, \ldots, S_d$ respectively. If we set $t = \widetilde{\mathcal{Q}}(1-\alpha; \hat{\mathcal{P}}_{f,\rho}) = \mathcal{Q}(g_{f,\rho}^{-1}(1-\alpha); \hat{F}_{\min})$, then for any $\epsilon > 0$, we obtain the following marginal coverage guarantee for $\widetilde{\mathcal{C}}$:*

$$\mathbb{P}\Big(Y_{n+1} \in \widetilde{\mathcal{C}}(X_{n+1})\Big) \geq \left(1 - 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}\right) g_{f,\rho^\star}\Big(g_{f,\rho}^{-1}(1-\alpha) - \epsilon\Big),$$

where the randomness is over the choice of the source examples and (X_{n+1}, Y_{n+1}) and

$$\hat{\mathcal{P}}_{f,\rho} := \Big\{ S \Big| \exists S_0 \in \mathcal{CH}(s\#\hat{S}_1, \cdots, s\#\hat{S}_d) \text{ s.t. } D_f(S\|S_0) \leq \rho \Big\}.$$

By Lemma 14 in the Appendix, $g_{f,\rho}(\beta)$ is non-increasing in ρ and non-decreasing in β, so $g_{f,\rho^\star}(g_{f,\rho}^{-1}(1-\alpha)-\epsilon) \geq g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)-\epsilon)$. In practice, we do not know $\rho^\star$, so we use $g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)-\epsilon)$ instead. Since $g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)-\epsilon) \leq g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)) = 1-\alpha$, we get guaranteed coverage $\left(1 - 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}\right) g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)-\epsilon) \leq 1-\alpha$. To achieve a marginal coverage with the level of at least $1-\alpha$, we need to correct the output set by replacing α with some $\alpha' < \alpha$ when running our confidence set predictor. The following corollary tells us how to choose α' to correct the prediction set.

Corollary 9 (Correct the prediction set to get a $(1-\alpha)$ marginal coverage)**.** *Let (X_{n+1}, Y_{n+1}), $\hat{F}_{\min}$, $\hat{\mathcal{P}}_{f,\rho}$ be defined as in Theorem 8. For arbitrary $\epsilon > 0$, if we set $t = \widetilde{\mathcal{Q}}(1-\alpha'; \hat{\mathcal{P}}_{f,\rho}) = \mathcal{Q}\left(g_{f,\rho}^{-1}(1-\alpha'); \hat{F}_{\min}\right)$, where*

$$\alpha' = 1 - g_{f,\rho}\left(\epsilon + g_{f,\rho}^{-1}\left(\frac{1-\alpha}{1 - 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}}\right)\right),$$

then we obtain the following marginal coverage guarantee:

$$\mathbb{P}\left(Y_{n+1} \in \widetilde{\mathcal{C}}(X_{n+1})\right) \geq 1 - \alpha.$$

Remark 3. *Corollary 9 tells us that we can take $t = \mathcal{Q}\left(g_{f,\rho}^{-1}(1-\alpha'); \hat{F}_{\min}\right) = \mathcal{Q}\left(\epsilon + g_{f,\rho}^{-1}\left(\frac{1-\alpha}{1-2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}}\right); \hat{F}_{\min}\right)$ to get a marginal coverage guarantee with confidence level $1-\alpha$. When $f(\cdot), s(\cdot,\cdot)$ are chosen and the numbers of examples that are used to estimate the source distributions, i.e., $m_1, \ldots, m_d$, are given, we solve the following optimization problem to find a desired t.*

$$\min_{0<\epsilon\leq 1} \ \mathcal{Q}\left(\epsilon + g_{f,\rho}^{-1}\left(\frac{1-\alpha}{1 - 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}}\right); \hat{F}_{\min}\right),$$
$$\text{s.t. } \epsilon + g_{f,\rho}^{-1}\left(\frac{1-\alpha}{1 - 2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}}\right) \leq 1.$$

Since the quantile function $\mathcal{Q}\left(\cdot; \hat{F}_{\min}\right)$ is non-decreasing, let $h(\epsilon) = \epsilon + g_{f,\rho}^{-1}\left(\frac{1-\alpha}{1-2\sum_{i=1}^{d} e^{-2m_i\epsilon^2}}\right)$, we solve the following problem instead:

$$\min \ \ h(\epsilon) \text{ s.t. } 0 < \epsilon \leq 1, h(\epsilon) \leq 1.$$

For some choices of f, the functions $g_{f,\rho}$ and $g_{f,\rho}^{-1}$ have closed forms (please refer to the examples in Section 5.3). For general f that we do not have a closed form of $g_{f,\rho}^{-1}$, the following lemma tells us that we can use a binary search algorithm to efficiently compute the value of $g_{f,\rho}^{-1}(\tau)$ for a given τ.

Lemma 10 ((Cauchois et al. 2020), The form of $g_{f,\rho}^{-1}$ that can be efficiently solved)**.** *Let $g_{f,\rho}, g_{f,\rho}^{-1}$ be defined as in Theorem 6. Then, for any $\tau \in [0,1]$, we have:*

$$g_{f,\rho}^{-1}(\tau) = \sup\left\{\beta \in [\tau,1] \,\middle|\, \beta f\left(\frac{\tau}{\beta}\right) + (1-\beta) f\left(\frac{1-\tau}{1-\beta}\right) \le \rho\right\}.$$

5.3 Examples

In this section, we present some examples of calculating $g_{f,\rho}$ and $g_{f,\rho}^{-1}$ for some important f-divergences.

Example 1 (χ^2-divergence)**.** Let $f(t) = (t-1)^2$; then, $D_f(P\|Q) = \mathbb{E}_Q\left[\left(\frac{dP}{dQ}-1\right)^2\right] = \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^2 - 1\right]$ is the χ^2-divergence. In this case, we have $g_{f,\rho}(\beta) = \left(\beta - \sqrt{\rho\beta(1-\beta)}\right)_+$, where $(x)_+ = \max\{0,x\}$. $g_{f,\rho}^{-1}(\tau)$ is the solution of the following optimization problem:

$$\max \beta \text{ s.t. } \begin{cases} \frac{\rho}{\rho+1} \le \beta \le 1 \\ \beta - \sqrt{\rho\beta(1-\beta)} \le \tau \end{cases}.$$

Example 2 (Total variation distance, (Cauchois et al. 2020))**.** Let $f(t) = \frac{1}{2}|t-1|$; then, $D_f(P\|Q) = \mathbb{E}_Q\left[\frac{1}{2}\left|\frac{dP}{dQ}-1\right|\right]$ is the total variation distance. In this case, we can provide analytic forms for $g_{f,\rho}$ and $g_{f,\rho}^{-1}$:

$$g_{f,\rho}(\beta) = (\beta-\rho)_+, \quad g_{f,\rho}^{-1}(\tau) = \min\{\tau+\rho, 1\}.$$

Example 3 (Kullback-Leibler divergence)**.** Let $f(t) = t\log t$; then, $D_f(P\|Q) = \mathbb{E}_Q\left[\frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)\right]$ is the Kullback-Leibler (KL) divergence (Solomon and Richard A 1951). Unfortunately, we cannot provide the analytic forms of $g_{f,\rho}$ and $g_{f,\rho}^{-1}$ for KL-divergence. Fortunately, according to Theorem 6 and Remark 3, we can compute the values $g_{f,\rho}(\beta)$ and $g_{f,\rho}^{-1}(\tau)$ by solving a one-dimensional convex optimization problem, which can be solved efficiently using binary search.

6 Experiments

In this section, we use simulated data to verify our theory and the validity of our constructed confidence set predictor (referred to as **OOD-SCP** in the remainder of this paper). We consider two cases: first, we verify the validity of OOD-SCP using the same settings as in Section 4; then, we construct a multi-source OOD confidence set prediction task and show that OOD-SCP is valid for this task.

According to Figure 2, unlike standard SCP, for all values of α, the violin for OOD-SCP is above the desired coverage line, which shows that OOD-SCP is empirically valid.

We next consider a multi-source OOD confidence set prediction task. Similar to Section 4, we consider the regression

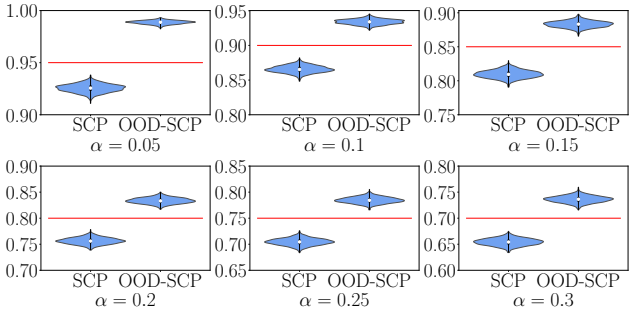

Figure 2: The violin plots for the coverage of the 1000 runs under the same data generation settings as in Section 4. We show results for $\alpha = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. Here, the red lines are the marginal coverage guarantees that we wish to achieve. The white point represents the median, while the two endpoints of the thick line are the 0.25 quantile and the 0.75 quantile.

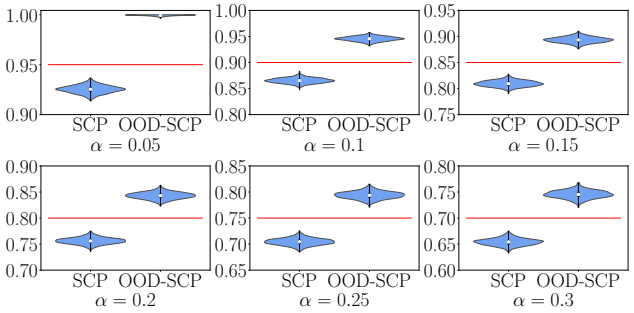

Figure 3: The violin plots for the coverage of the 1000 runs for the multi-source OOD confidence set prediction task. We show results for $\alpha = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. Here, the red lines are the marginal coverage guarantees that we wish to achieve. The white point represents the median, while the two endpoints of the thick line are the 0.25 quantile and the 0.75 quantile.

problem and set $\mathcal{X} = \mathbb{R}^l, \mathcal{Y} = \mathbb{R}$. Define the oracle linear predictor $L : \mathcal{X} \to \mathcal{Y}$ as $L(x) = \langle w^\star, x\rangle + b^\star$, where $w^\star \in \mathbb{R}^l$ and $b^\star \in \mathbb{R}$. We define the marginal distribution of X for the source domains S_1 and S_2 as

$$S_{1X} = \mathcal{N}(\mu_1, \sigma_{s,x}^2 I_l), \quad S_{2X} = \mathcal{N}(\mu_2, \sigma_{s,x}^2 I_l)$$

respectively, where $\mu_1, \mu_2 \in \mathbb{R}^l$ are the mean vectors, $\sigma_{s,x} > 0$ is a scalar, and $I_l \in \mathbb{R}^{l\times l}$ is the identity matrix with dimension $l \times l$. We define $Y|X = x \sim \mathcal{N}(L(x), \sigma_{s,y}^2)$ for both S_1 and S_2. For the target domain T, we define the marginal distribution of X as $T_X = \frac{S_{1X}+S_{2X}}{2}$ and the conditional distribution of Y given X as $Y|X = x \sim \mathcal{N}(L(x), \sigma_{t,y}^2)$. Here, $\sigma_{s,y}, \sigma_{t,y} > 0$ are the standard deviations and $\sigma_{s,y} \neq \sigma_{t,y}$.

Similar to Section 4, we sample $\frac{m_{\text{train}}}{2}$ examples from S_1 and $\frac{m_{\text{train}}}{2}$ examples from S_2 to train a linear predictor $\hat{L}(x) = \langle \hat{w}, x\rangle + \hat{b}$, where $\hat{w} \in \mathbb{R}^l$ and $\hat{b} \in \mathbb{R}$. We then define the nonconformity score as $s(x,y) = |\hat{L}(x) - y|$. We sample

$\frac{n}{2}$ examples from S_1, and $\frac{n}{2}$ examples from S_2 to construct the prediction set $\widetilde{\mathcal{C}}(x)$ and sample m_{test} examples from T to form the test data.

Figure 3 shows the results for the multi-source OOD confidence set prediction task. From the figure, we can see that the violins for the standard SCP are under the desired coverage lines, which means that the standard SCP is invalid in this case. By contrast, the violins for OOD-SCP are above the desired coverage lines, indicating that OOD-SCP is valid, which validates Corollary 9.

The reason we do not do experiments on real datasets is that do not know how to set the value of ρ for the existing OOD datasets. Our main claim is that **when the target domain satisfies $T \in \mathcal{T}_{f,\rho}$, the coverage of our method is guaranteed**. However, we claim **it is acceptable**. In many fields, we face the same problem.

In adversarial robustness (Szegedy et al. 2014), the theories (for example (Montasser, Hanneke, and Srebro 2019)) provide an upper bound of the test adversarial robustness $\mathop{\mathbb{P}}_{(x,y)\sim D}[\exists\|\delta\| \leq \epsilon : h(x+\delta) \neq y]$, where h is a classifier. The results just tell us that we have the guarantee for the test accuracy **if the test perturbation δ satisfies $\|\delta\| \leq \epsilon$**. However, what if $\|\delta\| > \epsilon$? It is out of the scope of their theories.

For distributional robustness optimization (DRO), the theories (Lee and Raginsky 2018) prove that **if the test distribution is in a Wasserstein ball with radius** r, then the test risk can be upper bounded. Formally, $\max_{D\in W(r)} \mathop{\mathbb{P}}_{(x,y)\sim D}[h(x) \neq y]$ is upper bounded, where $W(r)$ is a Wasserstein ball with radius. They do not know how to set r to make $W(r)$ contain the test distribution either, however, this does not overshadow their contribution to the DRO community. In other words, the issues of ρ do not overshadow our contribution to the OOD community.

7 Discussions

Our work is an extension of (Cauchois et al. 2020) to the multi-domain case. In this section, we discuss the differences between our work and (Cauchois et al. 2020).

7.1 The Necessity of Our Extension

In the multi-source setting, to make use of all the source domains, a trivial method is to regard the mixture of the source domains, as a domain $S = \sum_{i=1}^{d} \lambda_i S_i$ and use the method in (Cauchois et al. 2020). However, there are two issues:

- Given the empirical data from $S_1, \ldots, S_d$, we don't know the exact values of $\lambda_1, \ldots, \lambda_d$ for the mixed domain, so we don't know the set $\widetilde{\mathcal{P}} = \{s\#T | D_f(T\|S) \leq \rho\}$ for a given ρ. So we don't know the set that we are giving a coverage guarantee for.
- We may be not able to provide a coverage guarantee for data from one of the source domains. Take KL-divergence as an example, then drawing from S can be regarded as first drawing an index I from λ and then drawing an example from S_I. S_i can be seen as drawn from the same process with $\lambda = e_i$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^d$ with only the i-th element being 1. By the chain rule, $KL(S_i\|S) = KL(e_i\|\boldsymbol{\lambda}) + \mathbb{E}_{j\sim e_i} KL(S_j\|S_j) = \log(1/\lambda_i)$. If $\rho < \max_i \log(1/\lambda_i)$, then there exists a S_i s.t. $KL(S_i\|S) > \rho$, i.e., **we can not even get a coverage guarantee for the source domain S_i, which is unacceptable!** The problem gets worse if the source domain number d becomes larger since $\max_i \log(1/\lambda_i) \geq \log d$. However, in our generalization, there is no such problem even if we choose $\rho = 0$, which means **the method in (Cauchois et al. 2020) is not compatible with the multi-source setting, so our extension is necessary**.

7.2 The Difference in Proof Skills

In fact, our Theorem 6 is an extension of (Cauchois et al. 2020) to the OOD setting and mainly depends on Lemmas 17 and 18. Lemma 17 helps us reduce multi-input $g_{f,\rho}$ to a single-input case. The main idea of the proof of Lemma 18 comes from the argument in (Cauchois et al. 2020), however, **the extension is non-trivial**. In Lemma 18, let $h(z,\beta) = \beta f(z/\beta) + (1-\beta) f((1-z)/(1-\beta))$ we use the multi-input $g_{f,\rho}(\beta_1, \cdots, \beta_d) = \inf\{z \in [0,1] | \inf_{\lambda\in\Delta^{d-1}} h(z, \sum_{i=1}^{d} \lambda_i\beta_i) \leq \rho\}$, which involves taking infimum w.r.t. λ and is much more complicated than the single-input case in (Cauchois et al. 2020). We construct a set $\mathcal{P}^*_{f,\rho}$ that is more complicated than that in (Cauchois et al. 2020) and the proof is more difficult. Moreover, due to multiple inputs and the $\inf_{\lambda\in\Delta^{d-1}}$ operator, we need to consider 4 cases according to whether each $F_i(t)$ is 0 or 1.

Our Theorem 8 and its corresponding Corollary 9 are novel and quite different from Corollaries 2.1 and 2.2 in (Cauchois et al. 2020). The common point is that they all consider finite sample approximation. The proof of Corollary 2.1 in (Cauchois et al. 2020) relies on the exchangeability of the source examples, however, in the OOD setting, examples are drawn from different source domains and are not exchangeable. So the analysis techniques in (Cauchois et al. 2020) can not be applied in our case. To fill this gap, we use the decomposition technique and concentration inequalities.

8 Conclusion

We study the confidence set prediction problem in the OOD generalization setting. We first empirically show that SCP is not valid in the OOD generalization setting. We then develop a method for forming valid confident prediction sets in the OOD setting and theoretically prove the validity of our proposed method. Finally, we conduct experiments on simulated data to empirically verify both the correctness of our theory and the validity of our proposed method.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant 2023YFC3604702, the National Natural Science Foundation of China under Grant 61976161, the Fundamental Research Funds for the Central Universities under Grant 2042022rc0016.

References

Achim, K. 2013. *Probability theory: a comprehensive course*. Springer Science & Business Media.

Alfréd, R. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, 547–562.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *CoRR*, abs/1606.06565.

Angelopoulos, A. N.; Bates, S.; Jordan, M. I.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *ICLR*.

Bin, Y. 1994. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 94–116.

Cauchois, M.; Gupta, S.; Ali, A.; and Duchi, J. C. 2020. Robust Validation: Confident Predictions Even When Distributions Shift. *CoRR*, abs/2008.04267.

Fisch, A.; Schuster, T.; Jaakkola, T. S.; and Barzilay, R. 2021. Few-Shot Conformal Prediction with Auxiliary Tasks. In *ICML*, 3329–3339.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, 17: 59:1–59:35.

Gendler, A.; Weng, T.; Daniel, L.; and Romano, Y. 2022. Adversarially Robust Conformal Prediction. In *ICLR*.

Gibbs, I.; and Candès, E. J. 2021. Adaptive Conformal Inference Under Distribution Shift. In *NeurIPS*, 1660–1672.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

Jiang, H.; Kim, B.; Guan, M. Y.; and Gupta, M. R. 2018. To Trust Or Not To Trust A Classifier. In *NeurIPS*, 5546–5557.

Jiang, X.; Osl, M.; Kim, J.; and Ohno-Machado, L. 2012. Calibrating predictive model estimates to support personalized medicine. *J. Am. Medical Informatics Assoc.*, 19(2): 263–274.

Lee, J.; and Raginsky, M. 2018. Minimax Statistical Learning with Wasserstein distances. In *NeurIPS*, 2692–2701.

Li, D.; Yang, Y.; Song, Y.; and Hospedales, T. M. 2018a. Learning to Generalize: Meta-Learning for Domain Generalization. In *AAAI*, 3490–3497.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain Generalization With Adversarial Feature Learning. In *CVPR*, 5400–5409.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018c. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *ECCV*, volume 11219, 647–663.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Montasser, O.; Hanneke, S.; and Srebro, N. 2019. VC Classes are Adversarially Robustly Learnable, but Only Improperly. In *COLT*, 2512–2530.

Nagarajan, V.; Andreassen, A.; and Neyshabur, B. 2021. Understanding the failure modes of out-of-distribution generalization. In *ICLR*.

Oliveira, R. I.; Orenstein, P.; Ramos, T.; and Romano, J. V. 2022. Split Conformal Prediction for Dependent Data. arXiv:2203.15885.

Patrick, B. 2008. *Probability and measure*. John Wiley & Sons.

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.*, 9: 371–421.

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards Out-Of-Distribution Generalization: A Survey. *CoRR*, abs/2108.13624.

Solomon, K.; and Richard A, L. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *ECCV*, volume 9915, 443–450.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Tibshirani, R. J.; Barber, R. F.; Candès, E. J.; and Ramdas, A. 2019. Conformal Prediction Under Covariate Shift. In *NeurIPS*, 2526–2536.

Vovk, V. 2013. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2-3): 349–376.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag. ISBN 0387001522.

Wang, Q.; Wang, Y.; Zhu, H.; and Wang, Y. 2022. Improving Out-of-Distribution Generalization by Adversarial Training with Structured Priors. *CoRR*, abs/2210.06807.

Xiaohong, C.; Lars Peter, H.; and Marine, C. 2010. Nonlinearity and temporal dependence. *Journal of Econometrics*, 155: 155–169.

Xin, S.; Wang, Y.; Su, J.; and Wang, Y. 2022. Domain-wise Adversarial Training for Out-of-Distribution Generalization.

Zou, X.; and Liu, W. 2023. On the Adversarial Robustness of Out-of-distribution Generalization Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.