Certifiable Out-of-Distribution Generalization

Nanyang Ye[1]*, Lin Zhu[1], Jia Wang[2], Zhaoyu Zeng[1], Jiayao Shao[3], Chensheng Peng[1], Bikang Pan[4], Kaican Li[5], Jun Zhu[6]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] University of Cambridge, Cambridge, United Kingdom
[3] University of Warwick, Warwick, United Kingdom
[4] ShanghaiTech University, Shanghai, China
[5] Huawei Noah's Ark Lab, Hong Kong, China
[6] Tsinghua University, Beijing, China
ynylincoln@sjtu.edu.cn, zhulin_sjtu@sjtu.edu.cn, jw2054@cam.ac.uk, zlatanwilliams@sjtu.edu.cn, jiayao_shao@163.com, pesiterswift@sjtu.edu.cn, panbk@shanghaitech.edu.cn, mjust.lkc@gmail.com dcszj@mail.tsinghua.edu.cn

Abstract

Machine learning methods suffer from test-time performance degeneration when faced with out-of-distribution (OoD) data whose distribution is not necessarily the same as training data distribution. Although a plethora of algorithms have been proposed to mitigate this issue, it has been demonstrated that achieving better performance than ERM simultaneously on different types of distributional shift datasets is challenging for existing approaches. Besides, it is unknown how and to what extent these methods work on any OoD datum without theoretical guarantees. In this paper, we propose a certifiable out-of-distribution generalization method that provides provable OoD generalization performance guarantees via a functional optimization framework leveraging random distributions and max-margin learning for each input datum. With this approach, the proposed algorithmic scheme can provide certified accuracy for each input datum's prediction on the semantic space and achieves better performance simultaneously on OoD datasets dominated by correlation shifts or diversity shifts. Our code is available at https://github.com/ZlatanWilliams/StochasticDisturbanceLearning.

Introduction

Deep learning has achieved success in various domains, including computer vision and natural language processing. However, traditional algorithms only exhibit human-superior behaviors towards datasets that are independent and identically distributed (i.i.d.) (Silver et al. 2016), while the performance degenerates when exposed to out-of-distribution (OoD) data. This precludes many applications, especially in high-risk sectors, such as healthcare, autonomous driving and securities, where the distribution shift between training and testing data is ubiquitous, and the impact of machines' mistakes is severe.

Many previous works have been done for OoD generalization with empirical performance improvements (Ye et al. 2021; Ahuja et al. 2020a). However, due to the complexity of the OoD generalization problem where the model has to generalize across various unseen domains, it still remains largely unsolved. For example, it was recently found few methods can outperform the empirical risk minimization (ERM) with extensive data augmentation (Gulrajani and Lopez-Paz 2021). It has also been demonstrated that the proposed OoD generalization algorithms exhibit preferences on one type of distribution shift and fail on the other. Besides, many empirically well-performed algorithms are provided without performance guarantees (Ye et al. 2021; Wiles et al. 2022).

All this begs the following question:

Can we have a certifiable OoD generalization algorithm that may work well under multiple types of distribution shifts?

Previous pioneering work (Zhao et al. 2019; Ben-David et al. 2010) proposed upper bounds for generalization performances, while the bounds are not computable. In this paper, we propose the Stochastic Disturbance Learning (SDL) algorithm aiming to achieve SOTA performance for different kinds of distribution shifts that certifiably provide correct predictions within some sets around the input data in the *semantic space*. To achieve this, we derive the performance guarantee by measuring the performance of deep neural networks (DNNs) under random disturbances via a functional optimization framework, that is not reliant on the local convexity assumption of DNNs. Our main contributions are as follows:

1. We propose an OoD generalization algorithm that achieves better performances than ERM on OoD datasets dominated by correlation shifts and diversity shifts simultaneously. See Section for definitions of the mentioned shifts.
2. We provide a certification methodology with max-margin learning, that can provide theoretical performance guarantees. Ablation studies confirm the effectiveness of algorithmic components. The theoretical analysis also helps explain why the commonly-used dropout method can help improve DNN's generalization abilities.

Preliminaries

This section briefly summarizes the literature for the OoD generalization methods and introduces the topics that motivate the proposed algorithmic framework.

*Corresponding author. The full Appendix is available on Arxiv.

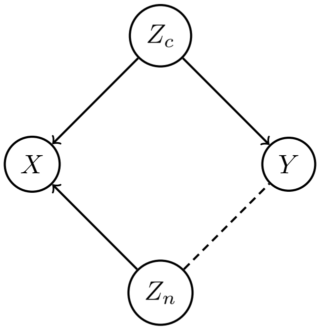


Figure 1: The causal graph for the mentioned random variables. The directed arrow (resp. dotted line) indicates causal (resp. confounding) relations between two random variables. More specifically, the directed arrow from A to B indicates a causal path from A to B.

OoD Generalization OoD generalization is the task of generalizing a model's performance under distribution shifts between the training and the unseen testing distributions. This is in stark contrast with adversarial defense, where the objective is to have a robust classifier against steered minor perturbations added to the images, which resemble noises in the images. The OoD generalization focuses on generalizing to data subsuming the same semantic information for classification but with different environments or styles information, arguably more commonly observed in practical scenarios than deliberate adversarial attacks. For example, the system has to generalize to unseen environments for safety in autonomous driving. Existing OoD generalization algorithms can be generally divided into four genres: domain generalization-based methods that focus on learning coherent patterns in data collected from diverse environments (Huang et al. 2020; Li et al. 2018b; Ganin et al. 2016); invariant learning-based methods that exclude spurious correlations that exist in the data (Arjovsky et al. 2019; Ahuja et al. 2020b, 2021); distributionally robust optimization methods that fabricate challenging data distributions from the original (Sagawa* et al. 2020); causal learning-based methods that leverage causal inference techniques (Shen et al. 2020b,a; Tran et al. 2016). These methods have demonstrated empirical improvements on OoD generalization tasks, while their theoretical performance on OoD generalization is largely untapped.

Diversity and Correlation Shifts It was recently found that multiple dimensions exist in OoD generalization datasets where algorithms typically perform better than ERM on one dimension but not as well on the other (Wiles et al. 2022; Ye et al. 2021). In (Ye et al. 2021), these dimensions are delineated as diversity shift or correlation shift. The *diversity shift* is formally defined as the difference in the environmental semantic feature's training and testing probability density functions (p.d.f.s) on the overall differences between two distributions' supports[1]. In comparison, the *correlation shift* is defined as the difference in training and testing marginal p.d.f.s over the environmental semantic feature on the intersection of training and testing distributions' supports. For example, consider data pairs (X, Y), where X is the covariate variable and Y is a dependent variable (*e.g*. label), we wish to model from X (i.e., we wish to find the function $f : \mathcal{X} \to \mathcal{Y}$). $\mathcal{Z}$ is defined to be the semantic/feature/latent[2] space with $Z_c \subset \mathcal{Z}$ and $Z_n \subset \mathcal{Z}$ be vectors of latent (unobserved) confounding causal and non-causal random variables. The causal graph in Figure 1 is an illustration of the settings described above. In other words, we can interpret X being caused by Z_c and Z_n and Y is determined by Z_c only. The reason for assuming this model is that in most real-world distribution shifts, the interpretation is valid and produces productive outcomes. Intuitively, The diversity distribution shifts summarise the traits of novel features between the training and testing distributions. In contrast, the correlation shifts summarise the spurious correlation between label Y and non-causal features Z_n, which can also be interpreted as the environmental semantic feature. It was found that very few methods can achieve better performance than ERM concurrently on two kinds of OoD shifts (Ye et al. 2021; Huang et al. 2022).

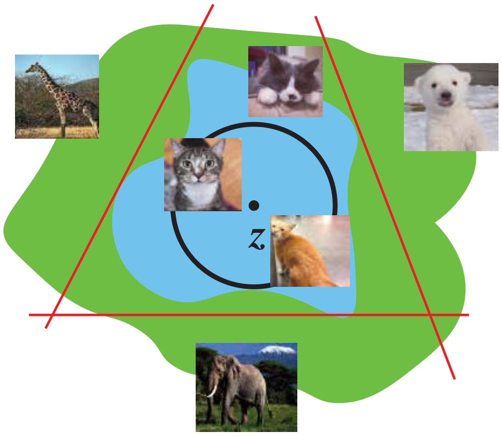


Figure 2: Summary of certifiable out-of-distribution generalization. The black center point is the living input datum in the semantic space, which should be classified as cats. The black circle denotes the certifiable range, inside which data are classified as cats with theoretical guarantees with our method. The blue region outside the black circle is the semantic space where input samples are also classified as cats but may have no guarantees. The green region is the semantic space where input samples are classified as other types, such as bears, elephants, and giraffes. The red lines denote the model's decision boundary shaped by the proposed max-margin training method to separate different categories.

Theoretical Results in OoD Generalization As mentioned in the previous paragraphs, there is no theoretical approach to solve the OoD generalization problem com-

[1] A function f's support is informally interpreted as the subset of the domain where f takes non-zero values.

[2] Note that semantic feature $\mathcal{Z}$ will be formally defined in the following section and could be viewed equivalently as latent space. This differs from some other papers where semantic feature means causal feature.

pletely, and existing theories mostly require assumptions and optimization under constraints. One of the most popular directions is the distributionally robust optimization (DRO) (Rahimian and Mehrotra 2019), which is under the robustness optimization framework. The ethics behind the DRO is minimizing the worst-case risk over an uncertainty distribution set centered on the training distribution. Under this scenario, there exhibits freedom of choosing sensible distance measure (*e.g.* f-Divergence (Weber et al. 2022; Duchi and Namkoong 2019), Wasserstein Distance (Gao and Kleywegt 2016; Sinha, Namkoong, and Duchi 2018), MMD (Staib and Jegelka 2019)) to define the uncertainty set which leads to various angles to tackle the OoD generalization problem. In (Gao and Kleywegt 2016), bounds for the worst-case risk are obtained under the minimum assumption of loss function l being bounded for any black-box machine learning functions. Another direction is Invariance-Based Optimization (Rojas-Carulla et al. 2018; Koyama and Yamaguchi 2020b) which defines an optimization problem based on information theory. Specifically, the Shannon mutual information (Cover and Thomas 2006) between two random variables is optimized under an invariance set. Despite the success of theoretical certificates, the limitations of using these optimization models are that these results are only non-trivial under a small ball of distribution shifts and the bounds derived are not numerically computable. (Weber et al. 2022). Sometimes strong assumptions are needed to be imposed on the loss function l or the machine learning models to guarantee validity (Gao and Kleywegt 2016; Sinha, Namkoong, and Duchi 2018; Staib and Jegelka 2019). These limitations have made these models hard to apply in real-world data where the distribution shifts are generally large and provide theoretical guarantees to existing algorithms.

Proposed Methodology

In this section, we will first introduce the certification methodology for providing certifiable guarantees for each input datum's prediction. Then, based on the theoretical results, the max-margin training is proposed and analyzed via the neural tangent kernel theory to improve the certification bounds. A summary of certifiable OoD generalization is illustrated in Figure 2. Later, we propose an instantiation of practical certifiable OoD generalization algorithms.

Certification Methodology

Notations and Settings We use $(\mathbf{X}, \mathbf{Y})$ to denote the dataset with n data pairs, where $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^n$. We denote $f_{\boldsymbol{\theta}}(\cdot) = f(\cdot;\boldsymbol{\theta}) = f_{L-1} \circ f_{L-2} \cdots \circ f_0$ as a L-layer DNN with the last layer as the classification layer, where $\boldsymbol{\theta} \in \mathcal{R}^{\Theta}$ is the parameters of the DNN. We also use $f(\cdot)$ for short in the next paragraphs. For simplicity and without loss of generality, we consider a 0-1 classification problem [3] and the range of f's output lies in $[0, 1]$. As previous research revealed that the intermediate representations learned by DNN exhibit semantic features of objects to be recognized (Zeiler and Fergus 2013), given an input datum $\mathbf{x}$, we define the semantic representation as the concatenation intermediate layers' representations: $\mathbf{z} = [f_0(\mathbf{x}), f_1(\mathbf{x}), \cdots, f_{L-2}(\mathbf{x})]$. Without loss of generality, we want to certify that if $f(\mathbf{z}) > 1/2$, then, it still holds for some set $\mathcal{B}$ around $\mathbf{z}$, i.e., for any change in the semantic information $\boldsymbol{\delta} \in \mathcal{B}$, $f(\mathbf{z} \otimes \boldsymbol{\delta}) \geq 1/2$, where $\otimes$ is an operator that is addition or multiplication. Formally, we have the following $\mathcal{B}$-Generalizable definition for OoD generalization:

Definition 1. **($\mathcal{B}$-Generalizable)** For a 0-1 classification problem with notations inherited from above where $f(\mathbf{z}) \in [0, 1]$, given $\mathcal{B}$ a closed set and $f(\mathbf{z}) > 1/2$, we say the function f is $\mathcal{B}$-Generalizable at $\mathbf{z}$, if for any perturbation $\boldsymbol{\delta}$ in the $\mathcal{B}$ set i.e., $\boldsymbol{\delta} \in \mathcal{B}$, $f(\mathbf{z} \otimes \boldsymbol{\delta}) > 1/2$.

Remark: For convenience we will simply write *"$f(\mathbf{z})$ is $\mathcal{B}$-Generalizable"*. For clarity, we posit the proposed definition for [0,1] classification problem. This can be also extended for multi-class cases. Next, we will introduce certifiable methods to elicit $\mathcal{B}$-Generalizable models and corresponding $\mathcal{B}$.

In the following, we introduce the proposed random disturbed version of the model. Assuming π_0 is the distribution of the stochastic disturbance, the randomized model is defined as the expectation of prediction averaged over the distribution of the semantic representation:

$$f_{\pi_0}(\mathbf{z}) := \mathbb{E}_{\boldsymbol{\eta}\sim\pi_0}[f(\mathbf{z} \otimes \boldsymbol{\eta})] \quad (1)$$

We want to prove that if the original classifier still gives the correct prediction under stochastic disturbance ($f_{\pi_0}(\mathbf{z}) > 1/2$), then for any perturbation within some range $\boldsymbol{\delta} \in \mathcal{B}$, the following inequality still holds:

$$\begin{aligned}\min_{\boldsymbol{\delta}\in\mathcal{B}} f_{\pi_0}(\mathbf{z} \otimes \boldsymbol{\delta}) &= \min_{\boldsymbol{\delta}\in\mathcal{B}} f_{\pi_{\boldsymbol{\delta}}}(\mathbf{z}) \\ &= \min_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{\eta}\sim\pi_0}[f(\mathbf{z} \otimes \boldsymbol{\eta} \otimes \boldsymbol{\delta})] > 1/2 \quad (2)\end{aligned}$$

where $\otimes$ is element-wise addition or element-wise multiplication and $\pi_{\boldsymbol{\delta}}$ is the distribution of $\boldsymbol{\eta} \otimes \boldsymbol{\delta}$. To derive a tractable lowerbound of $\min_{\boldsymbol{\delta}\in\mathcal{B}} f_{\pi_0}(\mathbf{z} \otimes \boldsymbol{\delta})$, we further relax f to the functional space $\mathcal{F} = \{\hat{f} : \hat{f}(\mathbf{z}) \in [0, 1], \forall \mathbf{z} \in \mathbb{R}^Z\}$ that is the set of all functions bounded in $[0, 1]$, along with a equality constraint at the original function f:

$$\begin{aligned}&\min_{\boldsymbol{\delta}\in\mathcal{B}} f_{\pi_0}(\mathbf{z} \otimes \boldsymbol{\delta}) \geq \min_{\hat{f}\in\mathcal{F}} \left\{ \min_{\boldsymbol{\delta}\in\mathcal{B}} \hat{f}_{\pi_0}(\mathbf{z} \otimes \boldsymbol{\delta}) \right\} \quad (3) \\ &\text{s.t. } \hat{f}_{\pi_0}(\mathbf{z}) = f_{\pi_0}(\mathbf{z})\end{aligned}$$

Remark 1: To make the lowerbound computable, as f is a high-dimensional non-linear function (deep neural networks) that is generally intractable, we further relax it to any function bounded in $[0, 1]$ but with an extra constraint that it should give the same prediction as the original function. Note that the above inequality can be solved with the Lagrangian method.

Theorem 1. **(Lagrangian)** Denote by $\pi_{\boldsymbol{\delta}}$ the distribution of $\boldsymbol{\eta} \otimes \boldsymbol{\delta}$, solving Inequality 3 is equivalent to solving the following problem:

$$\begin{aligned}\mathcal{L} &= \min_{\hat{f}\in\mathcal{F}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \max_{\lambda\in\mathbb{R}} \left\{ \hat{f}_{\pi_0}(\mathbf{z} \otimes \boldsymbol{\delta}) - \lambda(\hat{f}_{\pi_0}(\mathbf{z}) - f_{\pi_0}(\mathbf{z})) \right\} \\ &\geq \max_{\lambda\geq 0} \left\{ \lambda f_{\pi_0}(\mathbf{z}) - \max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_0, \pi_{\boldsymbol{\delta}}) \right\} \quad (4)\end{aligned}$$

[3] Indeed, a multi-class classification problem can be seen as multiple 0-1 classification problems (one versus others).

where $\mathbb{D}_{\mathcal{F}}(\lambda\pi_0, \pi_{\boldsymbol{\delta}})$ is:

$$\begin{aligned}&\mathbb{D}_{\mathcal{F}}(\lambda\pi_0, \pi_{\boldsymbol{\delta}})\\&= \max_{\hat{f}\in\mathcal{F}} \left\{\lambda\mathbb{E}_{\boldsymbol{\eta}\sim\pi_0}[\hat{f}(\mathbf{z}\otimes\boldsymbol{\eta})] - \mathbb{E}_{\boldsymbol{\eta}\sim\pi_{\boldsymbol{\delta}}}[\hat{f}(\mathbf{z}\otimes\boldsymbol{\eta})]\right\} \quad (5)\\&= \begin{cases}\int[\lambda\pi_0(\boldsymbol{\eta}) - \pi_{\boldsymbol{\delta}}(\boldsymbol{\eta})]_+ \mathrm{d}\boldsymbol{\eta}, \text{if } \pi \text{ is continuous,}\\ \sum[\lambda\pi_0(\boldsymbol{\eta}) - \pi_{\boldsymbol{\delta}}(\boldsymbol{\eta})]_+, \text{if } \pi \text{ is discrete.}\end{cases}\end{aligned}$$

Remark 2: The proof of Theorem 1 and following Propositions are shown in Appendix C. This theorem does not rely on additional assumptions of (local) convexity of deep neural networks, which is network architecture agnostic, which means the elicited bound can be applied to any black-box model. From this result, we can derive OoD generalizable set $\mathcal{B}$ via solving $\mathcal{L} > 1/2$. Based on Theorem 1, we have the following Propositions for various distributions.

Proposition 1. (Gaussian distribution) In this case, we instantiate π_0 as a Gaussian distribution that is centered at 0: $\mathcal{N}(0, \sigma^2 I)$, $\otimes$ as the addition operator, and $f_{\pi_0}(\mathbf{z}) > \frac{1}{2}$. Then $f_{\pi_0}(\mathbf{z})$ is $\mathcal{B}$-Generalizable for $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r\}$, with r satisfying the prediction confidence lowerbound:

$$\mathcal{L} \geq \Phi\left(\Phi^{-1}(f_{\pi_0}(\mathbf{z})) - \frac{r}{\sigma}\right) > \frac{1}{2} \quad (6)$$

$\Phi(\cdot)$ is the cumulative density function of standard Gaussian distribution, which gives:

$$r < \sigma\Phi^{-1}(f_{\pi_0}(\mathbf{z})) \quad (7)$$

Thus, the randomized classifier always gives correct predictions when the perturbation range r lies in the generalizable set $\mathcal{B}$. For Laplace distribution, the result is similar, which we leave to Appendix. We also consider the Bernoulli distribution, which can offer discrete stochastic disturbance to the semantic representations. Bernoulli distribution is the widely-used dropout trick for reducing overfitting in deep neural networks. Our results can shed some light on explaining why dropout works.

Proposition 2. (Bernoulli distribution) In this case, we instantiate π_0 as a Bernoulli distribution with a probability of setting to zero as $p \in [0, 1)$. $\otimes$ is instantiate as multiplication, and $f_{\pi_0}(\mathbf{z}) > \frac{1}{2}$. $\|\cdot\|_0$ represents the l_0-norm which counts the number of non-zero elements in a vector. Then $f_{\pi_0}(\mathbf{z})$ is $\mathcal{B}$-Generalizable for $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta} - \mathbf{1}\|_0 \leq r\}$, with r satisfying the prediction confidence lowerbound:

$$\mathcal{L} \geq \max\{f_{\pi_0}(\mathbf{z}) - 1 + p^r, 0\} > \frac{1}{2} \quad (8)$$

Which gives

$$r < \frac{\ln(1.5 - f_{\pi_0}(\mathbf{z}))}{\ln p} \quad (9)$$

The proof of Proposition 2 is challenging as it involves discrete distributions, which we leave for Appendix. From Proposition 2, it can be concluded that the radius of generalizable set $\mathcal{B}$ is propositional to the inverse of $\ln(p)$. This provides a plausible theoretical explanation of why the widely-used dropout methods can help avoid over-fitting and improve generalization performance. It also further reveals an inherent trade-off between selecting a higher dropout rate p and keeping $f_{\pi_0}(\mathbf{z}) > \frac{1}{2}$[4]. Additionally, this analysis also indicates a new research direction by automatically searching for the optimal parameters of random distributions *e.g.* the dropout rate to improve generalization abilities, which we leave for future works.

Remark 3: The above analysis equips us with methods for certifying OoD generalization algorithms. Though our analysis is based on a 0-1 classification problem, it can be directly extended to a multi-classes classification problem by constructing multiple one-vs-others classification problems. For the above propositions to hold, a necessary condition is $f_{\pi_0}(\mathbf{z}) > \frac{1}{2}$. Besides, in the binary classification setting, the closer $f_{\pi_0}(\mathbf{z})$ to 1 (i.e., further away from the decision boundary), the larger the allowed perturbation range r (i.e., the certifiable region $\mathcal{B}$) from Equation 7 [5].

Max-Margin Training

To find as large as possible certifiable region, we introduce max-margin training, which aligns with our moral of being away from the decision boundary. The maximal margin training is the task of finding the optimal hyperplane that linearly separates two separable classes. To ensure robustness, naturally, the optimal hyperplane is defined as the hyperplane that maximizes its distance ($= \frac{1}{2}$ margin) from the closest points of the two clouds of separable data. From the neural tangent kernel (NTK) theory in (Jacot, Gabriel, and Hongler 2018), we can approximate a high-dimensional non-linear deep neural network model $f_{\pi_0}(\mathbf{x}; \boldsymbol{\theta})$ with kernelized linear regression when the dimension of the network goes to infinity:

$$\begin{aligned}f_{\pi_0}(\mathbf{x}; \boldsymbol{\theta}) &\approx f_{\pi_0}(\mathbf{x}; \mathbf{w})\\&\approx f_{\pi_0}(\mathbf{x}; \mathbf{w}_0) + \Psi_{\pi_0}(\mathbf{x}; \mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0) \quad (10)\end{aligned}$$

where $\Psi_{\pi_0}(\mathbf{x}; \mathbf{w}_0)$ is the NTK, $\mathbf{w}$ is the last layer's parameter, $\mathbf{w}_0$ is the initialization for the last layer. Then, we derive the max-margin linear classifier for separating samples.

Theorem 2. (Max-margin classifier) If we wish to allow outliers (points within the margin, or even on the other side of decision boundary). The NTK max-margin classifier's parameters satisfy the following optimal conditions:

$$\begin{aligned}\min_{\boldsymbol{\xi},\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \quad (11)\\ \text{s.t.} \quad & y_i(f_{\pi_0}(\mathbf{x}_i; \mathbf{w}_0) + \Psi_{\pi_0}(\mathbf{x}_i; \mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0)) \geq 1 - \xi_i,\\ & \xi_i \geq 0, \;\; i = 1, \cdots, n\end{aligned}$$

where $\xi_i/\|\mathbf{w}\|$ is the distance of the furthest outlier point to the decision boundary (margin) along the i-th axis, C is the hyper-parameter controlling the costs of outliers. This is

[4] A too large dropout rate p may lead to wrong prediction results i.e., $f_{\pi_0}(\mathbf{z}) < \frac{1}{2}$.

[5] For Bernoulli case, the result is similar as maintaining higher $f_{\pi_0}(\mathbf{z})$ values can leave room for much smaller magnitude of $\ln(p)$.

equivalent to solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, \underbrace{1 - y_i(f_{\pi_0}(\mathbf{x}_i;\mathbf{w}_0) + \Psi_{\pi_0}(\mathbf{x}_i;\mathbf{w}_0)(\mathbf{w}-\mathbf{w}_0))}_{\text{Training loss}}) \tag{12}$$

We will show the equivalence of Equation 11 and Equation 12 in the Appendix. From the above equation, we can see that the essence of the max-margin classifier is the introduction of the Hinge loss which only keeps the training loss larger than a constant. Indeed, this can also be achieved by selecting samples that yield larger losses in a batch for training to avoid introducing an extra task-dependent hyper-parameter for practical reasons.

Practical Algorithm Instantiation

Based on the theoretical analysis, we propose an instantiation of the algorithm—Stochastic Disturbance Learning with Gaussian distribution as an example. For Gaussian distribution, the expectation of random disturbed loss for a data pair $(\mathbf{X}^i, \mathbf{Y}^i)$ is:

$$\mathbb{E}_{\pi_0}[\ell(f(\mathbf{X}^i;\boldsymbol{\theta}), \mathbf{Y}^i)] = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{N}(0,\sigma^2)}[\ell(f_{L-1}((\mathbf{z}+\boldsymbol{\eta});\boldsymbol{\theta}), \mathbf{Y}^i)] \tag{13}$$

where $\mathbf{z} = f_{L-2} \circ \cdots \circ f_0(\mathbf{X}^i;\boldsymbol{\theta})$. where σ is the added Gaussian distribution's variance. The proposed SDL algorithm is shown in Algorithm 1. Based on the derived propositions, an instantiation of the certification algorithm that obtains $\mathcal{B}$ is demonstrated in Appendix B. After training DNNs with Algorithm 1, we can use the Certification Algorithm in Appendix B to provide certified accuracies within the closed set $\mathcal{B}$ calculated by Propositions in the previous sections. This can provide a theoretical performance guarantee for high-stake applications.

Algorithm 1: Training procedure of stochastic disturbance learning

```
Require: Training set $(\mathbf{X}, \mathbf{Y})$, maximum number of epochs $T$, percentage of max-margin training epochs $\kappa$, percentage of top loss samples used in max-margin training $\eta$, batch-size $B$, variance of Gaussian distribution $\sigma$.
Ensure: The model's parameters $\boldsymbol{\theta}$.
while $t \leq (1-\kappa)T$ do
    Calculate the forward pass loss for every datum with added stochastic disturbances: $\mathbb{E}_{\pi_0}[\ell(f_{\boldsymbol{\theta}}(\mathbf{X}^i), \mathbf{Y}^i)]$ (See Eq 13).
    Select the top $\eta B$ loss samples for max-margin training $(X^i_{max}, Y^i_{max})$
    Train the model with $(X^i_{max}, Y^i_{max})$: $\mathbb{E}_{\pi_0}[\ell(f_{\boldsymbol{\theta}}(\mathbf{X}^i_{max}), \mathbf{Y}^i_{max})]$
end while
while $(1-\kappa)T < t \leq T$ do
    Calculate the forward pass loss for every datum $\mathbb{E}_{\pi_0}[\ell(f_{\boldsymbol{\theta}}(\mathbf{X}^i), \mathbf{Y}^i)]$
    Train the model with $(X, Y)$ with $\mathbb{E}_{\pi_0}[\ell(f_{\boldsymbol{\theta}}(\mathbf{X}^i), \mathbf{Y}^i)]$
end while
```

Experiments Results

This section will demonstrate the effectiveness of the proposed algorithmic framework with empirical experiments as it was found that benchmark results on OoD datasets are susceptible to hyper-parameters choices. For a fair comparison, we evaluate the effectiveness of our method with the OoD-Bench suit (Ye et al. 2021) based on the DomainBed implementation (Gulrajani and Lopez-Paz 2021). With OoD-Bench suit, we can evaluate the OoD generalization performances on datasets dominated by diversity shifts or correlation shifts. Next, ablation studies are conducted for further analysis.

OoD-Bench Results on Distribution Shifts Datasets

We follow the setting of OoD-Bench (Ye et al. 2021) for comparison experiments. Specifically, we have selected **PACS** (Li et al. 2017), **OfficeHome** (Venkateswara et al. 2017), **TerraIncognita** (Beery, Horn, and Perona 2018), and **Camelyon17-WILDS** (Koh et al. 2020) for benchmarking on the diversity shift datasets, and **Colored MNIST** (Arjovsky et al. 2019), **NICO** (He, Shen, and Cui 2020) , and a modified version of **CelebA** (Liu et al. 2015) for benchmarking on the correlation shift datasets. We use ResNet-18 for all experiments except the Colored MNIST dataset. For the Colored MNIST dataset, a multi-layer perceptron is used. For hyper-parameter search, we run twenty iterations for each algorithm and the search procedure is repeated three times. The means and standard deviations of accuracies are reported. For every dataset-algorithm pair, depending on whether the attained accuracy is lower than, within, or higher than the standard error bar of ERM accuracy on the same dataset, the ranking score -1, 0, +1 is assigned to the pair. Eighteen strong OoD generalization algorithms are compared, including the invariant risk minimization methods (e.g. IRM (Arjovsky et al. 2019), VREx (Krueger et al. 2020)), distributionally robust optimization method (e.g. GroupDRO (Sagawa* et al. 2020)), and domain generalization methods (e.g. MLDG (Li et al. 2018a), ERDG (Zhao et al. 2020)), etc. The results for diversity shifts dominated datasets are shown in Table 1. For correlation shifts dominated datasets, the results are shown in Table 2.

From Table 1 and Table 2, it can be observed that all methods except for the proposed SDL can only achieve better results than ERM on either type of distribution shift. From Table 1, we can observe that the proposed SDL method achieves better performances. Among the top performers, we observe that RSC, MMD, and SagNet outperform the standard empirical risk minimization method. This demonstrates that only a few methods can achieve better performances with systematic evaluation than ERM, revealing the inherent challenge of OoD generalization. From Table 2, the proposed method is still the best-performing method among all candidates. This answers an unsolved question in the previous paper (Ye et al. 2021) whether there exists an OoD generalization algorithm that can achieve better performances than ERM simultaneously on diversity and correlation shifts dominated datasets. For overall performance, the proposed SDL obtains a ranking score of +5 followed by RSC and MMD with the ranking score of +1, which means SDL achieves better performances stably for most OoD datasets in OoD-Bench.

Algorithm	PACS	OfficeHome	Terra Incognita	Camelyon17	Average	Ranking score
SDL(Proposed)	84.8 ± 0.6	63.9 ± 0.1	44.1 ± 1.1	95.4 ± 0.3	72.1	+4
RSC	82.8 ± 0.4	62.9 ± 0.4	43.6 ± 0.5	94.9 ± 0.2	71.1	+2
MMD	81.7 ± 0.2	63.8 ± 0.1	38.3 ± 0.4	94.9 ± 0.4	69.7	+2
SagNet	81.6 ± 0.4	62.7 ± 0.4	42.3 ± 0.7	95.0 ± 0.2	70.4	+1
ERM	81.5 ± 0.0	63.3 ± 0.2	42.6 ± 0.9	94.7 ± 0.1	70.5	0
IGA	80.9 ± 0.4	63.6 ± 0.2	41.3 ± 0.8	95.1 ± 0.1	70.2	0
CORAL	81.6 ± 0.6	63.8 ± 0.3	38.3 ± 0.7	94.2 ± 0.3	69.5	0
IRM	81.1 ± 0.3	63.0 ± 0.2	42.0 ± 1.8	95.0 ± 0.4	70.3	-1
VREx	81.8 ± 0.1	63.5 ± 0.1	40.7 ± 0.7	94.1 ± 0.3	70.0	-1
GroupDRO	80.4 ± 0.3	63.2 ± 0.2	36.8 ± 1.1	95.2 ± 0.2	68.9	-1
ERDG	80.5 ± 0.5	63.0 ± 0.4	41.3 ± 1.2	95.5 ± 0.2	70.1	-2
DANN	81.1 ± 0.4	62.9 ± 0.6	39.5 ± 0.2	94.9 ± 0.0	69.6	-2
MTL	81.2 ± 0.4	62.9 ± 0.2	38.9 ± 0.6	95.0 ± 0.1	69.5	-2
Mixup	79.8 ± 0.6	63.3 ± 0.5	39.8 ± 0.3	94.6 ± 0.3	69.4	-2
ANDMask	79.5 ± 0.0	62.0 ± 0.3	39.8 ± 1.4	95.3 ± 0.1	69.2	-2
ARM	81.0 ± 0.4	63.2 ± 0.2	39.4 ± 0.7	93.5 ± 0.6	69.2	-3
MLDG	73.0 ± 0.4	52.4 ± 0.2	27.4 ± 2.0	91.2 ± 0.4	61.0	-4
Average	80.8	62.6	39.8	94.6	69.4	–

Table 1: Performance of ERM and OoD generalization algorithms on datasets *dominated by diversity shift*. The baseline methods include RSC (Huang et al. 2020), MMD (Li et al. 2018b), SagNet (Nam et al. 2019), IGA (Koyama and Yamaguchi 2020a), CORAL (Sun and Saenko 2016), IRM (Arjovsky et al. 2019), VREx (Krueger et al. 2020), GroupDRO (Sagawa* et al. 2020), ERDG (Zhao et al. 2020), DANN (Ganin et al. 2016), MTL (Blanchard et al. 2017), Mixup (Yan et al. 2020), ANDMask (Parascandolo et al. 2021), ARM (Zhang et al. 2020), and MLDG (Li et al. 2018a). Every symbol ↓ denotes a score of -1, and every symbol ↑ denotes a score of +1; otherwise, the score is 0. The scores indicate how many datasets the candidate algorithm are performed better than ERM. Adding up the scores across all datasets produces the ranking score for each algorithm. The proposed SDL can achieve the best average performance and the highest ranking score.

Certified Accuracy on OoD Datasets and Ablation Studies

The certified accuracy is defined as the fraction of the test set samples which is certifiably correct within the maximum allowable generalizable set $\mathcal{B}$. We show the certified accuracies of the proposed algorithmic scheme taking PACS and OfficeHome as examples. We plot the certified accuracies of the proposed SDL against the radius of $\mathcal{B}$ varying the variance σ, which is shown in Figure 3 (a). We can observe that varying variances σ can yield different trade-offs between certified accuracies and the radius of the generalizable set $\mathcal{B}$. Larger variance generally leads to higher certified accuracies where semantic information deviates further. This aligns well with our theoretical analysis that larger variance can lead to a higher allowable radius (Equation 7). The certifying result with the Bernoulli distribution is shown in Appendix C.

The effectiveness of max-margin training and random noises in SDL is investigated for ablation studies. We compare SDL and its variants without max-margin training (M_{a1}) or random noises (M_{a2}), or without both of them (ERM). The results are visualized in Figure 3 (b). From Figure 3 (b), we can observe that SDL can achieve statistically significant better results than its variants on all radius selections. Furthermore, when the deviation degree increases, the certified accuracy reduces more significantly after removing the random noises part, confirming the algorithmic components' necessity.

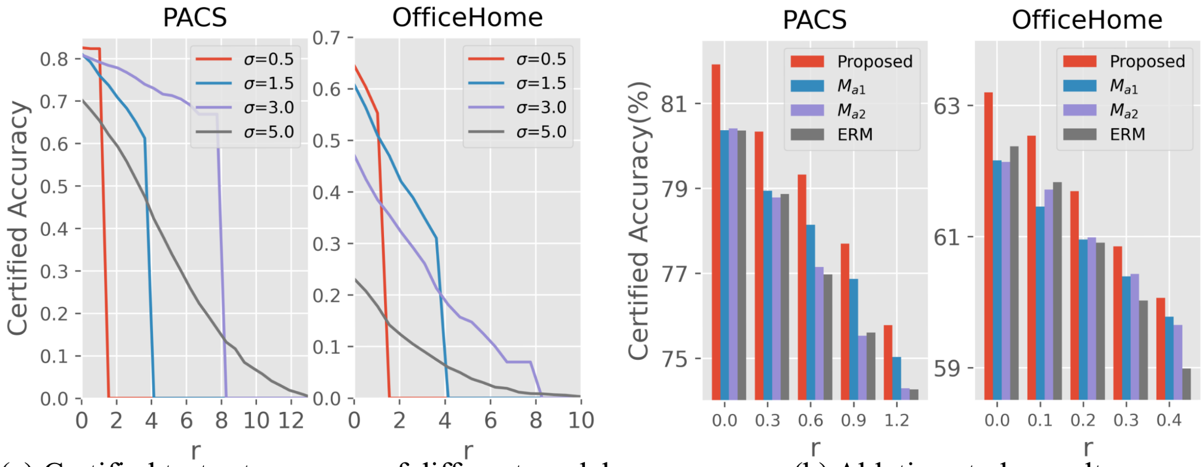
We have analyzed the relationship between the radius of $\mathcal{B}$ and the variance σ both theoretically and empirically (see Equation 7 and Figure 4(a)). The degree of domain shift between training data and test data affects $\mathcal{B}$ through $f_{\pi_0}(z)$. As for a dataset with a very large degree of distribution shift, it may be quite hard for the baseline to learn $f_{\pi_0}(z)$ well, which may be close to the classification bound of 0.5. We have addressed this issue by max-margin training which is quite efficient according to the ablation study. In fact, a very large domain shift does not essentially cause $\mathcal{B}$-set very small or disappear according to the experimental results. For example, for dataset PACS with a very large degree of diversity shift (about 0.8 in OoD-Bench), the certified test accuracies by SDL ($\sigma = 3.0$) start to drop significantly until the radius is up to 8.

Conclusion

In this paper, we propose a certifiable out-of-distribution generalization algorithm that can provide certified accuracy for each input datum's prediction on the semantic space. The proposed method simultaneously achieves the state-of-the-art empirical performance on datasets dominated by two types of distribution shifts with theoretically guaranteed performance. For future work, we will explore how to further apply this method to other tasks, such as autonomous driving or medical image processing, to improve OoD generalization performances.

Algorithm	Colored MNIST	CelebA	NICO	Average	Ranking score
SDL(Proposed)	58.8 ± 2.2	88.6 ± 0.5	71.7 ± 0.6	73.0	+2
VREx (Krueger et al. 2020)	56.3 ± 1.9	87.3 ± 0.2	71.5 ± 2.3	71.7	+1
GroupDRO (Sagawa* et al. 2020)	32.5 ± 0.2	87.5 ± 1.1	71.0 ± 0.4	63.7	+1
ERM (Vapnik 1998)	29.9 ± 0.9	87.2 ± 0.6	72.1 ± 1.6	63.1	0
IRM (Arjovsky et al. 2019)	60.2 ± 2.4	85.4 ± 1.2	73.3 ± 2.1	73.0	0
MTL (Blanchard et al. 2017)	29.3 ± 0.1	87.0 ± 0.7	70.6 ± 0.8	62.3	0
ERDG (Zhao et al. 2020)	31.6 ± 1.3	84.5 ± 0.2	72.7 ± 1.9	62.9	0
ARM (Zhang et al. 2020)	34.6 ± 1.8	86.6 ± 0.7	67.3 ± 0.2	62.8	0
MMD (Li et al. 2018b)	50.7 ± 0.1	86.0 ± 0.5	68.9 ± 1.2	68.5	−1
RSC (Huang et al. 2020)	28.6 ± 1.5	85.9 ± 0.2	74.3 ± 1.9	62.9	−1
IGA (Koyama and Yamaguchi 2020a)	29.7 ± 0.5	86.2 ± 0.7	71.0 ± 0.1	62.3	−1
CORAL (Sun and Saenko 2016)	30.0 ± 0.5	86.3 ± 0.5	70.8 ± 1.0	62.4	−1
Mixup (Yan et al. 2020)	27.6 ± 1.8	87.5 ± 0.5	72.5 ± 1.5	62.5	−1
MLDG (Li et al. 2018a)	32.7 ± 1.1	85.4 ± 1.3	66.6 ± 2.4	61.6	−1
SagNet (Nam et al. 2019)	30.5 ± 0.7	85.8 ± 1.4	69.8 ± 0.7	62.0	−2
ANDMask (Parascandolo et al. 2021)	27.2 ± 1.4	86.2 ± 0.2	71.2 ± 0.8	61.5	−2
DANN (Ganin et al. 2016)	24.5 ± 0.8	86.0 ± 0.4	69.4 ± 1.7	60.0	−3
Average	36.2	86.4	70.9	64.5	–

Table 2: Performance of ERM and OoD generalization algorithms on datasets *dominated by correlation shift*. Every symbol ↓ denotes a score of −1, and every symbol ↑ denotes a score of +1; otherwise, the score is 0. The scores indicate how many datasets the candidate algorithm performs better than ERM. Adding up the scores across all datasets produces the ranking score for each algorithm. The proposed SDL can achieve the best average performance and the highest ranking score.



(a) Certified test set accuracy of different models.

(b) Ablation study results.

Figure 3: (a) Certified accuracy on PACS and OfficeHome. The left line chart shows the certified accuracy under SDL with different σ of $\pi(\boldsymbol{\eta})$. The x-axis denotes the radius of generalizable set $\mathcal{B}$. (b) The right bar chart shows the ablation study results of removing max-margin training (M_{a1}) or random noises (M_{a2}).

Acknowledgements

Nanyang Ye was supported in part by National Natural Science Foundation of China under Grant No.62106139.

References

Ahuja, K.; Caballero, E.; Zhang, D.; Gagnon-Audet, J.-C.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. In *NeurIPS*.

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020a. Invariant risk minimization games. In *ICML*.

Ahuja, K.; Shanmugam, K.; Varshney, K. R.; and Dhurandhar, A. 2020b. Invariant Risk Minimization Games. arXiv:1907.02893.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv:1907.02893*.

Beery, S.; Horn, G. V.; and Perona, P. 2018. Recognition in Terra Incognita. In *ECCV*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning*, 79(1): 151–175.

Blanchard, G.; Deshmukh, A. A.; Dogan, U.; Lee, G.; and Scott, C. 2017. Domain generalization by marginal transfer learning. *arXiv:1711.07910*.

Cover, T. M.; and Thomas, J. A. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience. ISBN 0471241954.

Duchi, J.; and Namkoong, H. 2019. Learning Models with Uniform Performance via Distributionally Robust Optimization. *arXiv:1810.08750*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *JMLR*.

Gao, R.; and Kleywegt, A. J. 2016. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv:1604.02199*.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *ICLR*.

He, Y.; Shen, Z.; and Cui, P. 2020. Towards Non-IID Image Classification: A Dataset and Baselines. *Pattern Recognition*.

Huang, Z.; Wang, H.; Huang, D.; Lee, Y. J.; and Xing, E. P. 2022. The Two Dimensions of Worst-case Training and the Integrated Effect for Out-of-domain Generalization. *arXiv:2204.04384*.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. *arXiv:2007.02454*.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv:1806.07572*.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Beery, S.; et al. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421*.

Koyama, M.; and Yamaguchi, S. 2020a. Out-of-distribution generalization with maximal invariant predictor. *arXiv:2008.01883*.

Koyama, M.; and Yamaguchi, S. 2020b. When is invariance useful in an Out-of-Distribution Generalization problem ? *arXiv:2008.01883*.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Priol, R. L.; and Courville, A. 2020. Out-of-Distribution Generalization via Risk Extrapolation (REx). *arXiv:2003.00688*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018a. Learning to generalize: Meta-learning for domain generalization. In *AAAI*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, Broader and Artier Domain Generalization. In *ICCV*.

Li, H.; Jialin Pan, S.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *CVPR*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2019. Reducing domain gap via style-agnostic networks. *arXiv:1910.11645*.

Parascandolo, G.; Neitz, A.; Orvieto, A.; Gresele, L.; and Schölkopf, B. 2021. Learning explanations that are hard to vary. In *ICLR*.

Rahimian, H.; and Mehrotra, S. 2019. Distributionally Robust Optimization: A Review. *arXiv:1908.05659*.

Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; and Peters, J. 2018. Invariant Models for Causal Transfer Learning. *JMLR*, 19(36): 1–34.

Sagawa*, S.; Koh*, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks. In *ICLR*.

Shen, Z.; Cui, P.; Liu, J.; Zhang, T.; Li, B.; and Chen, Z. 2020a. Stable Learning via Differentiated Variable Decorrelation. In *KDD*.

Shen, Z.; Cui, P.; Zhang, T.; and Kuang, K. 2020b. Stable Learning via Sample Reweighting. In *AAAI*.

Silver, D.; Huang, A.; Maddison, C.; Guez, A.; Sifre, L.; Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–489.

Sinha, A.; Namkoong, H.; and Duchi, J. 2018. Certifiable Distributional Robustness with Principled Adversarial Training. In *ICLR*.

Staib, M.; and Jegelka, S. 2019. Distributionally Robust Optimization and Generalization in Kernel Methods. In *NeurIPS*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*.

Tran, D.; Ruiz, F. J.; Athey, S.; and M.Blei, D. 2016. Model Criticism for Bayesian Causal Inference. *arXiv:1610.09037*.

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *CVPR*.

Weber, M.; Li, L.; Wang, B.; Zhao, Z.; Li, B.; and Zhang, C. 2022. Certifying Out-of-Domain Generalization for Blackbox Functions. *arXiv: 2202.01679*.

Wiles, O.; Gowal, S.; Stimberg, F.; Rebuffi, S.-A.; Ktena, I.; Dvijotham, K. D.; and Cemgil, A. T. 2022. A Fine-Grained Analysis on Distribution Shift. In *ICLR*.

Yan, S.; Song, H.; Li, N.; Zou, L.; and Ren, L. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*.

Ye, N.; Li, K.; Bai, H.; Yu, R.; Hong, L.; Zhou, F.; Li, Z.; and Zhu, J. 2021. OoD-Bench: Benchmarking and Understanding Out-of-Distribution Generalization Datasets and Algorithms. *arXiv:2106.03721*.

Zeiler, M. D.; and Fergus, R. 2013. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901*.

Zhang, M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; and Finn, C. 2020. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Distribution Shift. *arXiv:2007.02931*.

Zhao, H.; des Combes, R. T.; Zhang, K.; and Gordon, G. J. 2019. On Learning Invariant Representation for Domain Adaptation. *CoRR*, abs/1901.09453.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020. Domain Generalization via Entropy Regularization. In *NeurIPS*.