Risk Variance Penalization: From Distributional Robustness to Causality

Xie Chuanlong
Huawei Noah's Ark Lab
xie.chuanlong@huawei.com

Chen Fei
Huawei Noah's Ark Lab
chen.f@huawei.com

Liu Yue
Huawei Noah's Ark Lab
liuyue52@huawei.com

Li Zhenguo*
Huawei Noah's Ark Lab
Li.Zhenguo@huawei.com

Abstract

Learning under multi-environments often requires the ability of out-of-distribution generalization for the worst-environment performance guarantee. Some novel algorithms, e.g. Invariant Risk Minimization [3] and Risk Extrapolation [25], build stable models by extracting invariant (causal) feature. However, it remains unclear how these methods learn to remove the environmental features. In this paper, we focus on the Risk Extrapolation (REx) and make attempts to fill this gap. We first propose a framework, Quasi-Distributional Robustness, to unify the Empirical Risk Minimization (ERM), the Robust Optimization (RO) and the Risk Extrapolation. Then, under this framework, we show that, comparing to ERM and RO, REx has a much larger robust region. Furthermore, based on our analysis, we propose a novel regularization method, Risk Variance Penalization (RVP), which is derived from REx. The proposed method is easy to implement, and has proper degree of penalization, and enjoys an interpretable tuning parameter. Finally, our experiments show that under certain conditions, the regularization strategy that encourages the equality of training risks has ability to discover relationships which do not exist in the training data. This provides important evidence to support that RVP is useful to discover causal models.

1 Introduction

Consider a learning task, in which the training data is collected from multiple environments, e.g. research centers, times, experimental conditions and so on. Thus the training data is structured,

$$\mathcal{S} = \{S_1, S_2, \ldots, S_m\}, \quad S_i = \{z_{i1}, z_{i2}, \ldots, z_{in}\} \text{ with } z_{ij} \sim P_i,$$

where $z = (x, y)$ is a data point consisting of an input x and the corresponding target y, and P_i is a data-generating distribution which represents the learning task under the i-th environment. Intuitively, the sample S_i can inherit some features from the environment behind P_i, which varies among the environments but can be highly related to the target under certain environments. In practice, without taking the structure of $\mathcal{S}$ into consideration, e.g. the shuffle operation, the learner will absorb all the correlations in the pooled $\mathcal{S}$ and learn a model $\hat{f}$ based on all features which are highly related to the target [21, 38, 39]. A classical example is the problem of classifying images of cows and camels [7]. Due to the selection bias, cows appear in pastures at most pictures and camels are taken in deserts. The common algorithms may learn to recognize cows and camels with background features and struggle to detect cows in the desert and camels in pastures. In other words, the learned model can have good performance on the learning tasks within the training environments, but may dramatically fail under some unseen environments.

We denote all possible environments by $\mathcal{P}$, which is a collection of distributions including P_i, $i = 1, \ldots, m$ and other unseen environments. Let $f \in \mathcal{F}$ be a hypothetical model that maps x to

Preprint.

$f(x)$ and $\ell(f, z)$ be the loss function measuring how poorly the output $f(x)$ predicts the target y. For a given model f, the risk and the empirical risk are defined by the expected loss and its empirical version respectively, denoted by

$$R(f, P_i) = \mathbb{E}_{Z\sim P_i}[\ell(f, Z)], \quad R(f, S_i) = \frac{1}{n}\sum_{j=1}^{n} \ell(f, z_{ij}).$$

The out-of-distribution generalization aims to learn a model f that has the best worst-environment performance, e.g.

$$\hat{f}_{OOD} = \arg\min_{f} \sup_{P\in\mathcal{P}} R(f, P).$$

Denote Q as a distribution on $\mathcal{P}$ that controls which environments the learner is likely to see in the learning procedure. Let $\bar{S}$ be the pooled $\mathcal{S}$. The Empirical Risk Minimization (ERM, [29]) is

$$\hat{f}_{ERM} \in \arg\min_{f} R(f, \bar{S}), \quad \text{and} \quad R(f, \bar{S}) := \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} \ell(f, z_{ij}),$$

where $R(f, \bar{S})$ is the empirical version of $R(f, \bar{P})$ with $\bar{P} = \int P dQ$. Thus ERM cannot uniformly minimize the risk $R(f, P)$ for any $P \in \mathcal{P}$. However, due to the sampling at the environment level, the ERM solution still shows stable performance on certain environments, especially for some domain generalization problems, e.g. [11, 14, 22, 28, 31, 36].

It is well known that the generalization performance of a causal model is more stable than that of a non-causal model if we can intervene on the input or change the environment [33]. The predictor based on causal features will in general work as well under interventions as for the raw data. The invariant risk minimization (IRM, [3]) proposes to learn a stable classifier by extracting the causal features, and suggests the objective:

$$\mathcal{R}_{IRM}(f) = R(f, \bar{S}) + \lambda \sum_{i=1}^{m} \left\|\nabla_w R(w * f, S_i)\big|_{w=1}\right\|^2,$$

where f is the entire hypothetical model, w is a scalar, and $\lambda \in [0, \infty)$ is a tuning parameter. Motivated by IRM and the robust optimization, [25] proposes the Minimax Risk Extrapolation (MM-REx), whose objective function is

$$\mathcal{R}_{MM-REx}(f) = \max_{\substack{\sum_i q_i=1,\\ q_i\geq -\lambda}} \sum_{P_i\in\mathcal{P}_{tr}} q_i R(f, S_i).$$

Notice that the penalty in IRM and the maximization of MM-REx are not easy to compute. Furthermore, a practical version, Variance Risk Extrapolation (V-REx, [25]), is proposed that minimizes the objective

$$\mathcal{R}_{V-REx}(f) = R(f, \bar{S}) + \lambda \text{Var}(f, \mathcal{S}),$$

where

$$\text{Var}(f, \mathcal{S}) = \frac{1}{m}\sum_{i=1}^{m} (R(f, S_i) - R(f, \bar{S}))^2.$$

However it is still unclear: (i) Is V-REx equivalent to MM-REx? (ii) How IRM and V-REx discover the invariant (causal) features via penalization? In this paper, we focus on the REx methods and make attempts to answer these questions. For problem (i), we propose a general framework, named Quasi-Distributional Robustness, to investigate the connection between the distributional robustness and the MM-REx. Then we suggest a new method, called Risk Variance Penalization (RVP), that minimizes the following objective function:

$$\mathcal{R}_{RVP}(f) = R(f, \bar{S}) + \lambda\sqrt{\text{Var}(f, \mathcal{S})}, \tag{1}$$

which is derived from MM-REx and is an upper bound of a distributional robustness problem. We further prove that, under mild conditions, the upper bound can be achieved, which also implies MM-REx can be exactly solved by minimizing (1). To answer (ii), we experimentally prove that, comparing to the robust optimization, V-REx and RVP can learn invariant features to obtain the out-of-distribution generalization. Furthermore, we design a learning scheme to illustrate that the penalization strategy of V-REx and RVP is able to discover relationships which do not exist in the training data.

2 Preliminaries

2.1 The OOD generalization of ERM

A hypothetical model f working well on a task P implies that the risk $R(f, P)$ should be small. However, the distribution P is generally unknown and a surrogate, the empirical distribution of a sample, is available. The generalization guarantee gives the upper bound of $R(f, P)$ based on the surrogate, e.g. [2, 5, 13, 16, 23, 30]. Under appropriate conditions, some classical works, e.g. [5, 6, 9, 10, 24], give the upper bound of $R(f, \bar{P})$ in the form of

$$R(f, \bar{S}) + C\sqrt{\frac{1}{mn}\mathrm{Var}_{\bar{P}}\big(\ell(f, Z)\big)} + O(\frac{1}{mn}),$$

where C is a constant that depends on $\mathcal{F}$ and the confidence level $1 - \delta$. Comparing to learning under a single-environment, the pooled ERM reduces the generalization bound due to the factor $1/m$, and at the same time, increases the variance term because

$$\mathrm{Var}_{\bar{P}}\big(\ell(f, Z)\big) = \mathbb{E}_Q\big[\mathrm{Var}_P(\ell(f, Z))\big] + \mathrm{Var}_Q\big(\mathbb{E}_P[\ell(f, Z)]\big).$$

Notice that $\mathbb{E}_Q\big[\mathrm{Var}_P(\ell(f, Z))\big]$ is the inner environment variance, and $\mathrm{Var}_Q\big(\mathbb{E}_P[\ell(f, Z)]\big)$ represents the cross environment variance introduced by pooling the training data $\mathcal{S}$.

2.2 Related work

Robust optimization. The robust optimization techniques consider the worst-environment loss over all possible convex combinations of the training environments or data points [4, 8, 15, 18, 27, 37]. A straightforward approach is to replace the population of the environments by the empirical distribution of the training environments, and then to minimize the worst empirical performance:

$$\mathcal{R}_{RO}(f) = \max_{P_i \in \mathcal{P}_{tr}} R(f, S_i) = \max_{\substack{\sum_i q_i = 1, \\ q_i \geq 0}} \sum_{P_i \in \mathcal{P}_{tr}} q_i R(f, S_i).$$

This assumes that the test distribution is a convex combination of all training distributions, which cannot extrapolate the training environments. Consequently, it may use unstable features of all training environments and fail to discover the invariant model[3].

Invariant feature learning. Finding the invariant feature or representation is a common-used method for the OOD generalization. [26, 28, 29, 34, 40] note that the OOD model can be viewed as a solution that minimizes the worst-environment risk under different distribution shifts. Invariant feature finding (selection) is also used in causal discovery. [33] proposes Invariant Causal Prediction (ICP) to estimate a set of the causal features. [19, 20] use multiple datasets from different environments for causal discovery.

3 Methodology

3.1 Quasi-distributional robustness

In this section, we propose a new framework, Quasi-Distributional Robustness, to investigate MM-REx from the view of distributional robustness, and prove that under mild conditions, MM-REx is also a variance-penalized method.

To proceed further, we need more notations. For a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ with $\phi(1) = 0$, the ϕ-divergence between two distributions Q and Q' is defined by $D_\phi(Q\|Q') = \int_{\mathcal{P}} \phi(dQ/dQ')dQ'$. In this paper, we use the χ^2-divergence, where $\phi(x) = (x-1)^2/2$. Denote $Q = (q_1, q_2, \ldots, q_m)$ and

$$\begin{aligned} \mathcal{R}(f) &= \big(R(f, S_1),\ R(f, S_2),\ \ldots,\ R(f, S_m)\big), \\ \mathcal{Q}_m(\lambda, \rho) &= \big\{Q;\ q_i \geq -\lambda,\ \sum_i q_i = 1,\ \frac{1}{2m}\sum_i \|mq_i - 1\|^2 \leq \frac{\rho}{m}\big\}. \end{aligned}$$

Here $\mathcal{Q}_m$ is a set of Q and stands for a robust region. For MM-REx, $\rho = \infty$. If $\lambda = 0$, $\mathcal{Q}_m$ is a set of distributions and can be rewritten as $\mathcal{Q}_m(0, \rho) = \{Q : D_\phi(Q\|Q_m) \leq \rho/m\}$, where Q_m is an

empirical distribution with $q_i = 1/m$, $i = 1, \ldots, m$. In the following, we still use D_ϕ to measure the distance between two elements of $\mathcal{Q}_m(\lambda, \rho)$. We further define a robust optimization problem,

$$min_f \mathcal{R}(f, \lambda, \rho) := \min_f \max_Q \langle Q, \mathcal{R}(f) \rangle, \quad \text{with} \quad Q \in \mathcal{Q}_m(\lambda, \rho),$$

where $\langle ., . \rangle$ stands for the inner product. This is a general version of distributional robustness since q_i can be negative. Thus we call this problem as quasi-distributional robustness (QDR). The choices of λ and ρ determine the size of $\mathcal{Q}_m$ and subsequently influence the robustness guarantees. It is easy to see that, both ERM and RO are special cases of $\mathcal{R}(f, \lambda, +\infty)$:

$$\mathcal{R}_{ERM}(f) = \mathcal{R}(f, -\frac{1}{m}, +\infty) \quad \text{and} \quad \mathcal{R}_{RO}(f) = \mathcal{R}(f, 0, +\infty).$$

Thus $\mathcal{R}(\lambda, +\infty)$ can represent ERM, RO and MM-REx with different values of λ. However, it is not easy to interpret λ, since $-\lambda$ is a lower bound of q_i. This motivates us to rewrite the objective of ERM, RO and MM-REx. Notice that, for ERM, $q_i \geq -\lambda = 1/m$. Thus $\mathcal{Q}(-1/m, +\infty)$ only contains one element Q_m and can be rewritten as $\mathcal{Q}(-1/m, 0)$. Further we have

$$\mathcal{R}_{ERM}(f) = \mathcal{R}(f, -\frac{1}{m}, +\infty) = \mathcal{R}(f, -\frac{1}{m}, 0).$$

In the following, we shall prove that λ and ρ govern each other, and for any λ, we can find a finite scalar C_λ such that $\mathcal{R}(f, \lambda, +\infty) = \mathcal{R}(f, \lambda, C_\lambda)$.

3.2 Analysis

We start with an upper bound of $\mathcal{R}(f, \lambda, \rho)$, which is independent to λ.

Proposition 1. *For any $f \in \mathcal{F}$, $\rho > 0$ and $-1/m \leq \lambda \leq +\infty$,*

$$\mathcal{R}(f, \lambda, \rho) \leq R(f, \bar{S}) + \sqrt{\frac{2\rho}{m} \text{Var}(f, \mathcal{S})}$$

Proof: For any $\langle Q, \mathcal{R}(f) \rangle$, we can rewrite it as

$$\langle Q, \mathcal{R}(f) \rangle = R(f, \bar{S}) + \langle Q - Q_m, \mathcal{R}(f) \rangle.$$

Notice that $\langle Q - Q_m, \mathbb{I}_m \rangle = 0$ with $\mathbb{I}_m = (1, 1, \ldots, 1)$. Thus

$$\langle Q, \mathcal{R}(f) \rangle = R(f, \bar{S}) + \langle Q - Q_m, \mathcal{R}^c(f) \rangle,$$

where $\mathcal{R}^c(f) = \mathcal{R}(f) - \mathbb{I}_m * R(f, \bar{S})$. Then we have

$$\begin{aligned} \max_Q \langle Q, \mathcal{R}(f) \rangle &= R(f, \bar{S}) + \max_Q \langle Q - Q_m, \mathcal{R}^c(f) \rangle \\ &\leq R(f, \bar{S}) + \|Q - Q_m\| \|\mathcal{R}^c(f)\|. \end{aligned}$$

It is easy to see $\|\mathcal{R}^c(f)\|^2 = m\text{Var}(f, \mathcal{S})$. By the definition of $\mathcal{Q}_m$, $\|Q - Q_m\|^2 \leq 2\rho/m^2$. Hence the proof is finished. □

Notice that the inequality holds uniformly for $-1/m \leq \lambda \leq +\infty$. This implies that, if the equality can be achieved, that is

$$\mathcal{R}(f, \lambda, \rho) = R(f, \bar{S}) + \sqrt{\frac{2\rho}{m} \text{Var}(f, \mathcal{S})}, \tag{2}$$

then $\mathcal{R}(f, \lambda, \rho)$ can deal with the robust optimization problem on $\mathcal{R}(f, +\infty, \rho)$. In other words, the objective function $\mathcal{R}(f, \lambda, \rho)$ can bound all possible linear combinations of the training risks $\langle Q, \mathcal{R}(f) \rangle$ on $\mathcal{Q}_m(\infty, \rho)$. The following result states that the equality in Proposition 1 can be achieved if λ and ρ are properly selected.

Proposition 2. *Suppose that the multi-environment training data $\mathcal{S}$ and the hypothetical model f are given and $\text{Var}(f, \mathcal{S}) > 0$. If λ and ρ satisfy*

$$\lambda \geq C(f, \mathcal{S}, \rho) := -\frac{1}{m} + \sqrt{\frac{2\rho}{m^2}} \frac{|\min_i R(f, S_i) - R(f, \bar{S})|}{\sqrt{m\text{Var}(f, \mathcal{S})}}, \tag{3}$$

then the equality in (2) holds.

Proof: According to the proof of Proposition 1, if (2) holds, the vector Q should satisfy: (i) $Q - Q_m$ and $\mathcal{R}^c(f)$ are in the same direction; (ii) $\|Q - Q_m\|^2 = 2\rho/m^2$. This implies that the i-th element of Q should be

$$q_i = \frac{1}{m} + \sqrt{\frac{2\rho}{m^2}} \frac{R(f, S_i) - R(f, \bar{S})}{\sqrt{m\text{Var}(f, \mathcal{S})}}.$$

In addition, $\mathcal{Q}_m(\lambda, \rho)$ requires $\min_i q_i \geq -\lambda$. Thus,

$$\frac{1}{m} + \sqrt{\frac{2\rho}{m^2}} \frac{\min_i R(f, S_i) - R(f, \bar{S})}{\sqrt{m\text{Var}(f, \mathcal{S})}} \geq -\lambda$$

Hence (3) is proved. □

According to the condition (3), $\mathcal{R}(f, \lambda, \rho) = \mathcal{R}(f, +\infty, \rho)$ when λ is sufficiently large. In fact, the scalar $C(f, \mathcal{S}, \rho)$ measures how hard it is to achieve robustness on $\mathcal{Q}_m(+\infty, \rho)$. If $C(f, \mathcal{S}, \rho) < 0$, RO is equivalent to MM-REx. From the expression of $C(f, \mathcal{S}, \rho)$, we can understand three factors m, ρ and $\text{Var}(f, \mathcal{S})$ that influence the QDR problem. If $C(f, \mathcal{S}, \rho) > 0$, more training environments (increase m) make the robustness easier to achieve. For simplicity, we call ρ/m as the radius of $\mathcal{Q}_m(+\infty, \rho)$. It is difficult to obtain a robust model for a large region $\mathcal{Q}_m(+\infty, \rho)$. Thus $C(f, \mathcal{S}, \rho)$ increases as ρ increases. For the third factor $\text{Var}(f, \mathcal{S})$, the variance represents the diversity of the training environments, which benefits the robust learning.

On the other hand, if λ is fixed and ρ is small enough such that

$$\rho \leq C'(f, \mathcal{S}, \lambda) := \frac{m(m\lambda + 1)^2 \text{Var}(f, \mathcal{S})}{2\big(\min_i R(f, S_i) - R(f, \bar{S})\big)^2},$$

then the equation (2) still holds. Thus λ also governs ρ. To proceed further, we denote

$$\rho_+ = D_\phi(Q^* \| Q_m), \quad \text{where} \quad Q^* = \big(1 + (m-1)\lambda, -\lambda, \ldots, -\lambda\big).$$

Here Q^* is a vertex of $\mathcal{Q}_m(\lambda, +\infty)$ and ρ_+ is the largest distance between Q and Q_m for $Q \in \mathcal{Q}_m(\lambda, +\infty)$. Now we are ready to state:

Proposition 3. *Suppose that the multi-environment training data $\mathcal{S}$ and the hypothetical model f are given and* $\text{Var}(f, \mathcal{S}) > 0$. *If λ is fixed and $\rho_- = C'(f, \mathcal{S}, \lambda)$, then*

$$\mathcal{R}(f, \lambda, \rho_-) \leq \mathcal{R}(f, \lambda, +\infty) \leq \mathcal{R}(f, \lambda, \rho_+).$$

Proof: The first inequality is trivial since $\mathcal{Q}_m(\lambda, \rho_-) \subset \mathcal{Q}_m(\lambda, +\infty)$. On the other hand, the center of $\mathcal{Q}_m(\lambda, +\infty)$ is Q_m. For any $Q \in \mathcal{Q}_m(\lambda, +\infty)$, the largest distance between Q and Q_m is achieved at the vertices. Note that

$$\rho_+ = D_\phi(Q^* \| Q_m) = \max_{Q \in \mathcal{Q}_m(\lambda, +\infty)} D_\phi(Q \| Q_m),$$

where $Q^* = \big(1 + (m-1)\lambda, -\lambda, \ldots, -\lambda\big)$. This implies the set $\mathcal{Q}_m(\lambda, +\infty)$ is covered by $\mathcal{Q}_m(\lambda, \rho_+)$. Then the second inequality is proved. □

One can find that for any given λ, there exists $\rho^* \in [\rho_-, \rho_+]$ such that $\mathcal{R}(f, \lambda, \infty) = \mathcal{R}(f, \lambda, \rho^*)$. Now let's focus on ρ_- and theoretically compare MM-REx and RO. According to Proposition 2 and 3, RO can bound all possible linear combinations of the training risks on $\mathcal{Q}_m(\infty, C'(f, \mathcal{S}, 0))$. At the same time, MM-REx can deal with the robust region $\mathcal{Q}_m(\infty, C'(f, \mathcal{S}, \lambda))$, which is much larger than that of RO. Hence the factor $(m\lambda + 1)^2$ in $\rho_- = C'(f, \mathcal{S}, \lambda)$ represents the potential benefit of risk extrapolation, which significantly enlarges $\mathcal{Q}_m(\infty, C'(f, \mathcal{S}, 0))$. In summary, REx is more robust than RO by enlarging the robust region $\mathcal{Q}_m$.

3.3 Risk Variance Penalization

Combining the arguments in Section 3.1 and 3.2, we know that: (1) the quasi-distributional robustness framework unifies ERM, RO and MM-REx; (2) the tuning parameters λ and ρ govern each other. These motivate us to consider the transformation from λ to ρ, and to use the following formulation:

$$\mathcal{R}(f, +\infty, \rho*) = \mathcal{R}(f, C(f, S, \rho^*), \rho*) = \mathcal{R}(f, \lambda, \rho^*) = \mathcal{R}(f, \lambda, +\infty).$$

Thus we suggest a novel method: $\hat{f} = \arg\min_f \mathcal{R}_{RVP}(f)$ with

$$\mathcal{R}_{RVP}(f) = R(f, \bar{S}) + \sqrt{\frac{2\rho}{m}\text{Var}(f, \mathcal{S})},$$

which is a risk-variance-penalized function and ρ is a tuning parameter. We call the method by Risk Variance Penalization (RVP).

According to Section 3.2, RVP is equivalent to MM-REx. In addition, $\mathcal{R}_{RVP}$ is very similar to $\mathcal{R}_{V-REx}$. This explains the connection between MM-REx and V-REx. The proposed objective function is simple and easy to compute, because its penalty has no maximization and gradient. This is also the advantage of V-REx. Comparing to IRM and the two REx methods, the biggest advantage of the proposed method is the interpretable tuning parameter ρ/m, which is the radius of the robust region $\mathcal{Q}_m(+\infty, \rho)$. This interpretation provides an insight to understand the relationship between the distributional robustness and the causality. From the view of distributional robustness, MM-REx (or RVP) just enlarges the robust region of RO. On the other hand, [25] experimentally proves that MM-REx obtains the causal performance on the colored MNIST while RO cannot. Similar to Theorem 1 of [34] and Section 1.1.2 of [35], we foresee that when $\mathcal{Q}_m$ is sufficiently large that contains all interventions on environmental features, REx and RVP obtain stable performance on $\mathcal{P}$ and discover causal features. This is an implication of the robustness of a causal model [1, 17, 32]. In one word, we think that MM-REx tunes λ to determine a robust region, which is sufficiently large to guarantee the causal performance, and RVP gives the upper bound of the MM-REx on the robust region.

4 Experiments

In this section, we provide three examples to show: (i) RVP can learn the invariant features; (ii) the comparison between RVP and V-REx; (iii) the causality behind the risk variance penalty. It is difficult to provide experimental evidence to support (i) and (iii). First, the dataset should contain ample environments. According to the arguments in Section 3, REx and RVP can tune λ to focus on the test environments. Thus the OOD generalization cannot be estimated by the performance on a single test environment. We shall report the test performance on both training and test environments. Second, the environments should be sufficiently diverse. In Section 2.1, we have discussed the OOD generalization of ERM. In some experiments of domain generalization, the transfer performance of ERM is not bad. Third, we should know the causal features and the non-causal features behind the learning task. To prove that a model is based on causal features, the stable test performance is a necessary condition rather than a sufficient condition. On the other hand, Section 3 points out that RVP and RO are not essentially different. Thus, without a non-causal baseline, there is no evidence to prove that the improvement on the test performance is caused by learning invariant (causal) features. Thus, the above mentioned three considerations can rule out many datasets.

In the following, we shall focus on the colored MNIST dataset from [3] to learn a binary classification task. The generating procedure of the colored MNIST is as follow: First, label each image from the original MNIST by its digit. If the digit is from 0 to 4, then label the image with $\tilde{y} = 0$; otherwise, label the image with $\tilde{y} = 1$. Second, obtain the final label y by flipping $\tilde{y}$ with probability P_ϵ. Third, generate the color label z by flipping the final label y with probability P_i, which represents the environment generating procedure. Finally, color the image based on the color label z: red for $z = 1$ and green for $z = 0$. It is easy to see that the digit determines the final label, and the final label determines the color. Thus the causal factor is the digit while the color is the spurious feature. To address the first and second considerations, we use 9 environments corresponding to $P_i = i/10$, $i = 1, \ldots, 9$, and denote $\mathcal{P} = \{P_1, \ldots, P_9\}$. For the third consideration, we take $P_\epsilon = 0.5$ as the non-causal baseline, which implies the final label is unrelated to the digit.

The code of this example is based on [3, 25], only replacing their objective functions with $\mathcal{R}_{RVP}$. We also follow the penalty anneal strategy of [25] that takes λ to be 1 at the first 100 epochs and then increases λ to $10,000$ from 101 to 500 epochs. Here $\lambda = \sqrt{2\rho/m}$ for RVP. The training environments are selected from $\mathcal{P}$. In the testing, we do not take the over-fitting into consideration. For example, if the test performance on a test environment is better than the test performance on training environments, we still think this result is reasonable. We will record the test performance on all environments and use the worst one to measure the OOD generalization.

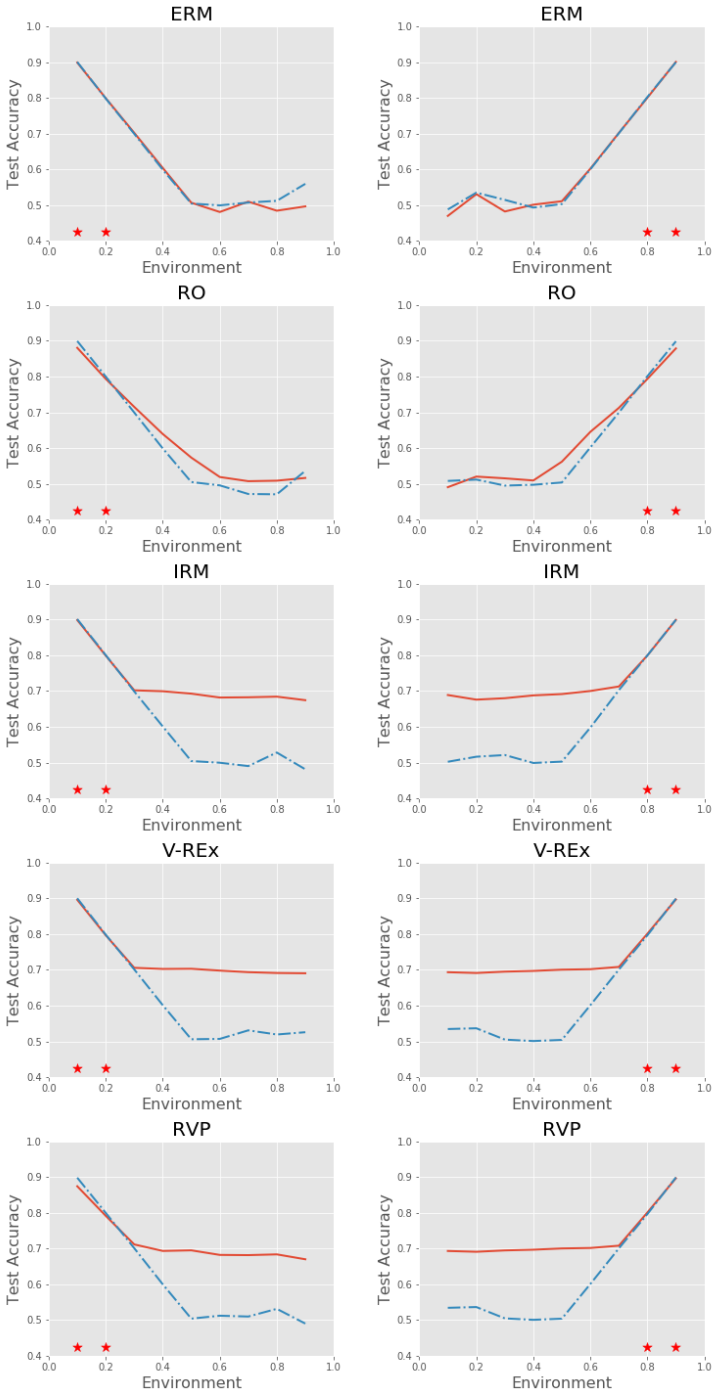

Figure 1: The best test accuracy before 500 epochs. The red solid line and blue dotted line stand for $P_\epsilon = 0.25$ and $P_\epsilon = 0.5$ respectively. The red stars represent the training environments.

Example 1. Consider $P_\epsilon = 0.25$ and $P_\epsilon = 0.5$, and take two training environments from $\mathcal{P}$. The noise level $P_\epsilon = 0.5$ implies that all causal features are removed from the data. Since the test accuracy highly depends on λ and other hyper-parameters, we report the best achievable accuracy during 500 epochs. The results are presented in Figure 1. Throughout this section, the accuracy under $P_\epsilon = 0.5$ is marked by the blue dotted line and the accuracy under $P_\epsilon = 0.25$ is marked by the red solid line. The stars represent the training environments. The performance gap between $P_\epsilon = 0.25$ and $P_\epsilon = 0.5$ represents what learns from the causal features (digit). One can see that IRM, V-REx and RVP can obtain good OOD performance via learning causal features while ERM and RO cannot. What's more, we consider five training environments from $\mathcal{P}$:

$$\mathcal{P}_{tr} = \{0.04, 0.08, 0.12, 0.16, 0.20\}, \quad \text{and} \quad \mathcal{P}'_{tr} = \{0.80, 0.84, 0.88, 0.92, 0.96\}.$$

Here we report the best achievable accuracy of ERM, IRM, V-REx and RVP during 500 epochs. The results are presented in Figure 2. One can see that IRM, REx and RVP can obtain good OOD performance by learning causal features.

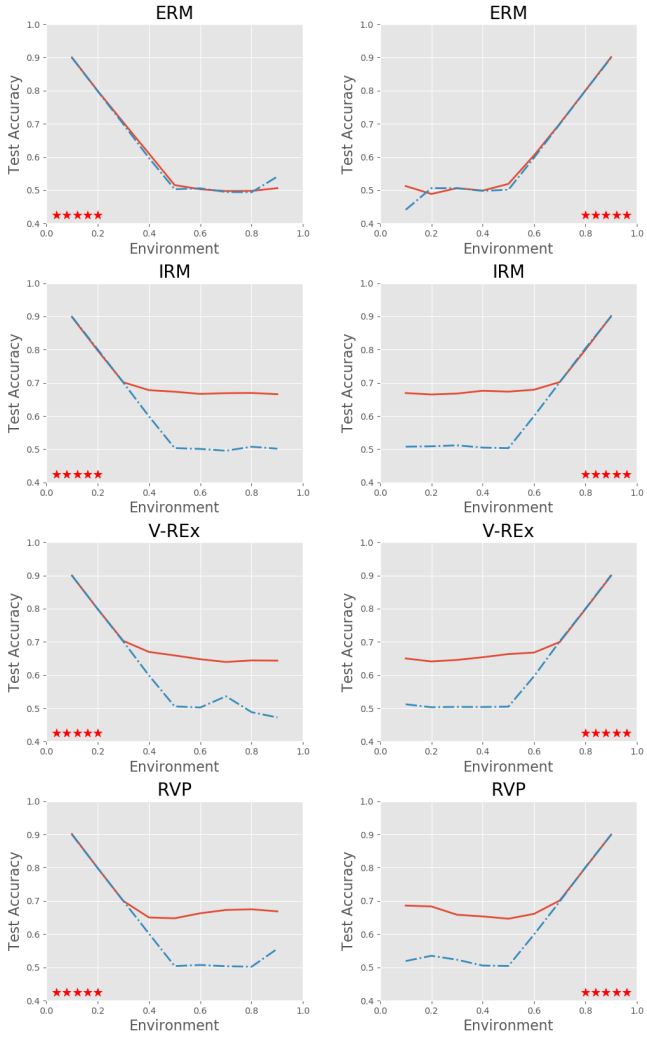

Figure 2: The best test accuracy during 500 epochs.

Example 2. Section 3 illustrates the equivalence between MM-REx and RVP. In this example, we would like to show that RVP uses a proper penalty and has more stable performance than V-REx. We take P_1 and P_2 as the training environments, and compute $R(f, S_1)$ and $R(f, S_2)$ during the training procedure. The trajectory $R(f, S_1)$ vs $R(f, S_2)$ is presented in Figure 3. One can find that the learning procedure of RVP is similar to that of IRM and more stable than the learning procedure of V-REx. What's more, the trajectories of V-REx, RVP and IRM are consistent to the two-stage hypothesis proposed by [25]. In the first stage, the training risks $R(f, S_1)$ and $R(f, S_2)$ keep decreasing by learning predictive features. Then, in the second stage, stable features are obtained by selecting or fine-tuning from the predictive features, which causes the increase of the risks.

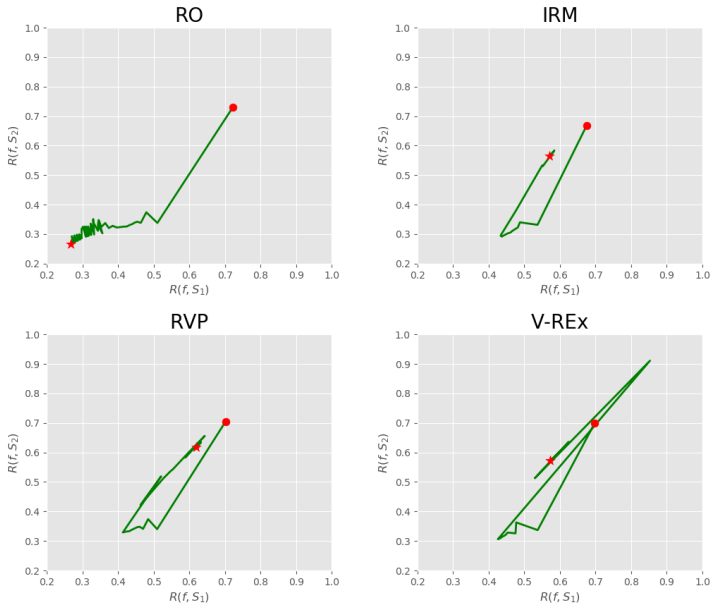

Figure 3: The trajectory $R(f, S_1)$ vs $R(f, S_2)$ during the training procedure. The risks of the random initialized model at epoch 0 is marked by the dot. The star stands for the risks of the final model at epoch 500.

Next we further show that RVP is not sensitive to the tuning procedure of λ. In Example 1, λ is taken to be 10000. In our tuning experiments, we increase λ from 100 to 3000 and fix other hyper-parameters. Then we take P_9 to be the test environment and check the test accuracy of the model at the 500-th epoch. According to the results in Figure 4, one can find that the test performance of RVP is very stable against the tuning parameter λ, even when λ is much smaller than 10000. In contrast, V-REx and IRM require that λ is sufficiently large to achieve the same performance level of $\lambda = 10000$. However, when λ is fixed, V-REx and IRM are more stable than RVP.

Example 3. The existing works [3, 8, 12] check the stable performance on $\mathcal{P}$ to confirm that IRM and REx can learn causal models/features. However the stable performance is a necessary condition and cannot provide further evidence to support that IRM and REx are essentially different to RO. In this example, we design a learning scheme to prove that (1) comparing to RO, REx and RVP remove the non-causal feature (color); (2) the variance penalization strategy is able to discover relationships which do not exist in the training data.

According to the results in Figure 3, we mark the final models on the risk plane $R(\hat{f}, S_1)$ vs $R(\hat{f}, S_2)$. All robust models lie around the line $R(\hat{f}, S_1) = R(\hat{f}, S_2) = r$, see Figure 5. For the three invariant models, $r \approx 0.57$ while $r \approx 0.27$ for RO. In the following, we shall confirm that, from $r \approx 0.27$ to $r \approx 0.57$, the non-causal features (color) are removed. In the training environments P_1 and P_2, y is positively related to z (color label), e.g. $y = z$ with high probability. If the positive relationship

Figure 4: The test accuracy against λ.

decreases as r increases, we may find a predictor based on the negative relationship, e.g. $y = -z$. To this purpose, we design a learning scheme to search models such that $R(f, S_1) \approx R(f, S_2) \approx r$ and push r to a large positive scalar.

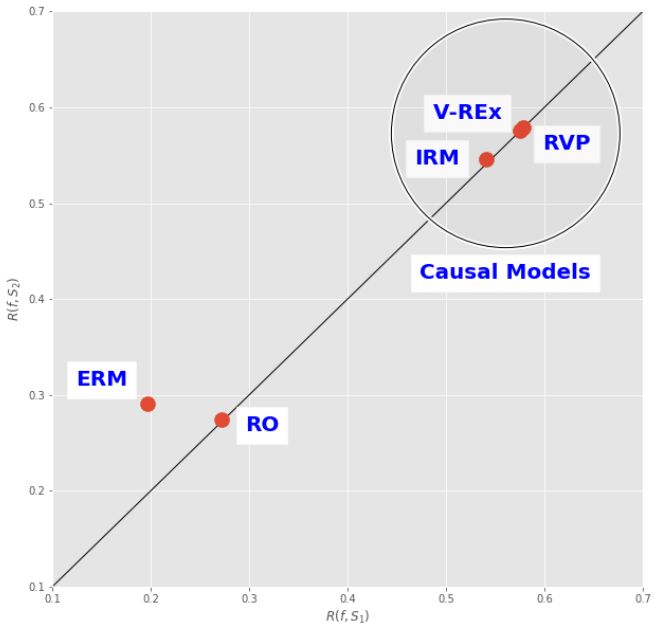

Figure 5: The final models. The black solid line represents $R(f, S_1) = R(f, S_2)$.

Consider the following problem, denoted by elastic learning,

$$\min_{f} \sum_{P_i \in \mathcal{P}_{tr}} R^2(f, S_i), \text{ subject to } R(f, \bar{S}) \geq \lambda$$

which enforces the equality of the risks across training environments. The objective function is formulated into

$$\mathcal{R}_E(f) = \sum_{P_i \in \mathcal{P}_{tr}} R^2(f, S_i) - \lambda R(f, \bar{S}).$$

We select two training environments from $\mathcal{P}$. At the first 100 epochs, $\lambda = 1$ enforces $R(f, S_1) = R(f, S_2) = 0.5$. After that, λ is taken to be $10,000$. Thus the elastic learning will obtain a sequence of models such that $R(\hat{f}, S_1) \approx R(\hat{f}, S_2) \approx r$, $0.5 \leq r \leq 5000$. The best achievable accuracy is presented in Figure 6.

One can find that the best achievable accuracy under $P_\epsilon = 0.25$ is almost the same to the best achievable accuracy under $P_\epsilon = 0.5$. This implies that the elastic learning ignores the causal features (digit). On the other hand, the elastic learning finds out two different prediction rules. At the environments P_1 to P_4, the learned model (models) use the positive correlation $y = z$ to predict the label y. This can be readily learned from the training data. At the environments P_6 to P_9, the learned model (models) use the negative correlation $y = -z$ to predict the label y. Furthermore, by extracting two models at epoch 100 and epoch 200, we find that the two models have learned two quite different prediction rules, which are closed to $y = z$ and $y = 1 - z$ respectively. To make sure that this inference procedure is not one directional, we also take P_8 and P_9 as training environments and also observe these two models, see Figure 6.

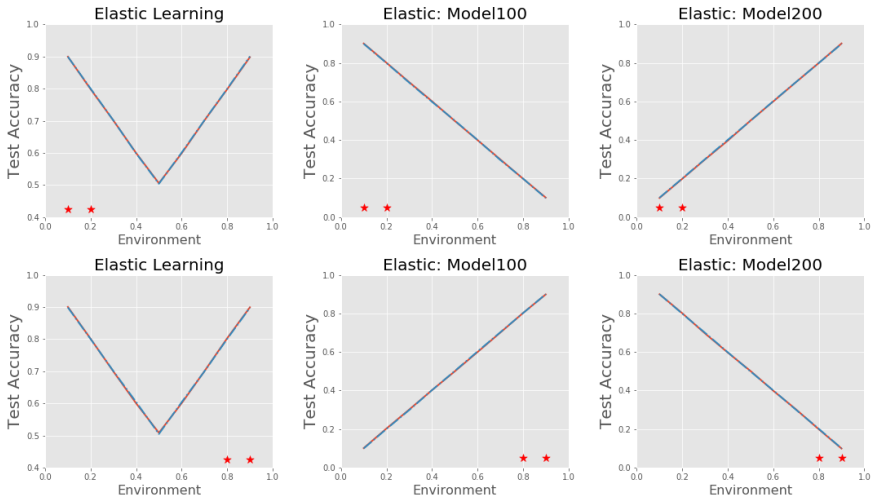

Figure 6: The left panel shows the best test accuracy before 500 epochs. The middle panel and the right panel report the test accuracy of the learned model at epoch 100 and epoch 200. The red stars represent the training environments.

5 Conclusion

This work provides a framework, Quasi-Distributional Robustness, and investigates the causality behind the Risk Extrapolation (REx) from the view of distributional robustness. Furthermore, we propose the Risk Variance Penalization (RVP), which is derived from MM-REx and has an interpretable tuning parameter. We experimentally show that under certain conditions, RVP can learn causal features and is not sensitive to the tuning parameter. In addition, the example 3 in Section 4 introduces a new learning scheme to check the causality from the view of counterfactual inference.

References

[1] John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

[2] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] J Andrew Bagnell. Robust supervised learning. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 714. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

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

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

[7] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[8] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

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

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

[11] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[12] Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*, 2020.

[13] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[14] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019.

[15] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[16] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[17] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

[18] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256, 2018.

[19] Shoubo Hu, Zhitang Chen, Vahid Partovi Nia, CHAN Laiwan, and Yanhui Geng. Causal inference and mechanism clustering of a mixture of additive noise models. In *Advances in Neural Information Processing Systems*, pages 5206–5216, 2018.

[20] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *arXiv preprint arXiv:1903.01672*, 2019.

[21] Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[22] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[23] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[24] Vladimir Koltchinskii *et al.* Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[26] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1617–1626, 2018.

[27] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696, 2018.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[30] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[31] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[32] Judea Pearl. *Causality*. Cambridge university press, 2009.

[33] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[34] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[35] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[36] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[37] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[38] Bob L Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[39] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.