Out-of-Distribution Generalization via Risk Extrapolation

David Krueger [1 2] **Ethan Caballero** [1 2] **Joern-Henrik Jacobsen** [3 4] **Amy Zhang** [1 5 6] **Jonathan Binas** [1 2]
Dinghuai Zhang [1 2] **Remi Le Priol** [1 2] **Aaron Courville** [1 2]

Abstract

Distributional shift is one of the major obstacles when transferring machine learning prediction systems from the lab to the real world. To tackle this problem, we assume that variation across training domains is representative of the variation we might encounter at test time, but also that *shifts at test time may be more extreme in magnitude*. In particular, we show that reducing differences in risk across training domains can reduce a model's sensitivity to a wide range of extreme distributional shifts, including the challenging setting where the input contains both causal and anti-causal elements. We motivate this approach, **Risk Extrapolation (REx)**, as a form of robust optimization over a perturbation set of extrapolated domains (MM-REx), and propose a penalty on the variance of training risks (V-REx) as a simpler variant. We prove that variants of REx can recover the causal mechanisms of the targets, while also providing robustness to changes in the input distribution ("covariate shift"). By trading-off robustness to causally induced distributional shifts and covariate shift, REx is able to outperform alternative methods such as Invariant Risk Minimization in situations where these types of shift co-occur.

1. Introduction

While neural networks often exhibit super-human generalization on the training distribution, they can be extremely sensitive to distributional shift, presenting a major roadblock for their practical application (Su et al., 2019; Engstrom et al., 2017; Recht et al., 2019; Hendrycks & Dietterich, 2019). This sensitivity is often caused by relying on "spurious" features unrelated to the core concept we are trying to learn (Geirhos et al., 2018). For instance, Beery et al. (2018) give the example of an image recognition model failing to correctly classify cows on the beach, since it has learned to make predictions based on the features of the background (e.g. a grassy field) instead of just the animal.

In this work, we consider **out-of-distribution (OOD) generalization**,also known as **domain generalization**, where a model must generalize appropriately to a new test domain for which it has neither labeled nor unlabeled training data. Following common practice (Ben-Tal et al., 2009), we formulate this as optimizing the worst-case performance over a **perturbation set** of possible test domains, $\mathcal{F}$:

$$\mathcal{R}^{\mathrm{OOD}}_{\mathcal{F}}(\theta) = \max_{e \in \mathcal{F}} \mathcal{R}_e(\theta) \tag{1}$$

Since generalizing to arbitrary test domains is impossible, the choice of perturbation set encodes our assumptions about which test domains might be encountered. Instead of making such assumptions *a priori*, we assume access to data from multiple training domains, which can inform our choice of perturbation set. A classic approach for this setting is **group distributionally robust optimization (DRO)** (Sagawa et al., 2019), where $\mathcal{F}$ contains all mixtures of the training distributions. This is mathematically equivalent to considering convex combinations of the training *risks*.

However, we aim for a more ambitious form of OOD generalization, over a larger perturbation set. Our method **minimax Risk Extrapolation (MM-REx)** is an extension of DRO where $\mathcal{F}$ instead contains *affine* combinations of training risks, see Figure 1. Under specific circumstances, MM-REx can be thought of as DRO over a set of extrapolated domains.[1] But MM-REx also unlocks fundamental new generalization capabilities unavailable to DRO.

In particular, focusing on supervised learning, we show that Risk Extrapolation can uncover invariant relationships between inputs X and targets Y. Intuitively, an **invariant relationship** is a statistical relationship which is maintained across all domains in $\mathcal{F}$. Returning to the cow-on-the-beach example, the relationship between the animal and the label is expected to be invariant, while the relationship between the background and the label is not. A model which bases its predictions on such an invariant relationship is said to perform **invariant prediction**.[2]

[1]Mila [2]University of Montreal [3]Vector [4]University of Toronto [5]McGill University [6]Facebook AI Research. Correspondence to: <david.scott.krueger@gmail.com>.

Proceedings of the 38^{th} *International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] We define "extrapolation" to mean "outside the convex hull", see Appendix B for more.

[2] Note this is different from learning an invariant representation

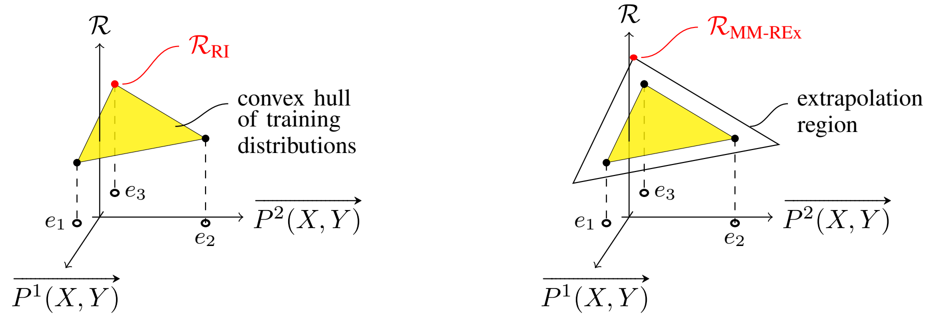
$\mathcal{R}$
$\mathcal{R}_{\text{RI}}$
convex hull of training distributions
e_3
e_1
e_2
$\overrightarrow{P^2(X,Y)}$
$\overrightarrow{P^1(X,Y)}$
$\mathcal{R}$
$\mathcal{R}_{\text{MM-REx}}$
extrapolation region
e_3
e_1
e_2
$\overrightarrow{P^2(X,Y)}$
$\overrightarrow{P^1(X,Y)}$

Figure 1. **Left**: Robust optimization optimizes worst-case performance over the convex hull of training distributions. **Right**: By extrapolating risks, REx encourages robustness to larger shifts and flattens the "risk plane" – the plane containing the training domains (e_1, e_2, and e_3). $\overrightarrow{P^1(X,Y)}$, $\overrightarrow{P^2(X,Y)}$ represent particular directions in the affine space of quasiprobability distributions over (X, Y).

Many domain generalization methods assume $P(Y|X)$ is an invariant relationship, limiting distributional shift to changes in $P(X)$, which are known as **covariate shift** (Ben-David et al., 2010b). This assumption can easily be violated, however. For instance, when Y causes X, a more sensible assumption is that $P(X|Y)$ is fixed, with $P(Y)$ varying across domains (Schölkopf et al., 2012; Lipton et al., 2018). In general, invariant prediction may involve an aspect of causal discovery. Depending on the perturbation set, however, other, more predictive, invariant relationships may also exist (Koyama & Yamaguchi, 2020).

The first method for invariant prediction to be compatible with modern deep learning problems and techniques is **Invariant Risk Minimization (IRM)** (Arjovsky et al., 2019), making it a natural point of comparison. Our work focuses on explaining how REx addresses OOD generalization, and highlighting differences (especially advantages) of REx compared with IRM and other domain generalization methods, see Table 1. Broadly speaking, REx optimizes for robustness to the forms of distributional shift that have been observed to have the largest impact on performance in training domains. This can be a significant advantage over the more focused (but also limited) robustness that IRM targets. For instance, unlike IRM, REx can better encourage robustness to covariate shift (see Section 3 and Figure 5).

Our experiments show that REx significantly outperforms IRM in settings that involve covariate shift and require invariant prediction, including modified versions of CMNIST and simulated robotics tasks from the Deepmind control suite. On the other hand, because REx does not distinguish between underfitting and inherent noise, IRM has an advantage in settings where some domains are intrinsically harder than others. Our contributions include:

1. MM-REx, a novel domain generalization problem formulation suitable for invariant prediction.
2. Demonstrating that REx solves invariant prediction tasks where IRM fails due to covariate shift.
3. Proving that equality of risk across domains can be a sufficient criteria for discovering causal structure.

Figure 1 illustrates how MM-REx encourages equality of risks: extrapolation magnifies any difference in risks that exists between training domains e_1, e_2, e_3. At the same time, encouraging equality of risks enables good OOD generalization to domains that vary in the same directions as the training domains. Extrapolated domains might correspond to more radical interventions than those observed during training. And they can help reveal which features are unreliable; see Figure 2 for a real example. While MM-REx provides a clear link between equalizing risks and OOD generalization, our experiments focus on a simpler method called V-REx, which simply penalizes the risks' variance.

2. Background & Related work

We consider multi-source domain generalization, where our goal is to find parameters θ that perform well on unseen domains, given a set of m training **domains**, $\mathcal{E} = \{e_1, .., e_m\}$, sometimes also called **environments**. We assume the loss function, ℓ is fixed, and domains only differ in terms of their data distribution $P_e(X, Y)$ and dataset D_e. The **risk function** for a given domain/distribution e is:

$$\mathcal{R}_e(\theta) \doteq \mathbb{E}_{(x,y)\sim P_e(X,Y)}\ell(f_\theta(x), y) \quad (2)$$

We refer to members of the set $\{\mathcal{R}_e | e \in \mathcal{E}\}$ as the **training risks** or simply **risks**. Changes in $P_e(X, Y)$ can be categorized as either changes in $P(X)$ (covariate shift), changes in $P(Y|X)$ (concept shift), or a combination. The standard approach to learning problems is **Empirical Risk Minimization (ERM)**, which minimizes the average loss across

(Ganin et al., 2016); see Section 2.2.

Method	Invariant Prediction	Covariate Shift Robustness	Suitable for Deep Learning
DRO	✗	✓	✓
(C-)ADA	✗	(✓)	✓
ICP	✓	✗	✗
IRM	✓	✗	✓
REx	✓	✓	✓

Table 1. A comparison of approaches for OOD generalization. (C-)ADA works for covariate shifts that do not also induce label shift.

all the training examples from all the domains:

$$\mathcal{R}_{\text{ERM}}(\theta) \doteq \mathbb{E}_{(x,y)\sim D}\ \ell(f_\theta(x), y) \quad (3)$$

$$= \frac{1}{D}\sum_{e} |D_e| \mathbb{E}_{(x,y)\sim D_e}\ \ell(f_\theta(x), y) \quad (4)$$

where $D \doteq \cup_{e\in\mathcal{E}} D_e$.

2.1. Robust Optimization

An approach more taylored to OOD generalization is **robust optimization** (Ben-Tal et al., 2009), which aims to optimize a model's worst-case performance over some **perturbation set** of possible data distributions, $\mathcal{F}$ (see Eqn. 1). When only a single training domain is available (**single-source domain generalization**), it is common to assume that $P(Y|X)$ is fixed, and let $\mathcal{F}$ be all distributions within some f-divergence ball of the training $P(X)$ (Hu et al., 2016; Bagnell, 2005). As another example, adversarial robustness can be seen as instead using a Wasserstein ball as a perturbation set (Sinha et al., 2017). The assumption that $P(Y|X)$ is fixed is commonly called the "covariate shift assumption" (Ben-David et al., 2010b); however, we assume that covariate shift and concept shift can co-occur, and refer to this assumption by the novel term **fixed relationship assumption (FRA)**.

In **multi-source domain generalization**, test distributions are often assumed to be mixtures (i.e. convex combinations) of the training distributions (Sagawa et al., 2019; Qian et al., 2018; Hu et al., 2016); this is equivalent to setting $\mathcal{F} \doteq \mathcal{E}$:

$$\mathcal{R}_{\text{RI}}(\theta) \doteq \max_{\substack{\Sigma_e \lambda_e = 1 \\ \lambda_e \geq 0}} \sum_{e=1}^{m} \lambda_e \mathcal{R}_e(\theta) = \max_{e\in\mathcal{E}} \mathcal{R}_e(\theta). \quad (5)$$

We call this objective **Risk Interpolation (RI)**, or, following Sagawa et al. (2019), **(group) Distributionally Robust Optimization (DRO)**. While single-source methods classically assume that the probability of each data-point can vary independently (Hu et al., 2016), DRO yields a much lower dimensional perturbation set, with at most one direction of variation per domain, regardless of the dimensionality of X and Y. It also does not rely on FRA, and can provide robustness to any form of shift in $P(X, Y)$ which occurs across training domains. Minimax-REx is an extension of this approach to affine combinations of training risks.

2.2. Invariant representations vs. invariant predictors

One approach to domain generalization, popular in deep learning, is to view it as a representation learning problem (Bengio et al., 2014).[3] We define an **equipredictive representation**, Φ, as a function of X with the property that $P_e(Y|\Phi)$ is equal, $\forall e \in \mathcal{F}$. In other words, the relationship between such a Φ and Y is fixed across domains. **Invariant relationships** between X and Y are then exactly those that can be written as $P(Y|\Phi(x))$ with Φ an equipredictive representation. A model $\hat{P}(Y|X=x)$ that learns such an invariant relationship is called an **invariant predictor**. Intuitively, an invariant predictor works equally well across all domains in $\mathcal{F}$. The principle of risk extrapolation aims to achieve invariant prediction by enforcing such equality across *training* domains $\mathcal{E}$, and does not rely on explicitly learning an equipredictive representation.

Koyama & Yamaguchi (2020) prove that a *maximal* equipredictive representation – that is, one that maximizes mutual information with the targets, $\Phi^* \doteq \text{argmax}_\Phi I(\Phi, Y)$ – solves the robust optimization problem (Eqn. 1) under fairly general assumptions.[4] When Φ^* is unique, we call the features it ignores **spurious**. The result of Koyama & Yamaguchi (2020) provides a theoretical reason for favoring invariant prediction over the common approach of learning **invariant *representations*** (Pan et al., 2010), which make $P_e(\Phi)$ or $P_e(\Phi|\hat{Y})$ equal $\forall e \in \mathcal{E}$. Popular methods here include **adversarial domain adaptation (ADA)** (Ganin et al., 2016) and **conditional ADA (C-ADA)** (Long et al., 2018). Unlike invariant predictors, invariant representations can easily fail to generalize OOD: ADA forces the predictor to have the same marginal predictions $\hat{P}(Y)$, which is a mistake when $P(Y)$ in fact changes across

[3] See Appendix I for more discussion of relevant work in deep learning.

[4] The first formal definition of an equipredictive representation we found was by Koyama & Yamaguchi (2020), who use the term "(maximal) invariant predictor". We prefer our terminology since: 1) it is more consistent with Arjovsky et al. (2019), and 2) Φ is a representation, not a predictor.

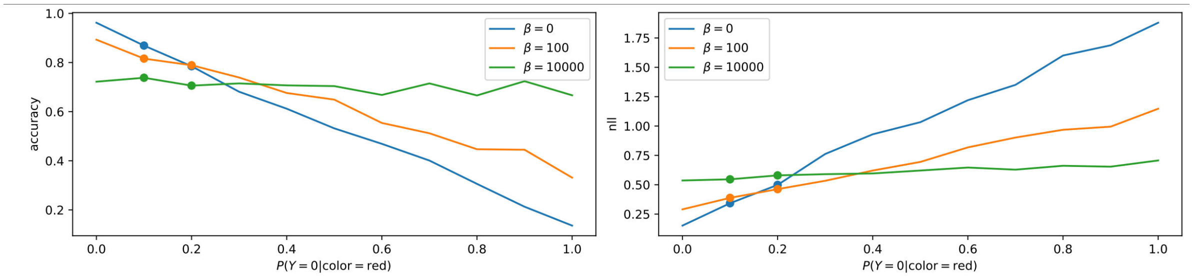

Figure 2. Training accuracies (**left**) and risks (**right**) on colored MNIST domains with varying $P(Y = 0|\text{color} = \text{red})$ after 500 epochs. Dots represent training risks, lines represent test risks on different domains. Increasing the V-REx penalty (β) leads to a flatter "risk plane" and more consistent performance across domains, as the model learns to ignore color in favor of shape-based invariant prediction. Note that $\beta = 100$ gives the best worst-case risk across the 2 training domains, and so would be the solution preferred by DRO (Sagawa et al., 2019). This demonstrates that REx's counter-intuitive propensity to *increase* training risks can be necessary for good OOD performance.

domains (Zhao et al., 2019); C-ADA suffers from the same issue (Tachet et al., 2020).

2.3. Invariance and causality

The relationship between cause and effect is a paradigmatic example of an invariant relationship. Here, we summarize definitions from causal modeling, and discuss causal approaches to domain generalization. We will refer to these definitions for the statements of our theorems in Section 3.2.

Definitions. A **causal graph** is a directed acyclic graph (DAG), where nodes represent variables and edges point from causes to effects. In this work, we use **Structural Causal Models (SCMs)** (sometimes called **Structural Equation Models (SEMs)**), which also specify how the value of a variable is computed given its parents. An SCM, $\mathfrak{C}$, is defined by specifying the **mechanism**, $f_Z : Pa(Z) \to dom(Z)$ for each variable Z.[5] Mechanisms are *deterministic*; noise in Z is represented explicitly via a special noise variable N_Z, and these noise variables are jointly independent. An **intervention**, ι is any modification to the mechanisms of one or more variables; an intervention can introduce new edges, so long as it does not introduce a cycle. $do(X_i = x)$ denotes an intervention which sets X_i to the constant value x (removing all incoming edges). Data can be generated from an SCM, $\mathfrak{C}$, by sampling all of the noise variables, and then using the mechanisms to compute the value of every node whose parents' values are known. This sampling process defines an **entailed distribution**, $P^{\mathfrak{C}}(\mathbf{Z})$ over the nodes $\mathbf{Z}$ of $\mathfrak{C}$. We overload f_Z, letting $f_Z(\mathbf{Z})$ refer to the conditional distribution $P^{\mathfrak{C}}(Z|\mathbf{Z} \setminus \{Z\})$.

Causal approaches to domain generalization. Instead of assuming $P(Y|X)$ is fixed (FRA), works that take a causal approach to domain generalization often assume that the *mechanism* for Y is fixed; our term for this is the **fixed mechanism assumption (FMA)**. Meanwhile, such works assume X may be subject to different (e.g. arbitrary) interventions in different domains (Bühlmann, 2018). We call changes in $P(X, Y)$ resulting from interventions on X **interventional shift**. Interventional shift can involve both covariate shift and/or concept shift. In their seminal work on **Invariant Causal Prediction (ICP)**, Peters et al. (2016) leverage this invariance to learn which elements of X cause Y. ICP and its nonlinear extension (Heinze-Deml et al., 2018) use statistical tests to detect whether the residuals of a linear model are equal across domains. Our work differs from ICP in that:

1. Our method is model agnostic and scales to deep networks.
2. Our goal is OOD generalization, not causal inference. These are not identical: invariant prediction can sometimes make use of non-causal relationships, but when deciding which interventions to perform, a truly causal model is called for.
3. Our learning principle only requires invariance of risks, not residuals. Nonetheless, we prove that this can ensure invariant causal prediction.

A more similar method to REx is **Invariant Risk Minimization (IRM)** (Arjovsky et al., 2019), which shares properties (1) and (2) of the list above. Like REx, IRM also uses a weaker form of invariance than ICP; namely, they insist that the optimal (e.g. linear) classifier must match across domains.[6] Still, REx differs significantly from IRM. While IRM specifically aims for invariant prediction, REx seeks robustness to *whichever* forms of distributional shift are present. Thus, REx is more directly focused on the problem of OOD generalization, and can provide robustness to a

[5] Our definitions follow *Elements of Causal Inference* (Peters et al., 2017); our notation mostly does as well.

[6] In practice, IRMv1 replaces this bilevel optimization problem with a gradient penalty on classifier weights.

wider variety of distributional shifts, including more forms of covariate shift. Also, unlike REx, IRM seeks to match $\mathbb{E}(Y|\Phi(X))$ across domains, not the full $P(Y|\Phi(X))$. This, combined with IRM's relative indifference to covariate shift, make it more effective in cases where different domains or examples are inherently more noisy.

2.4. Fairness

Equalizing risk across different groups (e.g. male vs. female) has been proposed as a definition of **fairness** (Donini et al., 2018), generalizing the equal opportunity definition of fairness (Hardt et al., 2016). Williamson & Menon (2019) propose using the absolute difference of risks to measure deviation from this notion of fairness; this corresponds to our MM-REx, in the case of only two domains, and is similar to V-REx, which uses the variance of risks. However, in the context of fairness, equalizing the risk of training groups is the *goal*. Our work goes beyond this by showing that it can serve as a *method* for OOD generalization.

2.5. On the effectiveness of invariant prediction and domain generalization in deep learning

The practical and theoretical (dis)advantages of various deep learning methods for domain generalization are not yet well understood. In particular, the effectiveness of invariant prediction has not been established, and several works provide negative results.

Theoretically, Rosenfeld et al. (2020) prove that IRM can sometimes successfully recover the optimal invariant predictor, but also that IRM (or REx) can fail to do so when provided too few training domains. Rosenfeld et al. (2020) also state that "IRM and its alternatives fundamentally do not improve over standard Empirical Risk Minimization" in the non-linear setting – non-linearity being a primary motivation for the development of IRM. However, their theorem only demonstrates the *existence* of a non-invariant model which approximately satisfies the IRM criterion. This is roughly analogous to demonstrating the existence of neural network parameters that fit the training set but don't generalize to the test set – neither is sufficient to establish that the method in question fails in practice.

Empirically, Gulrajani & Lopez-Paz (2020) perform a methodologically sound comparison of existing methods (including IRM) over a suite of popular benchmarks called DomainBed, and find that no methods outperform ERM on average;[7] this result suggests that many positive results in previous works could be the result of poor methodology (e.g. tuning on the test distribution). Prior to Gulrajani & Lopez-Paz (2020), we discussed this issue in a preprint version of this work (Krueger et al., 2020), and in private correspondence with the authors of Arjovsky et al. (2019), after noting that their CMNIST experiments tune on the test set.[8] Gulrajani & Lopez-Paz (2020) also suggest that domain generalization research might benefit from more realistic benchmarks. In more recent work, Koh et al. (2021) collect a set of such benchmarks called WILDS, and several works (Wald et al., 2021; Shi et al., 2021; Robey et al., 2021) demonstrate significant performance improvements are possible on WILDS. The method of Wald et al. (2021) in particular encourages invariant prediction by seeking a model that is calibrated on all training domains.

3. Risk Extrapolation

Before discussing algorithms for REx and theoretical results, we first expand on our high-level explanations of what REx does, what kind of OOD generalization it promotes, and how. The principle of Risk Extrapolation (REx) has two aims:

1. Reducing training risks
2. Increasing similarity of training risks

In general, these goals can be at odds with each other; decreasing the risk in the domain with the lowest risk also decreases the overall similarity of training risks. Thus methods for REx may seek to *increase* risk on the best performing domains. While this is counter-intuitive, it can be necessary to achieve good OOD generalization, as Figure 2 demonstrates. From a geometric point of view, encouraging equality of risks flattens the "risk plane" (the affine span of the training risks, considered as a function of the data distribution, see Figures 1 and 2). While this can result in higher training risks, it also means that the risk changes less if the distributional shifts between training domains are magnified at test time.

Figure 2 illustrates how flattening the risk plane can promote OOD generalization on real data, using the Colored MNIST (CMNIST) task as an example (Arjovsky et al., 2019). In the CMNIST training domains, the color of a digit is more predictive of the label than the shape is. But because the correlation between color and label is not invariant, predictors that use the color feature achieve different risk on different domains. By enforcing equality of risks, REx prevents the model from using the color feature enabling successful generalization to the test domain where the correlation between color and label is reversed.

Probabilities vs. Risks. Figure 3 depicts how the extrapolated risks considered in MM-REx can be translated into a corresponding change in $P(X, Y)$, using an example of pure covariate shift. Training distributions can be thought

[7] V-REx is no exception, see Section 4.3

[8] See the official code release.

of as points in an affine space with a dimension for every possible value of (X, Y); see Appendix C.1 for an example. Because the risk is linear w.r.t. $P(x, y)$, a convex combination of risks from different domains is equivalent to the risk on a domain given by the mixture of their distributions. The same holds for the affine combinations used in MM-REx, with the caveat that the negative coefficients may lead to negative probabilities, making the resulting $P(X, Y)$ a ***quasi*****probability distribution**, i.e. a signed measure with integral 1. We explore the theoretical implications of this in Appendix E.2.

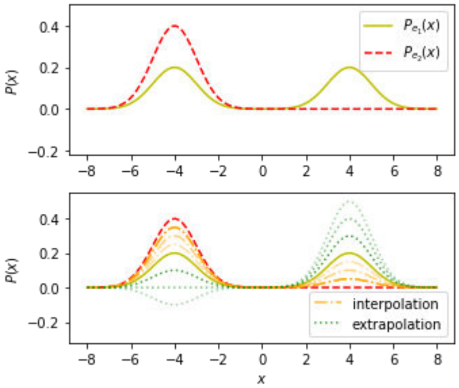

Figure 3. Extrapolation can yield a distribution with *negative* $P(x)$ for some x. **Top:** $P(x)$ for domains e_1 and e_2. **Bottom:** Pointwise interpolation/extrapolation of $P^{e_1}(x)$ and $P^{e_2}(x)$. Since MM-REx target worst-case robustness across extrapolated domains, it can provide robustness to such shifts in P(X) (covariate shift).

Covariate Shift. When only $P(X)$ differs across domains (i.e. FRA holds), as in Figure 3, then $\Phi(x) = x$ is already an equipredictive representation, and so *any* optimal predictor is an invariant predictor. Arjovsky et al. (2019) recognize this limitation of IRM in what they call the "realizable" case. Yet even when capacity is too limited to learn the optimal predictor, invariant prediction does not encourage spending more capacity on low-density regions of the input space, which can lead to poor performance if rare examples become more common; this issue can arise even when FRA does not hold. REx can provide such encouragement, however, as shown in Appendix C.2.

3.1. Methods of Risk Extrapolation

We now formally describe the **Minimax REx (MM-REx)** and **Variance-REx (V-REx)** techniques for risk extrapolation. While we use MM-REx to build geometric intuition and emphasize the relationship with prior work such as Sagawa et al. (2019), we believe V-REx is likely a more practical algorithm, as it has a smoother optimization landscape, see Figure 3.1.[9] See Appendix F for more on the relationship between V-REx and MM-REx.

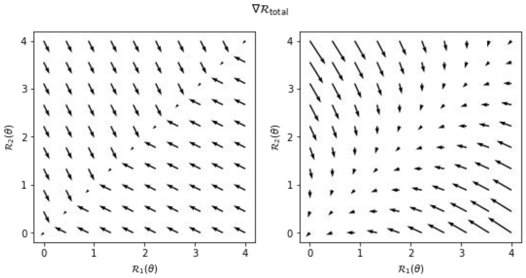

Figure 4. Vector fields of the gradient evaluated at different values of training risks $\mathcal{R}_1(\theta)$, $\mathcal{R}_2(\theta)$. We compare the gradients for $\mathcal{R}_{\text{MM-REx}}$ (**left**) and $\mathcal{R}_{\text{V-REx}}$ (**right**). Note that for $\mathcal{R}_{\text{V-REx}}$, the gradient vectors curve smoothly towards the direction of the origin, as they approach the diagonal (where training risks are equal); this leads to a smoother optimization landscape.

Minimax-REx performs robust learning over a perturbation set of *affine* combinations of training risks with bounded coefficients:

$$\mathcal{R}_{\text{MM-REx}}(\theta) \doteq \max_{\substack{\Sigma_e \lambda_e = 1 \\ \lambda_e \geq \lambda_{\min}}} \sum_{e=1}^{m} \lambda_e \mathcal{R}_e(\theta) \tag{6}$$

$$= (1 - m\lambda_{\min}) \max_e \mathcal{R}_e(\theta) + \lambda_{\min} \sum_{e=1}^{m} \mathcal{R}_e(\theta) \,, \tag{7}$$

where m is the number of domains, and the hyperparameter $\lambda_{\min}$ controls how much we extrapolate. For negative values of $\lambda_{\min}$, MM-REx places negative weights on the risk of all but the worst-case domain, and as $\lambda_{\min} \to -\infty$, this criterion enforces strict equality between training risks; $\lambda_{\min} = 0$ recovers risk interpolation (RI). Thus, like RI, MM-REx aims to be robust in the direction of variations in $P(X, Y)$ between test domains. However, negative coefficients allow us to extrapolate to more extreme variations. Geometrically, larger values of $\lambda_{\min}$ expand the perturbation set farther away from the convex hull of the training risks, encouraging a flatter "risk-plane" (see Figure 2).

While MM-REx makes the relationship to RI/RO clear, we found using the variance of risks as a regularizer (V-REx) simpler, stabler, and more effective:

$$\mathcal{R}_{\text{V-REx}}(\theta) \doteq \beta \ \mathrm{Var}(\{\mathcal{R}_1(\theta), ..., \mathcal{R}_m(\theta)\}) + \sum_{e=1}^{m} \mathcal{R}_e(\theta) \tag{8}$$

Here $\beta \in [0, \infty)$ controls the balance between reducing average risk and enforcing equality of risks, with $\beta = 0$ recovering ERM, and $\beta \to \infty$ leading V-REx to focus entirely on making the risks equal.

[9] Applying Algorithm 1 from Sagawa et al. (2019) to MM-REx could be an alternative method of reducing optimization difficulties.

While exact equality of risks might be desirable in principle in some settings (e.g. as discussed in Section 3.2, in practice, we treat β and $\lambda_{\min}$ as hyperparameters which effectively determine the size of the perturbation set. Conceptually, strict equality is undesirable in practice, since finite datasets make it impossible to determine if differences in training risks are due to the predictor being non-invariant or simply due to sample noise.

3.2. Theoretical Conditions for REx to Perform Causal Discovery

We now prove that exactly equalizing training risks (as incentivized by REx) leads a model to learn the causal mechanism[10] of Y under assumptions similar to those of Peters et al. (2016), namely:

1. The causes of Y are observed, i.e. $Pa(Y) \subseteq X$.
2. Domains correspond to interventions on X.
3. Homoskedasticity (a slight generalization of the additive noise setting assumed by Peters et al. (2016)). We say an SCM $\mathfrak{C}$ is **homoskedastic** (with respect to a loss function ℓ), if the Bayes error rate of $\ell(f_Y(x), f_Y(x))$ is the same for all $x \in \mathcal{X}$.[11]

The contribution of our theory (vs. ICP) is to prove that equalizing risks is sufficient to learn the causes of Y. In contrast, they insist that the entire distribution of error residuals (in predicting Y) be the same across domains. The primary purpose of these results is merely to help explain why REx can encourage invariant prediction. In particular, we do not provide any guarantees for the settings where we imagine REx being applied (and which our experiments tackle) – namely, deep networks and finitely many training domains. We provide proof sketches here and complete proofs in the appendix.

Theorem 1 demonstrates a practical result: we can identify a linear SCM model using REx with a number of domains linear in the dimensionality of X.

Theorem 1. *Given a Linear SEM, $X_i \leftarrow \sum_{j \neq i} \beta_{(i,j)} X_j + \varepsilon_i$, with $Y \doteq X_0$, and a predictor $f_\beta(X) \doteq \sum_{j:j>0} \beta_j X_j + \varepsilon_j$ that satisfies REx (with mean-squared error) over a perturbation set of domains that contains 3 distinct $do()$ interventions for each $X_i : i > 0$. Then $\beta_j = \beta_{0,j}, \forall j$.*

Proof Sketch. We adapt the proof of Theorem 4i from Peters et al. (2016). They show that matching the residual errors across observational and interventional domains forces the model to learn f_Y. We use the weaker condition of matching risks to derive a quadratic equation that the $do()$ interventions must satisfy for any model other than f_Y. Since there are at most 2 solutions to a quadratic equation, insisting on equality of risks across 3 distinct $do()$ interventions forces the model to learn f_Y.

Given the assumption that a predictor satisfies REx over *all* interventions that do not change the mechanism of Y, we can prove a much more general result. We now consider an arbitrary SCM, $\mathfrak{C}$, generating Y and X, and let $\mathcal{E}^I$ be the set of domains corresponding to arbitrary interventions on X, similarly to Peters et al. (2016).

Theorem 2. *Suppose ℓ is a (strictly) proper scoring rule. Then a predictor that satisfies REx over $\mathcal{E}^I$ uses $f_Y(x)$ as its predictive distribution on input x for all $x \in \mathcal{X}$.*

Proof Sketch. Since the distribution of Y given its parents doesn't depend on the domain, f_Y can make reliable point-wise predictions across domains. This translates into equality of risk across domains when the overall difficulty of the examples is held constant across domains, e.g. by assuming homoskedasticity.[12] While a different predictor might do a better job on *some* domains, we can always find a domain where it does worse than f_Y, and so f_Y is both unique and optimal.

Remark. Theorem 2 is only meant to provide insight into how the REx principle relates to causal invariance; the perturbation set in this theorem is uncountably infinite. Note, however, that even in this setting, the ERM principle does *not*, in general, recover the causal mechanism for Y. Rather, the ERM solution depends on the distribution over domains. For instance, if all but an $\epsilon \to 0$ fraction of the data comes from the CMNIST training domains, then ERM will learn to use the color feature, just as in original the CMNIST task. Furthermore, while access to this perturbation set implies access to the test domain, it does not mean we know *which* domain(s) we will encounter at test time; and thus, we cannot simply train on the test domain(s) of interest.

4. Experiments

We evaluate REx and compare with IRM on a range of tasks requiring OOD generalization. REx provides generalization benefits and outperforms IRM on a wide range of tasks, including: i) variants of the Colored MNIST (CMNIST) dataset (Arjovsky et al., 2019) with extra covariate shift, ii) continuous control tasks with partial observability and spurious features, iii) domain generalization tasks from

[10] See Section 2.3 for background on causality, including definitions and notation.

[11] Note that our definitions of **homoskedastic/heteroskedastic** do *not* correspond to the types of domains constructed in Arjovsky et al. (2019), Section 5.1, but rather are a generalization of the definitions of these terms as commonly used in statistics. Specifically, for us, *hetero*skedasticity means that the "predictability" (e.g. variance) of Y differs across inputs x, whereas for Arjovsky et al. (2019), it means the predictability of Y at a given input varies across *domains*; we refer to this second type as *domain*-homo/heteroskedasticity for clarity.

[12] Note we could also assume no covariate shift in order to fix the difficulty, but this seems hard to motivate in the context of interventions on X, which can change $P(X)$.

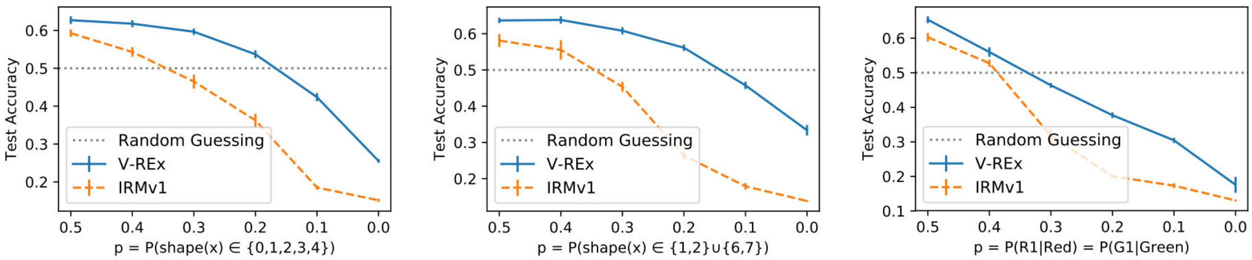

Figure 5. REx outperforms IRM on Colored MNIST variants that include covariate shift. The x-axis indexes increasing amount of shift between training distributions, with $p = 0$ corresponding to disjoint supports. **Left:** class imbalance, **Center:** shape imbalance, **Right:** color imbalance.

Method	**train acc**	**test acc**
V-REx (ours)	71.5 ± 1.0	~~$\mathbf{68.7 \pm 0.9}$~~
IRM	70.8 ± 0.9	~~66.9 ± 2.5~~
MM-REx (ours)	72.4 ± 1.8	~~66.1 ± 1.5~~
RI	88.9 ± 0.3	~~22.3 ± 4.6~~
ERM	87.4 ± 0.2	~~17.1 ± 0.6~~
Grayscale oracle	73.5 ± 0.2	~~73.0 ± 0.4~~
Optimum	75	75
Chance	50	50

Table 2. Accuracy (percent) on Colored MNIST. REx and IRM learn to ignore the spurious color feature. ~~Strikethrough~~ results achieved via tuning on the test set.

the DomainBed suite (Gulrajani & Lopez-Paz, 2020). On the other hand, when the inherent noise in Y varies across environments, IRM succeeds and REx performs poorly.

4.1. Colored MNIST

Arjovsky et al. (2019) construct a binary classification problem (with 0-4 and 5-9 each collapsed into a single class) based on the MNIST dataset, using color as a spurious feature. Specifically, digits are either colored red or green, and there is a strong correlation between color and label, which is reversed at test time. The goal is to learn the causal "digit shape" feature and ignore the anti-causal "digit color" feature. The learner has access to three domains:

1. A training domain where green digits have a 80% chance of belonging to class 1 (digits 5-9).
2. A training domain where green digits have a 90% chance of belonging to class 1.
3. A test domain where green digits have a 10% chance of belonging to class 1.

We use the exact same hyperparameters as Arjovsky et al. (2019), only replacing the IRMv1 penalty with MM-REx or V-REx penalty.[13] These methods all achieve similar performance, see Table 2.

CMNIST with extra covariate shift. To test our hypothesis that REx should outperform IRM under covariate shift, we construct 3 variants of the CMNIST dataset. These experiments include the original interventional shift of the original CMNIST (i.e. $P(Green|Y = 1)$ still differs across training domains) plus these extra forms of covariate shift:

1. **Class imbalance:** varying $p = P(\text{shape(x)} \in \{0, 1, 2, 3, 4\})$; as in Wu et al. (2020).
2. **Digit imbalance:** varying $p = P(\text{shape(x)} \in \{1, 2\} \cup \{6, 7\})$; digits 0 and 5 are removed.
3. **Color imbalance:** We use 2 versions of each color, for 4 total channels: R_1, R_2, G_1, G_2. We vary $p = P(R_1|Red) = P(G_1|Green)$.

While (1) also induces change in $P(Y)$, (2) and (3) induce *only* covariate shift in the causal shape and anti-causal color features (respectively). We compare across several levels of imbalance, $p \in [0, 0.5]$, using the same hyperparameters from Arjovsky et al. (2019), and plot the mean and standard error over 3 trials.

V-REx significantly outperforms IRM in every case, see Figure 5. In order to verify that these results are not due to bad hyperparameters for IRM, we perform a random search that samples 340 unique hyperparameter combinations for each value of p, and compare the the number of times each method achieves better than chance-level (50% accuracy). Again, V-REx outperforms IRM; in particular, for small values of p, IRM never achieves better than random chance performance, while REx does better than random in 4.4%/23.7%/2.0% of trials, respectively, in the class/digit/color imbalance scenarios for $p = 0.1/0.1/0.2$. This indicates that REx can achieve good OOD generalization in settings involving both interventional shift and more intense covariate shift, whereas IRM struggles to do so.

4.2. Toy Structural Equation Models (SEMs)

REx's sensitivity to covariate shift can also be a weakness when reallocating capacity towards domains with higher

[13] When there are only 2 domains, MM-REx is equivalent to a penalty on the Mean Absolute Error (MAE), see Appendix F.2.2.

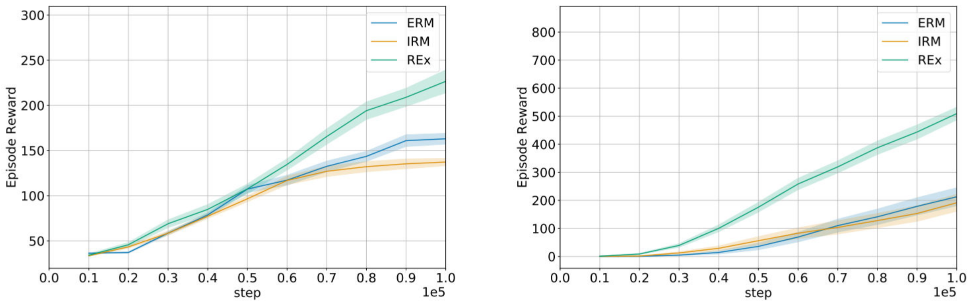

Figure 6. Performance and standard error on `walker_walk` (**top**), `finger_spin` (**bottom**).

Algorithm	ColoredMNIST	VLCS	PACS	OfficeHome
ERM	52.0 ± 0.1	77.4 ± 0.3	85.7 ± 0.5	67.5 ± 0.5
IRM	51.8 ± 0.1	78.1 ± 0.0	84.4 ± 1.1	66.6 ± 1.0
V-REx	52.1 ± 0.1	77.9 ± 0.5	85.8 ± 0.6	66.7 ± 0.5

Table 3. REx, IRM, and ERM all perform comparably on a set of domain generalization benchmarks.

risk does not help the model reduce their risk, e.g. due to irreducible noise. We illustrate this using the linear-Gaussian structural equation model (SEM) tasks introduced by Arjovsky et al. (2019). Like CMNIST, these SEMs include spurious features by construction. They also introduce 1) heteroskedasticity, 2) hidden confounders, and/or 3) elements of X that contain a mixture of causes and effects of Y. These three properties highlight advantages of IRM over ICP (Peters et al., 2016), as demonstrated empirically by Arjovsky et al. (2019). REx is also able to handle (2) and (3), but it performs poorly in the heteroskedastic tasks. See Appendix G.2 for details and Table 5 for results.

4.3. Domain Generalization in the DomainBed Suite

Methodologically, it is inappropriate to assume access to the test environment in domain generalization settings, as the goal is to find methods which generalize to *unknown* test distributions. Gulrajani & Lopez-Paz (2020) introduced the DomainBed evaluation suite to rigorously compare existing approaches to domain generalization, and found that no method reliably outperformed ERM. We evaluate V-REx on DomainBed using the most commonly used training-domain validation set method for model selection. Due to limited computational resources, we limited ourselves to the 4 cheapest datasets. Results of baseline are taken from Gulrajani & Lopez-Paz (2020), who compare with more methods. Results in Table 3 give the average over 3 different train/valid splits.

4.4. Reinforcement Learning with partial observability and spurious features

Finally, we turn to reinforcement learning, where covariate shift (potentially favoring REx) and heteroskedasticity (favoring IRM) both occur naturally as a result of randomness in the environment and policy. In order to show the benefits of invariant prediction, we modify tasks from the Deepmind Control Suite (Tassa et al., 2018) to include spurious features in the observation, and train a Soft Actor-Critic (Haarnoja et al., 2018) agent. REx outperforms both IRM and ERM, suggesting that REx's robustness to covariate shift outweighs the challenges it faces with heteroskedasticity in this setting, see Figure 6. We average over 10 runs on `finger_spin` and `walker_walk`, using hyperparameters tuned on `cartpole_swingup` (to avoid overfitting). See Appendix for details and further results.

5. Conclusion

We have demonstrated that REx, a method for robust optimization, can provide robustness and hence out-of-distribution generalization in the challenging case where X contains both causes and effects of Y. In particular, like IRM, REx can perform causal identification, but REx can also perform more robustly in the presence of covariate shift. Covariate shift is known to be problematic when models are misspecified, or when training data is limited or does not cover areas of the test distribution (Ben-David et al., 2010b). As such situations are inevitable in practice, REx's ability to outperform IRM in scenarios involving a combination of covariate shift and interventional shift makes it a powerful approach.

References

Albuquerque, I., Naik, N., Li, J., Keskar, N., and Socher, R. Improving out-of-distribution generalization via multitask self-supervised pretraining. *arXiv:2003.13525*, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views. *arXiv:1906.00910*, 2019.

Bagnell, J. A. Robust Supervised Learning. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, pp. 714–719. AAAI Press, 2005. ISBN 157735236x.

Beery, S., Van Horn, G., and Perona, P. Recognition in Terra Incognita. *Lecture Notes in Computer Science*, pp. 472–489, 2018. ISSN 1611-3349.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010a.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, 2010b.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton University Press, 2009.

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538*, 2014.

Bühlmann, P. Invariance, Causality and Robustness. *arXiv:1812.08233*, 2018.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. AutoAugment: Learning Augmentation Policies from Data. *arXiv:1805.09501*, 2018.

Desjardins, G., Simonyan, K., Pascanu, R., and Kavukcuoglu, K. Natural neural networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada*, pp. 2071–2079, 2015. arXiv:1507.00210.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical Risk Minimization under Fairness Constraints. *arXiv:1802.08626*, 2018.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. *arXiv:1712.02779*, 2017.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231*, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

Gowal, S., Qin, C., Huang, P.-S., Cemgil, T., Dvijotham, K., Mann, T., and Kohli, P. Achieving Robustness in the Wild via Adversarial Mixing with Disentangled Representations. *arXiv:1912.03192*, 2019.

Gulrajani, I. and Lopez-Paz, D. In Search of Lost Domain Generalization. *arXiv:2007.01434*, 2020.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Haffner, P. Escaping the Convex Hull with Extrapolated Vector Machines. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 753–760, Cambridge, MA, USA, 2001. MIT Press.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413*, 2016.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

He, Y., Shen, Z., and Cui, P. Towards Non-I.I.D. Image Classification: A Dataset and Baselines. *arXiv:1906.02899*, 2019.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6(2), Sep 2018. ISSN 2193-3685. doi: 10.1515/jci-2017-0016. URL http://dx.doi.org/10.1515/jci-2017-0016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv:1903.12261*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv:1812.04606*, 2018.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. *arXiv:1906.12340*, 2019a.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv:1912.02781*, 2019b.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670*, 2018.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does Distributionally Robust Supervised Learning Give Robust Classifiers? *arXiv:1611.02041*, 2016.

Ilse, M., Tomczak, J. M., and Forré, P. Designing Data Augmentation for Simulating Interventions. *arXiv:2005.01856*, 2020.

Johansson, F. D., Sontag, D., and Ranganath, R. Support and Invertibility in Domain-Invariant Representations. *arXiv:1903.03448*, 2019.

King, G. and Zeng, L. The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159, 2006.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. Wilds: A benchmark of in-the-wild distribution shifts, 2021.

Koyama, M. and Yamaguchi, S. Out-of-Distribution Generalization with Maximal Invariant Predictor. *arXiv:2008.01883*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. L., and Courville, A. Out-of-Distribution Generalization via Risk Extrapolation (REx). *arXiv:2003.00688v1*, 2020.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Lipton, Z. C., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. *arXiv:1802.03916*, 2018.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 1647–1657, Red Hook, NY, USA, 2018. Curran Associates Inc.

Meinshausen, N., Bühlmann, P., et al. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Mouli, S. C. and Ribeiro, B. Neural networks for learning counterfactual g-invariances from single environments, 2021.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

Peters, J., Bühlmann, P., and Meinshausen, N. 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.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Qian, Q., Zhu, S., Tang, J., Jin, R., Sun, B., and Li, H. Robust Optimization over Multiple Domains. *arXiv:1805.07588*, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? *arXiv:1902.10811*, 2019.

Robey, A., Pappas, G. J., and Hassani, H. Model-based domain generalization. *arXiv preprint arXiv:2102.11436*, 2021.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The Risks of Invariant Risk Minimization. *International Conference on Learning Representations*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv:1911.08731*, 2019.

Sahoo, S. S., Lampert, C. H., and Martius, G. Learning Equations for Extrapolation and Control. *arXiv:1806.07259*, 2018.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On Causal and Anticausal Learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 459–466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Shi, Y., Seely, J., Torr, P. H. S., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization, 2021.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571*, 2017.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Tachet, R., Zhao, H., Wang, Y.-X., and Gordon, G. Domain adaptation with conditional distribution matching and generalized label shift, 2020.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. DeepMind Control Suite. Technical report, DeepMind, jan 2018. arXiv:1801.00690.

Tian, Y., Krishnan, D., and Isola, P. Contrastive Multiview Coding. *arXiv:1906.05849*, 2019.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

van den Oord, A., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*, 2018.

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. On calibration and out-of-domain generalization, 2021.

Wang, H., He, Z., Lipton, Z. C., and Xing, E. P. Learning robust representations by projecting superficial statistics out. *arXiv:1903.06256*, 2019.

Williamson, R. C. and Menon, A. K. Fairness risk measures. *arXiv:1901.08665*, 2019.

Wu, X., Guo, Y., Chen, J., Liang, Y., Jha, S., and Chalasani, P. Representation Bayesian Risk Decompositions and Multi-Source Domain Adaptation. *arXiv:2004.10390*, 2020.

Xu, K., Zhang, M., Li, J., Du, S. S., ichi Kawarabayashi, K., and Jegelka, S. How neural networks extrapolate: From feedforward to graph neural networks, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412*, 2017.

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