On Calibration and Out-of-domain Generalization

Yoav Wald[*]
Johns Hopkins University
yoav.wald@gmail.com

Amir Feder[*]
Technion
amirfeder@gmail.com

Daniel Greenfeld
Jether Energy Research
danielgreenfeld3@gmail.com

Uri Shalit
Technion
urishalit@technion.ac.il

Abstract

Out-of-domain (OOD) generalization is a significant challenge for machine learning models. Many techniques have been proposed to overcome this challenge, often focused on learning models with certain invariance properties. In this work, we draw a link between OOD performance and model calibration, arguing that calibration across multiple domains can be viewed as a special case of an invariant representation leading to better OOD generalization. Specifically, we show that under certain conditions, models which achieve *multi-domain calibration* are provably free of spurious correlations. This leads us to propose multi-domain calibration as a measurable and trainable surrogate for the OOD performance of a classifier. We therefore introduce methods that are easy to apply and allow practitioners to improve multi-domain calibration by training or modifying an existing model, leading to better performance on unseen domains. Using four datasets from the recently proposed WILDS OOD benchmark [23], as well as the Colored MNIST dataset [21], we demonstrate that training or tuning models so they are calibrated across multiple domains leads to significantly improved performance on unseen test domains. We believe this intriguing connection between calibration and OOD generalization is promising from both a practical and theoretical point of view.

1 Introduction

Machine learning models have recently displayed impressive success in a plethora of fields [19, 9, 41]. However, as models are typically only trained and tested on in-domain (ID) data, they often fail to generalize to out-of-domain (OOD) data [23]. The problem is especially pressing when deploying machine learning models in the wild, where they are required to perform well under conditions that were not observed during training. For instance, a medical diagnosis system trained on patient data from a few hospitals could fail when deployed in a new hospital.

Many methods have been proposed to improve the OOD generalization of machine learning models. Specifically, there is rapidly growing interest in learning models that display certain invariance properties under distribution shifts and do not rely on spurious correlations in the training data [34, 17, 1]. While highlighting the need for learning robust models, so far these attempts have limited success scaling to realistic high-dimensional data, and in learning truly invariant representations [37, 11, 20].

In this paper, we argue that an alternative and relatively simple approach for learning invariant representations could be achieved through model calibration across multiple domains. Calibration asserts that the probabilities of outcomes predicted by a model match their true probabilities. Our claim is that simultaneous calibration over several domains can be used as an observable indicator

[*]Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

for favorable performance on unseen domains. For example, if we take all patients for whom a classifier outputs a probability of 0.9 for being ill, and in one hospital the true probability of illness in these patients is 0.85 while in the other it is 0.95, then we may suspect the classifier relies on spurious correlations. Intuitively, the features which lead the classifier to predict a probability of 0.9 imply different results under different experimental conditions, suggesting that their correlation with the label is potentially unstable. Conversely, if the true probabilities in both hospitals match the classifier's output, it may be a sign of its robustness.

Our contributions are as follows: We prove that in Gaussian-linear models, under a general-position condition, being concurrently calibrated across a sufficient number of domains guarantees a model has no spurious correlations. We then introduce three methods for encouraging multi-domain calibration in practice. These are, in ascending order of complexity: (i) model selection by a multi-domain calibration score, (ii) robust isotonic regression as a post-processing tool, and (iii) directly optimizing deep nets with a multi-domain calibration objective, based on the method introduced by Kumar et al. [26]. We show that multi-domain calibration achieves the correct invariant classifier in a learning scenario presented by Kamath et al. [20], unlike the objective proposed in Invariant Risk Minimization [1]. Finally, we demonstrate that the proposed approaches lead to significant performance gains on the WILDS benchmark datasets [23], and also succeed on the colored MNIST dataset [21].

2 Calibration and Invariant Classifiers

2.1 Problem Setting

Consider observable features X, a label Y and an environment (or domain) E with sample spaces $\mathcal{X}, \mathcal{Y}, \mathcal{E}$ accordingly. We mostly focus on regression and binary classification, therefore $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = \{0, 1\}$. To lighten notation, our definitions will be given for the binary classification setting and we will point out adjustments to regression where necessary. There is no explicit limitation on $|\mathcal{E}|$, but we assume that training data that has been collected from a finite subset of the possible environments $E_{\text{train}} \subset \mathcal{E}$. The number of training environments is denoted by k, and $E_{\text{train}} = \{e_i\}_{i=1}^{k} \subset \mathcal{E}$, so that our training data is sampled from a distribution $P[X, Y \mid E = e_i] \quad \forall i \in [k]$. Our goal is to learn models that will generalize to new, unseen environments in $\mathcal{E}$.

Ideally, we would like to learn a classifier that is optimal for all environments $\mathcal{E}$. Unfortunately, we only observe data from the limited set E_{train} and even if this set is extremely large, the Bayes optimal classifiers on each environment do not necessarily coincide. Following other recent work [34, 17, 1] we therefore aim for a different goal – learning classifiers whose per-instance output will be stable across environments E, as we explain below.

We assume the data generating process for E, X, Y follows the causal graph in Fig. 1. [2] We differentiate between causal and anti-causal components of X, and further differentiate between the anti-causal variables which are affected or unaffected by E, denoted as $X_{\text{ac-spurious}}$ and $X_{\text{ac-non-spurious}}$, respectively. As an illustrative example, consider again predicting illness across different hospitals. When predicting lung cancer, Y, from patient health records, X_{causal} could be features like smoking. $X_{\text{ac-non-spurious}}$ are symptoms of Y such as infections that appear in chest X-rays, while $X_{\text{ac-spurious}}$ can be marks that technicians put on X-rays as in [51]. Smoking habits may vary across hospital populations, as might lab test policy or X-ray markings; but the influence of smoking on cancer and the manifestation of cancer in an X-ray do not vary by hospital.

We do not assume to know how to partition X into $X_{\text{causal}}, X_{\text{ac-spurious}}, X_{\text{ac-non-spurious}}$. The main assumptions made in the causal graph in Fig. 1 are that there are no hidden variables, and that there is no edge directly from environment E to the label Y. Such an arrow would imply the conditional distribution of Y given X can be arbitrarily different in an unseen environment E, compared to those present in the training set. Note that for simplicity we do not include arrows from X_{causal} to $X_{\text{ac-spurious}}$ and $X_{\text{ac-non-spurious}}$ but they may be included as well.

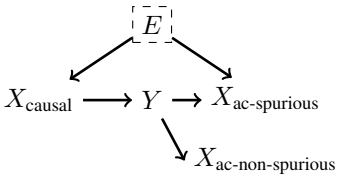


Figure 1: Learning in the presence of causal and anti-causal features. Anti-causal features can be either spurious ($X_{\text{ac-spurious}}$), or non-spurious ($X_{\text{ac-non-spurious}}$).

[2] See Appendix A.5 for a brief introduction to causal graphs.

We will say a representation $\Phi(X)$ contains a *spurious correlation* with respect to the environments E and label Y, if $Y \not\perp\!\!\!\perp E \mid \Phi(X)$; this motivates our naming of $X_{\text{ac-spurious}}$ and $X_{\text{ac-non-spurious}}$ in Fig. 1, as $Y \not\perp\!\!\!\perp E \mid X_{\text{ac-spurious}}$ but $Y \perp\!\!\!\perp E \mid X_{\text{ac-non-spurious}}$. Similar observations have been made by [17, 1]. Having a spurious correlation implies that the relation between $\Phi(X)$ and Y depends on the environment – it is not transferable nor stable across environments. In this work we will simply consider the output $f(X)$ of a classifier $f : \mathcal{X} \to [0, 1]$ as a representation. The crux of this paper is the observation that having $\mathbb{E}[Y \mid f(X), E = e] = f(X)$ for every value of E, i.e. f being a ***calibrated*** classifier across all environments, is equivalent up-to a simple transformation to having $Y \perp\!\!\!\perp E \mid f(X)$, and thus to f having ***no spurious correlations*** with respect to E. We prove this assertion in section 2.2, and as a demonstration of this principle we prove (section 3) that linear models which are calibrated across a diverse set of environments E are guaranteed to discard $X_{\text{ac-spurious}}$ as viable features for prediction.

2.2 Invariance and Calibration on Multiple Domains

We define calibration, along with a straightforward generalization to the multiple environment setting.

Definition 1. *Let $P[X, Y]$ be a joint distribution over the features and label, and $f : \mathcal{X} \to [0, 1]$ a classifier. Then $f(\mathbf{x})$ is calibrated w.r.t to P if for all $\alpha \in [0, 1]$ in the range of f, $\mathbb{E}_P[Y \mid f(X) = \alpha] = \alpha$. In the multiple environments setting, $f(\mathbf{x})$ is calibrated on E_{train} if for all $e_i \in E_{train}$ and α in the range of f restricted to e_i, $\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \alpha$.*

For regression problems, we consider regressors that output estimates for the mean and variance of Y, and say they are calibrated if they match the true values similarly to the definition above. The precise definition can be found in the supplementary material.

We now tie the notion of calibration on multiple environments with OOD generalization, starting with its correspondence with our definition of spurious correlations. Recall that a representation $\Phi(X)$ does not contain spurious correlations if $Y \perp\!\!\!\perp E \mid \Phi(X)$. Treating the output $f(X)$ of a classifier as a representation of the data, and considering classifiers satisfying the above conditional independence with respect to training environments, we arrive at a definition of an invariant classifier.

Definition 2. *Let $f : \mathcal{X} \to [0, 1]$. f is an* invariant classifier *w.r.t E_{train} if for all $\alpha \in [0, 1]$ and environments $e_i, e_j \in E_{train}$, where α is in the range of f restricted to each of them:*

$$\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \mathbb{E}[Y \mid f(X) = \alpha, E = e_j]. \tag{1}$$

Lemma 1 gives the correspondence between invariant classifiers and classifiers calibrated on multiple environments. The proof is in Section A.1 of the supplementary material.

Lemma 1. *If a binary classifier f is invariant w.r.t E_{train}, then there exists some $g : \mathbb{R} \to [0, 1]$ such that (i) $g \circ f$ is calibrated on all training environments, and (ii) the mean squared error of $g \circ f$ on each environment does not exceed that of f. On the other hand, if a classifier is calibrated on all training environments it is also invariant w.r.t E_{train}.*

Now, we can note how the above notion of invariance relates to that of Invariant Risk Minimization [1], where invariance of a representation $\Phi : \mathcal{X} \to \mathcal{H}$ is linked to a shared classifier $\mathbf{w}^* : \mathcal{H} \to [0, 1]$, $\mathbf{w}^* \circ \Phi$ being optimal on all environments w.r.t a loss $l : [0, 1] \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Under the representation $\Phi(X) = f(X)$, and the cross-entropy or squared losses it turns out that the original IRM definition coincides with Equation (1) [3]. Hence we aim for a similar notion of conditional independence, yet we approach it from the point-of-view of calibration. In Section 5 we will see that taking this approach leads to different methods that are highly effective in achieving and assessing invariance. We further note that the original IRM objective was deemed too difficult to optimize by the original IRM authors, leading them to propose an alternative called IRMv1. This alternative however does not capture the full set of required invariances, as shown by [20], whereas we show in section 6.1 that multi-domain calibration does indeed capture the required invariances.

Having established the connection between calibration on multiple environments and invariance, there are several interesting questions and points to consider:

Calibration and sharpness. Calibration alone is not enough to guarantee that a classifier performs well; on a single environment, always predicting $\mathbb{E}[Y]$ will give a perfectly calibrated classifier.

[3] See Observation 2 in [20] for a proof.

Hence, multi-domain calibration should be combined with some sort of guarantee on accuracy. In the calibration literature, this is often referred to as sharpness. To this end, in Section 5 we will propose regularizing models during training or fine-tuning with Calibration Loss Over Environments (CLOvE). Combining this regularizer with standard empirical loss functions helps balance between sharpness and multi-domain calibration. Even without training a new model, we will propose methods for model selection and post-processing that are very easy to apply and help improve multi-domain calibration without a significant effect on the sharpness of the models.

Generalization and dependence on $X_{\text{ac-spurious}}$. Suppose that $f(X)$ is calibrated on E_{train}. Under what conditions does this imply it is calibrated on $\mathcal{E}$? It is easy to show that calibration on several environments entails calibration on any distribution which can be expressed as a linear combination of the distributions underlying said environments. However, can we go beyond that? Given a general set $\mathcal{E}$ we would like to know what conditions and how many training environments are required for calibration to generalize. We also wish to understand when does calibration over a finite set of training environments indeed guarantee that a classifier is free of spurious correlations. We now turn to answer these questions in the setting of linear-Gaussian models.

3 Motivation: a Linear-Gaussian Model

Let us consider data where X is a multivariate Gaussian. Since we will be considering Gaussian data, the set of all environments $\mathcal{E}$ will be parameterized using pairs of real vectors expressing expectations and positive definite matrices of an appropriate dimension expressing covariances: $\mathcal{E} = \{(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\}$.

For two scenarios ((a) and (b) in Figure 2) we prove that when provided with data from k training environments, where k is linear in the number of features, and the environments satisfy some mild non-degeneracy conditions, any predictor that is calibrated on all training environments will not rely on any of the spurious features $X_{\text{ac-sp}}$, and will also be calibrated on all $e \in \mathcal{E}$.

In scenario (a), we take Y to be a binary variable drawn from a Bernoulli distribution with parameter $\eta \in [0, 1]$, and observed features are generated conditionally on Y. The features $\mathbf{x}_{\text{ac-ns}} \in \mathbb{R}^{d_{\text{ns}}}$ are invariant, meaning their conditional distribution given Y is the same for all environments, whereas $\mathbf{x}_{\text{ac-sp}} \in \mathbb{R}^{d_{\text{sp}}}$ are spurious features, as their distribution may shift between environments, altering their correlation with Y. The data generating process for training environment $i \in [k]$ in Fig. 2(a) is given by:

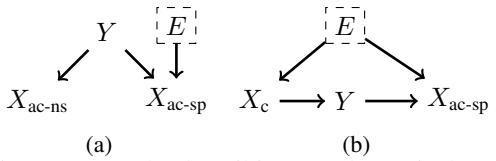


Figure 2: Graphs describing our theoretical analysis. Features are: (a) anti-causal, some spurious while others invariant. (b) causal and may undergo covariate shift, or anti-causal and spurious.

$$y = \begin{cases} 1 & \text{w.p } \eta \\ 0 & \text{o.w} \end{cases} \qquad \begin{aligned} X_{\text{ac-ns}} \mid Y = y &\sim \mathcal{N}\left((y - 1/2)\mu_{\text{ns}}, \Sigma_{\text{ns}}\right), \\ X_{\text{ac-sp}} \mid Y = y &\sim \mathcal{N}\left((y - 1/2)\mu_i, \Sigma_i\right). \end{aligned} \tag{2}$$

For $\mathbf{x} = [\mathbf{x}_{\text{ac-ns}}, \mathbf{x}_{\text{ac-sp}}]$ we consider a linear classifier $f(\mathbf{x}; \mathbf{w}, b) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$, where $\sigma : \mathbb{R} \to [0, 1]$ is some invertible function (e.g. a sigmoid). Since the mean of spurious features, μ_i, is determined by y, these features can help predict the label in some environments. Yet, these correlations do not carry to all environments, and $f(\mathbf{x})$ might rely on spurious correlations whenever the coefficients in $\mathbf{w}$ corresponding to $\mathbf{x}_{\text{ac-sp}}$ are non-zero. Any such classifier can suffer an arbitrarily high loss in an unseen environment, because a new environment can reverse and magnify the correlations observed in E_{train}. Using these definitions, we may now state our result for this case:

Theorem 1. *Given $k > 2d_{sp}$ training environments where data is generated according to Equation (2) with parameters $\{\mu_i, \Sigma_i\}_{i=1}^k$, we say they lie in general position if for all non-zero $\mathbf{x} \in \mathbb{R}^{d_{sp}}$:*

$$\dim\left(\operatorname{span}\left\{\begin{bmatrix} \Sigma_i \mathbf{x} + \mu_i \\ 1 \end{bmatrix}\right\}_{i \in [k]}\right) = d_{sp} + 1.$$

If a linear classifier is calibrated on k training environments which lie in general position, then its coefficients for the features $\mathbf{x}_{ac\text{-}sp}$ are zero. Moreover, the set of training environments that do not lie in general position has measure zero in the set of all possible training environments $\mathcal{E}^k$.

As a corollary, we see that calibration on training environments generalizes to calibration on $\mathcal{E}$. The proof of this theorem is given in the supplementary material, Section A.3. The data generating process closely resembles the one considered by [37], who use diagonal covariance matrices.

In the second scenario we consider the addition of causal features subject to covariate shift $\mathbf{x}_{\text{c}} \in \mathbb{R}^{d_{\text{c}}}$, as shown in Figure 2b. The covariate shift is induced when the environments E alter the distribution of the causal features $\mathbf{x}_{\text{c}}$ [40]. In this case, we analyze a regression problem since it is amenable to exact analysis. The data generating process for training environment $i \in [k]$ is:

$$X_c \sim \mathcal{N}(\mu_i^c, \Sigma_i^c);\ Y = \mathbf{w}_c^{*\top}\mathbf{x}_c + \xi,\ \xi \sim \mathcal{N}(0, \sigma_y^2)$$
$$X_{\text{ac-sp}} = y\mu_i + \eta,\ \eta \sim \mathcal{N}(\mathbf{0}, \Sigma_i). \tag{3}$$

For $\mathbf{x} = [\mathbf{x}_c, \mathbf{x}_{\text{ac-sp}}]$ it turns out that in this case, calibration on multiple domains forces $f(\mathbf{x})$ to discard $\mathbf{x}_{\text{ac-sp}}$, but also forces it to use $\mathbf{w}_c^*$, since it characterizes $P(Y \mid \mathbf{x}_c)$ which is the invariant mechanism in this scenario. The exact statement and proof are in Section A.4 of the supplement.

Theorem 2 (informal)**.** *Let $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}$ be a linear regressor and assume we have $k > \max\{d_c + 2, d_{sp}\}$ training environments where data is generated according to Equation (3). Under mild non-degeneracy conditions, if the regressor is calibrated across all training environments then the coefficients corresponding to X_c equal $\mathbf{w}_c^*$ and those that correspond to $X_{ac\text{-}sp}$ are zero.*

Together, these results show calibration can generalize across environments, given that the number of environments is approximately that of the spurious features. They also show that for the settings above, the relatively stable and well-known notion of calibration implies avoiding spurious correlations.

4 Related Work

As discussed in Section 2, multi-domain calibration is an instance of an invariant representation [1]. Many extensions to the above work have been proposed, e.g. [24, 3]. Yet, recent work claims that many of these approaches still fail to find invariant relations in cases of interest [20, 37, 13], where a significant challenge seems to be the gap between what is achieved by the regularization term used in practice and the goal of conditional independence $Y \perp\!\!\!\perp E \mid \Phi(X)$. Gulrajani et al. [11] give a sobering view on methods for OOD generalization, emphasizing the power of ERM and data augmentation, and the challenge of model selection. We claim that compared to the above approaches, multi-domain calibration studied here is a simpler form of invariance. Furthermore, calibration is attractive because there are standard tools to quantify it such as calibration scores [31] and a vast literature on its properties and how it can be obtained [50, 47, 30, 26, 45, 14, 36].

Learning models which generalize OOD is a fruitful area of research with many recent developments. Most work focuses on the case of Domain Adaptation where unlabeled samples are available from the target domain, including recent work on OOD calibration [48]. However, important work has also been done on the area of our focus – the so-called "proactive" case [43], where no OOD samples are available whatsoever [28, 17, 38, 34, 39].

Calibration also plays an important role in uncertainty estimation for deep networks [12], and recently in fairness, where calibration on subgroups of populations is sought [35]. This has interesting resemblance to the multiple environments calibration we consider here. A more general notion of multi-calibration has also been studied in this context [16], with recent results on sample complexity [42] which may provide tools to finite sample analysis of domain generalization. Finally, multiple methods for training calibrated models [26, 29, 36] have also been proposed. In Section 5 we propose a generalization of [26] to the multi-domain case to achieve multi-domain calibration.

5 Proactively Achieving Multi-Domain Calibration

So far we have seen a general argument why calibration can limit spurious correlations, and that in linear-Gaussian models multi-domain calibration guarantees OOD generalization. Now we turn to a more applied perspective and show how can we optimize models so they achieve this type of calibration in practice. We propose three approaches: (1) using calibration measures for model selection, (2) post-processing calibration, and (3) a calibration objective building on a method proposed by [26]. Section A.1 in the supplementary provides a slightly broader introduction to notions we use here. We will assess model calibration by the Expected Calibration Error (*ECE*) of the calibration curve [7], which is the average deviation between model accuracy and model confidence.

5.1 Model selection with average ECE

Model selection is challenging when aimed at OOD generalization. As recently observed by [11], since OOD accuracy is often at odds with In-Domain (ID) accuracy, selection based on ID validation error eliminates the advantage of domain generalization methods over vanilla ERM with data augmentation. We suggest that model selection towards OOD generalization should balance ID validation error with another observable surrogate for the stability of a model to distribution shifts between domains. Motivated by multi-domain calibration, we propose using the average ECE across training environments as this surrogate. Concretely, we propose choosing a model with lowest average ECE from those obtaining ID validation accuracy that is above a certain user-defined threshold.

5.2 Post-Processing Calibration

Practitioners interested in (single-domain) calibrated models often apply post-processing calibration methods to binary classifiers, where the most widely used approach is Isotonic Regression Scaling [50, 30]. Unlike standard calibration problems, in our case there are multiple domains to calibrate over. We give two ways of extending Isotonic Regression to the multi-domain setting, which we term "naive calibration" and "robust calibration". **Naive Calibration** takes predictions of a trained model f on validation data pooled from all domains and fits an isotonic regression z^*. We then report the performance of $z^* \circ f$ on the OOD test set.

Robust Calibration: In a multiple domain setting, Naive calibration may produce a model that is well calibrated on the pooled data, but uncalibrated on individual environments. Since our goal is simultaneous calibration, the following alternative attempts to bound the worst-case miscalibration across training environments. For each environment $e \in E_{\text{train}}$, we denote the number of validation examples we have from it by N_e, and by $f_{e,i}$ the prediction of the model on the i-th example. Then in a similar vein to robust optimization, we fit an isotonic regressor that solves: $z^* = \arg\min_z \max_{e \in E_{\text{train}}} \frac{1}{N_e} \sum_{i=1}^{N_e} \left(z(f_{e,i}) - y_i\right)^2$. Since Isotonic Regression can be formulated as a quadratic program, and Equation (5.2) minimizes a pointwise maximum over such objectives, we can cast Eq. 5.2 as a convex program and solve with standard optimizers. We then evaluate the OOD performance of $z^* \circ f$.

5.3 Learning with Multi-Domain Calibration Error

The above model selection and post-processing methods are easy to apply and (as we will soon see) surprisingly effective. However, both are limited in their power to learn a model that is truly well-calibrated across multiple domains. We now propose a more powerful approach: an objective function that directly penalizes calibration errors on multiple domains during training. Specifically, we propose learning a parameterized classifier $f_\theta(\mathbf{x})$ using a learning rule of the form: $\min_\theta \sum_{e \in E_{\text{train}}} l^e(f_\theta) + \lambda \cdot r(f_\theta)$, where $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is an empirical loss function (e.g. cross-entropy) and $l^e(f_\theta)$ denotes the expected loss over data from training environment e, and $r(f_\theta)$ is a regularization term over multiple environments. Using this notation the method proposed by [1] learns a classifier $f = w \circ \Phi$ with a regularizer given by $r(f) = \sum_{e \in E_{\text{train}}} r^e_{\text{IRMv1}}(f)$, where $r^e_{\text{IRMv1}}(f) = \|\nabla_{w|w=1} l^e(w \cdot \Phi)\|^2$.

Our proposed regularizer $r(f_\theta)$ is based on the work of Kumar et al. [26], who introduce a method they call Maximum Mean Calibration Error (MMCE). MMCE harnesses the power of universal kernels to express the ECE as an Integral Probability Measure, and works as follows: For a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^m$, denote the confidence of a classifier on the i-th example by $f_{\theta;i} = \max\{f_\theta(x_i), 1 - f_\theta(x_i)\}$ and its correctness by $c_i = \mathbb{1}_{|y_i - f_{\theta;i}| < \frac{1}{2}}$. For a given universal kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, MMCE over the dataset D is given by: $r^D_{\text{MMCE}}(f_\theta) = \frac{1}{m^2} \sum_{i,j \in D} (c_i - f_{\theta;i})(c_j - f_{\theta;j}) k(f_{\theta;i}, f_{\theta;j})$.

Calibration Loss Over Environments (CLOvE). Given multiple training domains with a dataset D^e for each $e \in E_{\text{train}}$, we arrive at our proposed regularizer by aggregating MMCE over them: $r_{\text{CLOvE}}(f_\theta) = \sum_{e \in E_{\text{train}}} r^{D_e}_{\text{MMCE}}(f_\theta)$. A key property of CLOvE is that its minima correspond to perfectly calibrated classifiers over all training domains, a consequence of the correspondence between MMCE and perfect calibration.

Corollary 1 (of Thm. 1 in [26])**.** *CLOvE is a proper scoring rule. That is, it equals* 0 *if and only if* $f_\theta(\mathbf{x})$ *is perfectly calibrated for every* $e \in E_{\text{train}}$.

Additional properties of CLOvE, such as large deviation bounds and relation to ECE, can also be derived; see results in [26] for further details. In the following section, we will see how these properties translate into favorable OOD generalization in practice when training with CLOvE.

6 Experiments and Results

6.1 Colored MNIST and Two-Bit Environments

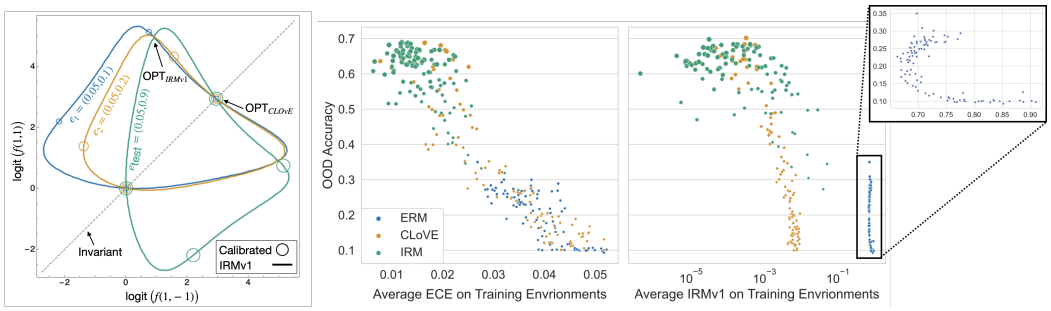
In order to explore the challenges of OOD generalization and how they relate to learning from multiple environments, [1] used the colored MNIST dataset [21]. In this dataset certain digits tend to be colored either red or green in the train set, but the correlation between colors and digits is flipped in the OOD test set, making color a spurious feature. This dataset was then further simplified into "Two-Bit" environments by [20], who proved that the IRMv1 penalty proposed in [1] does *not* in fact achieve the correct invariant solution on the simplified setting. The Two-Bit environments problem setting has two binary features, $X_1, X_2 \in \{-1, 1\}$, corresponding respectively to digit identity ($0-4$ or $5-9$) and digit color in the original colored MNIST. The environments are parameterized by $e = (\alpha, \beta) \in [0, 1]^2$ controlling the correlation of the features with the label:
$Y \leftarrow \mathrm{Rad}(0.5),\ X_1 \leftarrow Y \cdot \mathrm{Rad}(\alpha), X_2 \leftarrow Y \cdot \mathrm{Rad}(\beta)$, where $\mathrm{Rad}(\delta)$ is a random variable equal to -1 with probability δ and 1 with probability $1-\delta$. At training we are given data from two environments $e_1 = (\alpha, \beta_1), e_2 = (\alpha, \beta_2)$, $\beta_1 \neq \beta_2$. The learned model is tested on a new environment $e_3 = (\alpha, \beta_3)$ with β_3 significantly different from β_1, β_2. Only a model discarding the spurious feature X_2 will maintain its accuracy moving from train to OOD test.
Calibration discards spurious correlation in Two-Bit environments. Figure 3(a), which we adapt from Figure 6 in Appendix B of [20], illustrates the merits of CLOvE in this setting. The figure shows the space of odd classifiers, i.e. those for which $f(1,-1) = -f(-1,1)$, and $f(1,1) = -f(-1,-1)$.[4] The true invariant classifiers are those for which in addition $f(1,1) = f(1,-1)$, corresponding to models lying on the diagonal of Figure 3(a), denoted by the dashed gray line. In the figure, we plot in solid lines the classifiers for which $r^e_{\text{IRMv1}}(f)$ equals 0, and in solid circles the classifiers for which $r^e_{\text{MMCE}}(f)$ equals 0 (due to Corollary 1 these coincide with calibrated classifiers on environment e). Note that in this parameterization, the zeros of $r^e_{\text{IRMv1}}(f)$ are lines whereas the zeros of $r^e_{\text{MMCE}}(f)$ are isolated points. Intersections of the zeros of $r^e_{\text{IRMv1}}(f)$ denote solutions for which the corresponding regularization terms are 0 on all respective environments, while intersection of zeros of $r^e_{\text{MMCE}}(f)$ are the zeros of $r_{\text{CLOvE}}(f)$. As observed by [20], when $E_{\text{train}} = \{e_1, e_2\}$ the solution denoted by $\text{OPT}_{\text{IRMv1}}$ has the lowest empirical loss, yet this solution has a spurious correlation with X_2 and thus will incur a higher loss on the test environment e_3. This means the corresponding IRMv1 learning rule cannot retrieve the optimal invariant classifier. On the other hand, learning with CLOvE does retrieve the optimal invariant classifier in this case, in addition to the trivial, constant classifier. This means CLOvE discards spurious correlations in cases where IRMv1 does not. In Section C we present experiments reproducing the above scenario on the Colored MNIST dataset.

Model selection based on average ECE We train models with varying hyperparameters on Colored MNIST using ERM, CLOvE and IRM, (100 models with each algorithm, see Section C of the supplement for details). We then calculate the ECE and IRMv1 penalties of each model over a held-out validation set from each training environment, and evaluate the average of these against OOD accuracy. Figure 3(b) presents the results across all trained models. The ID ECE penalty displays a very strong correlation across the entire range and every training regime (Pearson corr. = -0.92), while ID IRMv1 behaves more erratically (Pearson corr. = -0.59). Since quantities used for model selection should be agnostic to choices made at training time, we suggest that ID ECE is a better choice for use in model selection. Further results on model selection can be found in the supplement, Section C.

6.2 WILDS Benchmarks

WILDS is a recently proposed benchmark of in-the-wild distribution shifts from several data modalities and applications[5]. Table 1 presents the four WILDS datasets we experiment with, chosen to represent diverse OOD generalization scenarios. We follow the models and training algorithms

[4] As explained in [20], the optimal solutions are odd so we may focus on them for visualization purposes.
[5] `https://wilds.stanford.edu`



(a) (b)

Figure 3: (a) Zeros of MMCE and IRMv1 are indicated by circles and by solid lines respectively, in a color corresponding to each environment. The dashed diagonal is the space of invariant solutions. Some zeros intersect across environments, and these are therefore the domain-invariant solutions. Among the domain-invariant solutions, $\text{OPT}_{\text{IRMv1}}$ has the lowest empirical loss when training on e_1, e_2. Hence learning with IRMv1 will prefer this model over $\text{OPT}_{\text{CLOvE}}$, which discards the spurious correlation with X_2. (b) Correspondence between observable criteria and OOD accuracy in CMNIST. Each point corresponds to a model trained with some training algorithm (marked by color) and hyperparameter setting. Size of marker is proportional to the ratio between OOD and ID accuracies.

proposed by [23]. In order to perform multi-domain calibration we modify the splits to include a multi-domain validation set whenever possible. See supplemental Section B for details and for additional results on Amazon Reviews. As in [23], we use three different training algorithms to train our models: **ERM**, **IRM**, **DeepCORAL**, and further use **GroupDRO** for one of the datasets, compatible with WILDS version 1.0.0. We apply three calibration approaches described in 5.2 and 5.3 above to each trained model: **naive calibration** and **robust calibration**, which are post-processing methods and therefore applied on the models' outputs; and **CLOvE**, which we apply as a fine-tuning approach to the top layers of each trained model. We train each (algorithm $\times$ calibration) combination four times with different random seeds, and report average results and their standard deviations.

Dataset	Type	Label (y)	Input (x)	Domain (e)	Model ($f(x)$)
PovertyMap	Regression	Asset Wealth Index	Satellite Image	Country	ResNet
Camelyon17	Binary	Tumor Tissue	Histopathological Image	Hospital	DenseNet
CivilComments	Binary	Comment Toxicity	Online Comment	Demographics	BERT
FMoW	Multi-class	Land Use Type	Satellite Image	Region	DenseNet

Table 1: Description of each of the datasets used in our WILDS experiments.

Table 2 presents our main results on the *FMoW* (left) and *Camelyon17* (right) datasets. On both datasets, robust calibration already improves performance, and CLOvE then significantly outperforms robust calibration, improving performance by 7% and 2.8% (absolute) over the strongest alternative on *FMoW* and *Camelyon17*, respectively. When compared to the original model, the performance of CLOvE is even more striking, with CLOvE outperforming it by more than 10% (absolute) on *FMoW* and 6% on *Camelyon17*. Another appealing property of CLOvE is the low variance exhibited across different runs. Indeed, CLOvE has lower variance than both naive and robust calibration approaches, and has lower variance than the original (uncalibrated) model on 4 of the 6 experiments.

	FMoW				*Camelyon17*			
Algorithm	Orig.	Naive Cal.	Rob. Cal.	CLOvE	Orig.	Naive Cal.	Rob. Cal.	CLOvE
ERM	32.63	33.09	37.19	**44.16**	66.66	71.23	71.22	**75.75**
	(0.016)	(0.021)	(0.035)	(0.018)	(0.144)	(0.089)	(0.086)	(0.049)
DeepCORAL	31.73	31.75	33.86	**40.05**	72.44	75.97	76.8	**79.96**
	(0.01)	(0.01)	(0.016)	(0.009)	(0.044)	(0.054)	(0.065)	(0.039)
IRM	31.33	31.81	34.41	**42.24**	70.87	73.25	73.4	**73.95**
	(0.012)	(0.016)	(0.015)	(0.014)	(0.068)	(0.066)	(0.069)	(0.061)

Table 2: Left: worst unseen region accuracy on OOD test set in *FMoW*. Right: Accuracy on unseen hospital test set in *Camelyon17*. Orig.: original algorithm, no changes applied. Best OOD result for each domain in **bold**. Standard deviation across runs in brackets, lowest OOD std. is underlined.

Analysis. As can be seen in Figure 4, improvements in ID calibration are associated with better OOD performance. Interestingly, when our post-processing does not improve OOD performance, it

is often linked to our inability to substantially improve ID calibration. This is most visible in IRM experiments, where robust calibration is unable to outperform naive calibration both in terms ID calibration and in OOD performance. Finally, we find it interesting that merely post-processing the data (as in robust calibration) can already have such a marked effect on OOD accuracy, though still inferior to actually optimizing for multi-domain calibration as done by CLOvE.

Results on alternative settings. While our theoretical analysis is focused on OOD generalization of classification models, we also experiment with alternative settings from *WILDS* to test the power of ID calibration in improving OOD performance. Specifically, we experiment with the *PovertyMap* dataset, which introduces a regression task, and the *CivilComments* dataset, which introduces a subpopulation shift scenario for a binary classifier. As can be seen in Table 3, results on the *CivilComments* dataset (right), show that calibration consistently improves worst-case performance, with an average improvement of 21.5% across training algorithms. While CLOvE does outperform naive and robust calibration on average, the gain is lower in comparison to *FMoW* and *Camelyon17*.

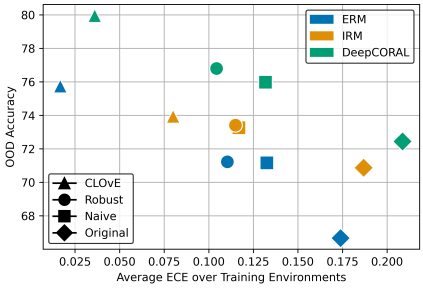


Figure 4: OOD accuracy as a function of average ECE over training domains, for all models on the *Camelyon17* dataset.

In *PovertyMap* (left), the model solves a regression task, so we cannot use CLOvE to improve OOD performance. Still, robust calibration improves performance across all experiments, though by a smaller margin. In the case of models pre-trained by IRM, robust calibration improves OOD performance substantially, outperforming the original model by 0.08% (absolute). Interestingly, calibration also leads to more stable results both in *PovertyMap* and in *CivilComments*, as can be seen in the standard deviation across different model runs.

	PovertyMap				*CivilComments*			
Algorithm	Orig.	Naive Cal.	Rob. Cal.	Algorithm	Orig.	Naive Cal.	Rob. Cal.	CLOvE
ERM	0.832	0.827	**0.834**	ERM	63.65	76.98	78.99	**80.39**
	(0.011)	(0.014)	(0.006)		(0.026)	(0.005)	(0.008)	(0.007)
IRM	0.735	0.812	**0.815**	IRM	40.61	**68.97**	68.92	68.45
	(0.117)	(0.016)	(0.015)		(0.16)	(0.013)	(0.013)	(0.02)
DeepCORAL	0.832	0.835	**0.837**	GroupDRO	71.67	76.2	78.54	**80.07**
	(0.011)	(0.009)	(0.012)		(0.007)	(0.013)	(0.008)	(0.003)

Table 3: Left: Pearson correlation r on in-domain (ID) and OOD (unseen countries) test sets in *PovertyMap*. Right: worst-case group accuracy on the test set in the *CivilComments* dataset.

7 Conclusion

In this paper we highlight a novel connection between multi-domain calibration and OOD generalization, arguing that such calibration can be viewed as an invariant representation. We proved in a linear setting that models calibrated on multiple domains are free of spurious correlations and therefore generalize out of domain. We then proposed multi-domain calibration as a practical and measurable surrogate for the OOD performance of a classifier. We demonstrated that actively tuning models to achieve multi-domain calibration significantly improves model performance on unseen test domains, and that in-domain calibration on a validation set is a useful criterion for model selection. A major limitation of our work is that our theoretical findings are limited to linear models in a population (as opposed to finite-sample) setting; we thus consider them more as a motivation rather than a full justification of using multi-domain calibration in practice as we do. We look forward to expanding the scope of theoretical understanding of the conditions under which multi-domain calibration can provably guarantee out-of-domain generalization, including the finite-sample setting and the analysis of specific algorithms. We also expect new practical methods, building on our findings, will help push forward the real-world ability to generalize to unseen test domains.

Acknowledgments

We wish to thank Ira Shavitt for his helpful comments and to Alexandre Ramé for pointing us to an error in the original manuscript. This research was partially supported by the Israel Science Foundation (grant No. 1950/19).

References

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

[2] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[3] A. Bellot and M. van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

[4] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[5] G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[6] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[7] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[8] S. Desai and G. Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[10] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[11] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[13] R. Guo, P. Zhang, H. Liu, and E. Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

[14] C. Gupta, A. Podkopaev, and A. Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[16] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[17] C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[18] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In *AISTATS*, 2021.

[21] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[24] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[25] V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.

[26] A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814, 2018.

[27] J. M. Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.

[28] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10869–10879, 2018.

[29] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

[30] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[31] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41. Computer Vision Foundation / IEEE, 2019.

[32] J. Pearl. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*, pages 454–462. Elsevier, 1994.

[33] J. Pearl. *Causality*. Cambridge university press, 2009.

[34] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[35] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[36] A. Rahimi, A. Shaban, C.-A. Cheng, R. Hartley, and B. Boots. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[37] E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[38] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[39] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[40] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[41] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. Improved protein structure prediction using potentials from deep learning. *Nat.*, 577(7792):706–710, 2020.

[42] E. Shabat, L. Cohen, and Y. Mansour. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

[43] A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

[44] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[45] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[47] V. Vovk, G. Shafer, and I. Nouretdinov. Self-calibrating probability forecasting. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 1133–1140, 2003.

[48] X. Wang, M. Long, J. Wang, and M. I. Jordan. Transferable calibration with lower bias and variance in domain adaptation. *arXiv preprint arXiv:2007.08259*, 2020.

[49] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

[50] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 609–616. Morgan Kaufmann, 2001.

[51] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A Proofs for Theoretical Claims

We begin by supplementing the definition of multiple domain calibration, extending it for the case of regression, then we provide proofs of the theorems in the paper.

A.1 Definition of Calibration

Recall our definition of a calibrated classifier for binary tasks.

Definition S1. *Let $f : \mathcal{X} \to [0, 1]$ and $P[X, Y]$ be a joint distribution over the features and label. Then $f(\mathbf{x})$ is calibrated w.r.t to P if for all $\alpha \in [0, 1]$ in the range of f:*

$$\mathbb{E}_P[Y \mid f(X) = \alpha] = \alpha.$$

In the multiple environments setting, $f(\mathbf{x})$ is calibrated on E_{train} if for all $e_i \in E_{train}$ and α in the range of f restricted to e_i:

$$\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \alpha.$$

Let us prove the connection between multi-domain calibration and invariance, we repeat the statement of the lemma from the main paper for convenience.

Lemma S1 (Lemma 1 in main paper)**.** *If a binary classifier f is invariant w.r.t E_{train} then there exists some $g : \mathbb{R} \to [0, 1]$ such that $g \circ f$ is calibrated on all training environments and its mean squared error on each environment does not exceed that of f. On the other hand, if a classifier is calibrated on all training environments it is also invariant w.r.t E_{train}.*

Proof. Assume that the classifier is invariant w.r.t E_{train}, let $e_i \in E_{\text{train}}$ and note that:

$$\mathbb{E}[(Y - f(X))^2 \mid E = e_i] \geq \min_{g:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(Y - g \circ f(X))^2 \mid E = e_i].$$

The solution to the RHS is to take $g(\hat{\alpha}) = \mathbb{E}[Y \mid f(X) = \hat{\alpha}, E = e_i]$ for all $\hat{\alpha} \in [0, 1]$ and it results in a classifier $g \circ f$ that is calibrated w.r.t e_i. Due to invariance, for all $\hat{\alpha} \in \mathbb{R}$ the expectation $\mathbb{E}[Y \mid f(X) = \hat{\alpha}]$ is identical across all $e_i \in E_{\text{train}}$ where $\hat{\alpha}$ is in the range of f restricted to e_i. Therefore there exists a single function g that solves the RHS simultaneously over all environments. The resulting $g \circ f$ is indeed calibrated over all training domains and its mean squared error does not exceed that of f (note that since the square loss is Bayes-consistent, this claim also holds for the classification error). The other part of the statement that a calibrated classifier on all E_{train} is invariant follows easily from the definitions. □

For regression tasks, one may consider a function that outputs a full CDF on Y and define a calibrated classifier as one where all quantiles of the CDF match the true quantiles of Y as the number of examples approached infinity. This leads to the definition in [25], and one may follow this to analyze more general cases than the scenario we will consider in this work.

Since in this section we consider Gaussian distributions and linear regressors, a definition based on the first two moments of the distribution (instead of all quantiles of a CDF) will suffice. Hence we will be working the following definition:

Definition S2. *Let $f : \mathcal{X} \to \mathbb{R}^2$ and $P[X, Y]$ a joint distribution over the features and label. Then $f(\mathbf{x})$ is calibrated w.r.t to P if for all $(\alpha, \beta) \in \mathbb{R}^2$ in the range of f:*

$$\mathbb{E}[Y \mid f(X)_1 = \alpha] = \alpha,\ \mathbb{E}\big[Y^2 \mid f(X)_2 = \beta\big] = \beta.$$

In the multiple environments setting, $f(\mathbf{x})$ is calibrated on E_{train} if for all $e_i \in E_{train}$ and (α, β) in the range of f restricted to e_i:

$$\mathbb{E}[Y \mid f(X) = (\alpha, \beta), E = e_i] = \alpha,\ \mathbb{E}\big[Y^2 \mid f(X) = (\alpha, \beta), E = e_i\big] = \beta. \tag{4}$$

A.2 Details about ECE, MMCE and Post-Processing Methods

To evaluate calibration and optimize our models towards multi-domain calibration, we use the Expected Calibration Error (ECE) and the Maximum Mean Calibration Error (MMCE) [26].

The ECE is a scalar summary of the calibration plot, used throughout the literature to assess how well calibrated is a given classifier. **Calibration plots** [7] are a visual representation of model calibration

in the case of binary labels. Each example $\mathbf{x}$ is placed into one of B bins that partition the $[0, 1]$ interval, in which the output, or *confidence*, of the classifier $f(\mathbf{x})$ falls. For each bin b, the accuracy of f on the bin's examples $acc(b)$ is calculated along with the average confidence $conf(b)$. These are plotted against each other to form a curve, where deviations from a diagonal represent miscalibration. **ECE score** summarizes the calibration curve by averaging the deviation between accuracy and confidence:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|. \tag{5}$$

n_b is the number of examples in bin b, N is the total number of examples. In all of our experiments we used $B = 10$ bins of equal size.

To handle the miscalibration that is often observed in models such as neural networks [12], the MMCE was proposed in [26] as a method to improve calibration at training time. Recalling the definition of this loss: We consider a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^m$, a binary classifier parameterized by a vector θ which we denote $f_\theta :\to [0, 1]$. The confidence of f_θ on the i-th example is $f_{\theta;i} = \max\{f_\theta(x_i), 1 - f_\theta(x_i)\}$ and its correctness is $c_i = \mathbb{1}_{|y_i - f_{\theta;i}| < \frac{1}{2}}$. Then we fix a kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, associated with a feature map $\phi : [0, 1] \to \mathcal{H}$, and MMCE over the dataset D is given by:

$$r^D_{\text{MMCE}}(f_\theta) = \frac{1}{m^2} \sum_{i,j \in D} (c_i - f_{\theta;i})(c_j - f_{\theta;j}) k(f_{\theta;i}, f_{\theta;j}). \tag{6}$$

In our experiments we use an RBF kernel $k(r, r') = \exp(-\gamma(r - r')^2)$ with $\gamma = 2.5$. Equation (6) is the finite sample approximation of the following:

$$MMCE(f_\theta; P[X, Y]) = \|\mathbb{E}_{(\mathbf{x},y)\sim P}[(c - f_\theta(\mathbf{x}))\phi(f_\theta(\mathbf{x}))]\|_{\mathcal{H}}. \tag{7}$$

Here c is the correctness of f_θ on $(\mathbf{x}, y)$ as defined for Equation (6). Attractive properties of the MMCE include it being a proper scoring rule:

Theorem (Adapted from Thm. 1 in [26])**.** *Let $P[X, Y]$ be a probability measure defined on the space $(\mathcal{X} \times \{0, 1\})$ such that the conditionals on the pushforward measure $P[r, c] = f_\theta \sharp P$,*[6] *$P(r \mid c = 1)$ over $([0, 1] \times \{0, 1\})$, $P(r \mid c = 0)$ are Borel probability measures, and let k be a universal kernel. The MMCE in Equation (7) is 0 if and only if f_θ is calibrated w.r.t P.*

Corollary 1 in the paper follows by considering $\sum_{e \in E_{\text{train}}} MMCE(f_\theta; P[X, Y \mid E = e])$ and applying the theorem to each summand. For more details on the MMCE, its derivation as an integral probability measure analogue of the ECE and its properties, we refer the reader to [26].

Another popular metric for calibration in binary classification problems is the Brier score, which is simply the squared error between the predicted probability and the outcome [5]:

$$BS(f) = \frac{1}{m} \sum_{i=1}^{m} (f(\mathbf{x}_i) - y_i)^2.$$

The Isotonic Regression [30] post-processing methods that we use in the paper minimize the Brier score using a monotonic post-processing function. Hence we consider a classifier f and a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^m$. Denote the prediction of f on $\mathbf{x}_i$ by f_i, then isotonic regression solves:

$$\min_{z: f_i \le f_j \Rightarrow z(f_i) \le z(f_j)} \frac{1}{m} \sum_{i=1}^{m} (z(f_i) - y_i)^2.$$

A motivation for using this as a post-processing calibration method is the decomposition of the Brier score to a refinement and calibration score. We may denote the set of prediction values that are obtained by f across the dataset by $F = \{f_i \mid i \in [m]\}$. For each such value $\tilde{f} \in F$ then denote $N_{\tilde{f}} = |\{i \mid f_i = \tilde{f}\}|$ as the number of points for which we obtain this prediction and $y_{\tilde{f}} = \frac{1}{N_{\tilde{f}}} \sum_{i: f_i = \tilde{f}} y_i$ the average outcome over them:

$$BS(f) = CAL(f) + REF(f) = \frac{1}{m} \sum_{\tilde{f} \in F} N_{\tilde{f}} (\tilde{f} - y_{\tilde{f}})^2 + \frac{1}{m} \sum_{\tilde{f} \in F} N_{\tilde{f}} (y_{\tilde{f}}(1 - y_{\tilde{f}}))$$

[6] we note the abuse of notation here, as $f_\theta \sharp P$ is used to denote the measure that we get by applying f_θ to X to obtain r and c is obtained by calculating its correctness w.r.t to Y.

The calibration score measures how far is the average prediction value from the average outcome, while refinement gives a measure of their sharpness (i.e. it raises the score of uncertain prediction). Due to the monotonicity constraint of isotonic Regression, it is usually thought of as not changing the $REF(f)$ too much, which means it minimizes the Brier score mainly by reducing $CAL(f)$. In the multi-domain cases we are interested in, note that this vanilla isotonic regression does not take domains into account. In our experiments we use it simple by pooling the dataset on all environments and performing post-processing calibration on this dataset using isotonic regression. This procedure could output a classifier that is perfectly calibrated for the entire dataset, but not on single environments.

To give a simple variant that does post-processing while taking environments into account, we proposed a Robust Isotonic Regression method. The method minimizes the Brier score on the worst-case environment, thus aiming to bound the worst miscalibration on each environment. While in practice it will usually not provide perfect calibration on each environment, the method trades off the error between environments so it is better geared towards simultaneous calibration of the classifier on all domains. Formally we solve:

$$z^* = \underset{z: f_i \leq f_j \Rightarrow z(f_i) \leq z(f_j)}{\arg\min} \max_{e \in E_{\text{train}}} \frac{1}{N_e} \sum_{i=1}^{N_e} \left(z(f_{e,i}) - y_i\right)^2. \tag{8}$$

Where N_e are the number of data points in environment $e \in E_{\text{train}}$ and $f_{e,i}$ is the output of f on point i in the environment.

Equipped with these definitions, we now turn to the proofs of the theorems in the paper.

A.3 Classification with Invariant Features

We first consider the classification task from the main paper, where the data generating process is described in Figure S1. Recall that we are considering linear classifiers of the form $f(\mathbf{x}; \mathbf{w}, b) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$. Our environments here are defined by the parameters of the multivariate Gaussian distributions that generate the spurious features $\{\mu_i, \Sigma_i\}_{i=1}^k$. As a first step we will derive the algebraic form of the constraints that calibration imposes on $\mathbf{w}$ and the parameters defining the environments. For convenience, we modify the notation from the main paper and consider a binary label where $\mathcal{Y} = \{-1, 1\}$ instead of $\mathcal{Y} = \{0, 1\}$.

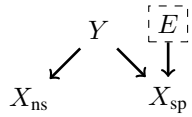


Figure S1: Diagram for data generating process in the invariant features scenario.

Lemma S2. *Assume we have k environments with means and covariance matrices for environmental features $\mu_i \in \mathbb{R}^{d_e}, \Sigma_i \in \mathbb{S}_{++}^{d_e}, i \in [k]$ and a common covariance matrix $\Sigma_{ns} \in \mathbb{S}_{++}^{d_{ns}}$ for invariant features, where data is generated according to:*

$$y = \begin{cases} 1 & \textit{w.p } \eta \\ -1 & \textit{otherwise} \end{cases}, \qquad \begin{aligned} \mathbf{x}_{ns} \mid Y = y &\sim \mathcal{N}(y\mu_{ns}, \Sigma_{ns}), \\ \mathbf{x}_{sp} \mid Y = y &\sim \mathcal{N}(y\mu_i, \Sigma_i), \end{aligned}$$

and $\mathbf{x}_{ns}, \mathbf{x}_{sp}$ are drawn independently. Let $\sigma : \mathbb{R} \to (0, 1)$ be an invertible function and define the classifier:

$$f(\mathbf{x}; \mathbf{w}, b) = \sigma(\mathbf{w}^\top \mathbf{x} - b).$$

Decompose the weights $\mathbf{w} = [\mathbf{w}_{ns}, \mathbf{w}_{sp}]$ to the coefficients of the invariant and spurious features accordingly. Then if the classifier is calibrated on all environments, it holds that either $\mathbf{w} = \mathbf{0}$ or there exists $t \neq 0$ such that:

$$\frac{\mathbf{w}_{ns}^\top \mu_{ns} + \mathbf{w}_{sp}^\top \mu_i}{\mathbf{w}_{ns}^\top \Sigma_{ns} \mathbf{w}_{ns} + \mathbf{w}_{sp}^\top \Sigma_i \mathbf{w}_{sp}} = t \quad \forall i \in [k]. \tag{9}$$

Proof. Let $i \in [k]$, the joint distribution of features in the environment is Gaussian with mean $\hat{\mu}_i = [\mu_{\text{ns}}, \mu_i]$, covariance $\hat{\Sigma}_i = \begin{bmatrix} \Sigma_{\text{ns}} & 0 \\ 0 & \Sigma_i \end{bmatrix}$. Hence the output of the affine function corresponding to the classifier is a random variable with probability density function:

$$P[\sigma^{-1}(f(X)) = \alpha \mid Y = y, E = e_i] = (2\pi \mathbf{w}^\top \hat{\Sigma}_i \mathbf{w})^{-\frac{1}{2}} \exp\left(\frac{\left(\alpha - y\mathbf{w}^\top \hat{\mu}_i + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} \right).$$

Hence the conditional probability of Y is given by:

$$P[Y = 1 \mid \sigma^{-1}(f(X)) = \alpha, E = e_i] = \frac{\eta \exp\left(\frac{\left(\alpha - \mathbf{w}^\top \hat{\mu}_i + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} \right)}{\eta \exp\left(\frac{(\alpha - \mathbf{w}^\top \hat{\mu}_i + b)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} \right) + (1-\eta) \exp\left(\frac{(\alpha + \mathbf{w}^\top \hat{\mu}_i + b)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} \right)}.$$

Note that unless $\mathbf{w} = \mathbf{0}$ (which results in a calibrated classifier that satisfies Equation (9)), the variance of $\sigma^{-1}(f(X))$ is strictly positive since $\hat{\Sigma}_i \succ 0$, so above conditional probabilities are well-defined. Now it is easy to see that if the classifier is calibrated across environments, we need to have equality in the log-odds ratio for each i, j and all $\alpha \in \mathbb{R}$:

$$\frac{\left(\alpha - \mathbf{w}^\top \hat{\mu}_i + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} - \frac{\left(\alpha + \mathbf{w}^\top \hat{\mu}_i + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} = \frac{\left(\alpha - \mathbf{w}^\top \hat{\mu}_j + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_j \mathbf{w}} - \frac{\left(\alpha + \mathbf{w}^\top \hat{\mu}_j + b\right)^2}{2\mathbf{w}^\top \hat{\Sigma}_j \mathbf{w}} \quad \forall \alpha \in \mathbb{R}.$$

After dropping all the terms that cancel out in the subtractions we arrive at:

$$\frac{\mathbf{w}^\top \hat{\mu}_i}{\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} = \frac{\mathbf{w}^\top \hat{\mu}_j}{\mathbf{w}^\top \hat{\Sigma}_j \mathbf{w}}.$$

This may also be written as a system of equations with an additional scalar variable $t \in \mathbb{R}$:

$$\frac{\mathbf{w}^\top \hat{\mu}_i}{\mathbf{w}^\top \hat{\Sigma}_i \mathbf{w}} = t \quad \forall i \in [k].$$

Now because we assumed $\Sigma_i \succ 0$ for all environments, for any solution to the above system with $t = 0$, we must have:

$$\mathbf{w}^\top \hat{\mu}_i = 0 \quad \forall i \in [k].$$

Furthermore we will have for any $\alpha \in \mathbb{R}$:

$$P[Y = 1 \mid \sigma^{-1}(f(X)) = \alpha, E = e_i] = \eta.$$

Since we assume f is calibrated and the right hand side needs to equal α, this is only possible if $f(\mathbf{x}; \mathbf{w}, b)$ is a constant function. Again, because $\Sigma_i \succ 0$, this is only possible if $\mathbf{w} = \mathbf{0}$. Hence we conclude with our desired result, as can be seen by decomposing $\mathbf{w}$ to the parts corresponding to invariant and spurious features. □

We now give a result for the special case where the covariance matrices of the spurious features satisfy $\Sigma_i = \sigma_i^2 \mathbf{I}$, considered in [37]. The nice correspondence here is that we will see that calibration demands one more environment than IRM to discard all spurious features. This matches the intuition that each environment reduces a degree of freedom from the set of invariant classifiers, while risk minimization reduces one more degree of freedom.

Lemma S3. *Assume we have $k \geq d_{sp} + 2$ environments and define $M\left(\{\mu_i, \sigma_i\}_{i=1}^k\right) \in \mathbb{R}^{k \times d_e + 2}$:*

$$M(\{\mu_i, \sigma_i\}_{i=1}^k) = \begin{bmatrix} \mu_1^\top & \sigma_1^2 & 1 \\ & \vdots & \\ \mu_k^\top & \sigma_k^2 & 1 \end{bmatrix}.$$

If the matrix has full rank, then for any invariant predictor the linear coefficients on spurious features are zero.

Proof. According to Lemma S2, writing down the conditional probability $P[Y \mid \sigma^{-1}(f(\mathbf{x})), E = e]$ and demanding calibration results in the constraint that either $\mathbf{w} = \mathbf{0}$, and then the linear coefficients on spurious features are indeed 0; or that for some $t \neq 0$:

$$\frac{\mathbf{w}_{\text{ns}}^\top \mu_{\text{ns}} + \mathbf{w}_{\text{sp}}^\top \mu_i}{\mathbf{w}_{\text{ns}}^\top \Sigma_{\text{ns}} \mathbf{w}_{\text{ns}} + \sigma_i^2 \|\mathbf{w}_{\text{sp}}\|_2^2} = t \quad \forall i \in [k].$$

Without loss of generality we can phrase these constraints as:

$$\frac{\mathbf{w}_{\text{ns}}^\top \mu_{\text{ns}} + \mathbf{w}_{\text{sp}}^\top \mu_i}{\mathbf{w}_{\text{ns}}^\top \Sigma_{\text{ns}} \mathbf{w}_{\text{ns}} + \sigma_i^2 \|\mathbf{w}_{\text{sp}}\|_2^2} = 1 \quad \forall i \in [k].$$

This is true since if $\mathbf{w}$ is a solution to this system of equations where the right hand side is some $t \in \mathbb{R}$ then $t\mathbf{w}$ is a solution to the system where t is replaced by 1. Rewrite the constraints again to isolate the parts depending on $\mathbf{w}_{\text{sp}}$:

$$\sigma_i^2 \|\mathbf{w}_{\text{sp}}\|_2^2 - \mu_i^\top \mathbf{w}_{\text{sp}} = \mathbf{w}_{\text{ns}}^\top \Sigma_{\text{ns}} \mathbf{w}_{\text{ns}} - \mathbf{w}_{\text{ns}}^\top \mu_{\text{ns}} \quad \forall i \in [k].$$

To find whether this system has a solution where $\mathbf{w}_{\text{sp}}$ is non-zero we can replace the right hand side with a scalar variable $t \in \mathbb{R}$, and ask whether the following system has a non-zero solution:

$$\sigma_i^2 \|\mathbf{w}_{\text{sp}}\|_2^2 - \mu_i^\top \mathbf{w}_{\text{sp}} = t \quad \forall i \in [k].$$

For the above equations to have a non-zero solution, the following linear system must also have such a solution:

$$M(\{\mu_i, \sigma_i\}_{i=1}^k)\mathbf{x} = \mathbf{0}.$$

But from our non-degeneracy condition, such a solution does not exist. □

Next we generalize the above to prove the result from the main paper, namely when the matrices $\{\Sigma_i\}_{i=1}^k$ are not diagonal. For this purpose we introduce a definition of general position for environments, similar to the one given in [1].

Definition S3. *Given $k > 2d_{sp}$ environments with mean parameters $\{\Sigma_i, \mu_i\}_{i=1}^k$, we say they are in general position if for all non-zero $\mathbf{x} \in \mathbb{R}_{sp}^d$:*

$$\dim\left(\text{span}\left\{\begin{bmatrix}\Sigma_i \mathbf{x} + \mu_i \\ 1\end{bmatrix}\right\}_{i \in [k]}\right) = d_e + 1.$$

Equipped with this notion of general position, we now need to show that if it holds then the only predictors that satisfy the conditions of Lemma S2 are those with $\mathbf{w}_{\text{sp}} = \mathbf{0}$. Another claim we will need to prove is that the subset of environments which do not lie in general position have measure zero in the set of all possible environment settings. Hence generic environments are expected to lie in general position. This argument will follow the lines of the one given in [1], adapted to our case with the fixed coordinate 1 added in the above definition.

Theorem 1. *Under the setting of Lemma S2, if the environments lie in general position then all classifiers that are calibrated across environments satisfy $\mathbf{w}_{sp} = \mathbf{0}$.*

Proof. According to Lemma S2, if the predictor is calibrated then Equation (9) must hold. Following the same arguments laid out in the proof at the main paper, we get that $\mathbf{w}_{sp}$ needs to be a solution for the following system of equations:

$$\mathbf{w}_{sp}^\top \Sigma_i \mathbf{w}_{sp} - \mu_i^\top \mathbf{w}_{sp} - t = 0 \quad \forall i \in [k]. \tag{10}$$

Now, let $\mathbf{w}_{sp} \in \mathbb{R}^{d_{\text{sp}}}$ be a non-zero vector and let us define the $k \times d_e + 1$ matrix:

$$M(\{\mu_i, \Sigma_i\}_{i=1}^k, \mathbf{w}_{sp}) = \begin{bmatrix} \mathbf{w}_{sp}^\top \Sigma_1 - \mu_1^\top & 1 \\ \vdots & \\ \mathbf{w}_{sp}^\top \Sigma_k - \mu_k^\top & 1 \end{bmatrix}$$

If the environments are in general position, the above matrix has full rank for any non-zero $\mathbf{w}_{sp}$. Similarly to the proof of Lemma S3, if Equation (10) has a non-zero solution then the following system must also have a solution:

$$M(\{\mu_i, \Sigma_i\}_{i=1}^k, \mathbf{w}_{sp})\mathbf{x} = \mathbf{0}.$$

Which is of course impossible due to $M(\{\mu_i, \Sigma_i\}_{i=1}^k, \mathbf{w}_{sp})$ having full rank. □

We conclude with the statement about the measure of sets of environments which do not lie in general position, this will follow the lines of [1].

Lemma S4. *Let $k > 2d_{sp}$ and $\{\mu_i\}_{i=1}^k$ be arbitrary fixed vectors, then the set of matrices $\{\Sigma_i\}_{i=1}^k \in (\mathbb{S}_{++}^{d_{sp}})^k$ for which $\{\Sigma_i, \mu_i\}_{i=1}^k$ do not lie in general position has measure zero within the set $(\mathbb{S}_{++}^{d_{sp}})^k$.*

Proof. We assume $k > 2d_{\text{sp}}$ and denote by $LR(k, d_{\text{sp}}, r)$ the matrices of dimensions $k \times d_{\text{sp}}$ and rank r. Also for any d denote by $\mathbf{1}_d$ the vector in $\mathbb{R}^d$ where all entries equal 1. Define $\mathbf{M}_*^1(k, d_{\text{sp}})$ as the set of $k \times d_{\text{sp}}$ matrices of full column-rank whose columns span the vector of ones $\mathbf{1}_k$:

$$\mathbf{M}_*^1(k, d_{\text{sp}}) = \{A \in LR(k, d_{\text{sp}}, d_{\text{sp}}) \mid \mathbf{1}_k \in \text{colsp}(A)\}.$$

Let $\{\Sigma_i\}_{i=1}^k \in (\mathbb{S}_{++}^{d_{\text{sp}}})^k$ and define $\mathbf{W} \subseteq \mathbb{R}^{k \times d_{sp}}$ as the image of the mapping $G : \mathbb{R}^{d_{sp}} \setminus \{0\} \to \mathbb{R}^{k \times d_{sp}}$:

$$(G(\mathbf{x}))_{i,l} = (\Sigma_i \mathbf{x} - \mu_i)_l$$

By the definition of general position given in the paper, the environments defined by $\{\Sigma_i, \mu_i\}_{i=1}^k$ lie in general position if $\mathbf{W}$ does not intersect $LR(k, d_{\text{sp}}, r)$ for all $r < d_{\text{sp}}$ and $\mathbf{M}_*^1(k, d_{\text{sp}})$. We would like to show that this happens for all but a measure zero of $\left(\mathbb{S}_{++}^{d_{sp}}\right)^k$.

Due to the exact same arguments in Thoerem 10 of [1], we have that $\mathbf{W}$ is transversal to any submanifold of $\mathbb{R}^{k \times d_{\text{sp}}}$ and also does not intersect $LR(k, d_{\text{sp}}, r)$ where $r < d_{\text{sp}}$, for all $\{\Sigma_i\}_{i=1}^k$ but a measure zero of $\left(\mathbb{S}_{++}^{d_{sp}}\right)^k$.

It is left to show that it also does not intersect $\mathbf{M}_*^1(k, d_{\text{sp}})$ for all but a measure zero of $\left(\mathbb{S}_{++}^{d_{sp}}\right)^k$. Because $\mathbf{M}_*^1(k, d_{\text{sp}})$ is a submanifold of $\mathbb{R}^{k \times d_{\text{sp}}}$, it intersects transversally with $\mathbf{W}$ for generic $\{\Sigma_i\}_{i=1}^k$. Then by transversality they cannot intersect if $\dim(\mathbf{W}) + \dim(\mathbf{M}_*^1(k, d_{\text{sp}})) - \dim(\mathbb{R}^{k \times d_{\text{sp}}}) < 0$. We will claim that $\dim(\mathbf{M}_*^1(k, d_{\text{sp}})) = k(d_{\text{sp}} - 1) + d_{\text{sp}}$ and then since $k > 2d_{\text{sp}}$ we may obtain:

$$\begin{aligned}
\dim(\mathbf{W}) + \dim(\mathbf{M}_*^1(k, d_{\text{sp}})) - \dim(\mathbb{R}^{k \times d_{\text{sp}}}) &\leq d_{\text{sp}} + k(d_{\text{sp}} - 1) + d_{\text{sp}} - kd_{\text{sp}} \\
&= 2d_{\text{sp}} - k \\
&< 0.
\end{aligned}$$

The negativity of the dimension implies that if $\mathbf{W}$ and $\mathbf{M}_*^1(k, d_{\text{sp}})$ are transversal then they do not intersect, and we may conclude our desired result that the environments lie in general position for all but a measure zero of $\left(\mathbb{S}_{++}^{d_{\text{sp}}}\right)^k$.

To show that $\dim(\mathbf{M}_*^1(k, d_{\text{sp}})) = k(d_{\text{sp}} - 1) + d_{\text{sp}}$, consider a matrix $A \in \mathbf{M}_*^1(k, d_{\text{sp}})$. Since it has full rank, it has a $d_{\text{sp}} \times d_{\text{sp}}$ minor that is invertible. Assume this minor is just the first d_{sp} rows of A, otherwise there is a linear isomorphism that transforms it into such a matrix and the arguments that follow still apply (see [27], Example 5.30; our proof follows a similar line of reasoning). Now write A as a block matrix using $B \in \mathbb{R}^{d_{\text{sp}} \times d_{\text{sp}}}, C \in \mathbb{R}^{(k - d_{\text{sp}}) \times d_{\text{sp}}}$:

$$A = \begin{bmatrix} B \\ C \end{bmatrix}.$$

Denoting by $\mathbf{U}$ the set of $k \times d_{\text{sp}}$ matrices whose first d_{sp} rows are invertible, we consider the mapping $F : \mathbf{U} \to \mathbb{R}^{k - d_{\text{sp}}}$:

$$F(A) = \mathbf{1}_{k - d_{\text{sp}}} - CB^{-1}\mathbf{1}_{d_{\text{sp}}}.$$

Clearly $F^{-1}(\mathbf{0}) = \mathbf{M}_*^1(k, d_{\text{sp}})$ and F is smooth. We will show that it is a submersion by observing that its differential $DF(U)$ is surjective for each $U \in \mathbf{U}$. To this end, for a given $U = \begin{bmatrix} B \\ C \end{bmatrix}$ and any $X \in \mathbb{R}^{(k - d_{\text{sp}}) \times d_{\text{sp}}}$ define a curve $\gamma : (-\epsilon, \epsilon) \to \mathbf{U}$ by:

$$\gamma(t) = \begin{bmatrix} B \\ C + \gamma X \end{bmatrix}.$$

We have that:

$$(F \circ \gamma)'(t) = \frac{d}{dt}|_{t=0}(\mathbf{1}_{k-d_{\text{sp}}} - (C + tX)B^{-1}\mathbf{1}_{d_{\text{sp}}}) = XB^{-1}\mathbf{1}_{d_{\text{sp}}}.$$

Since $B^{-1}\mathbf{1}_{d_{\text{sp}}}$ is not the zero vector, and $X \in \mathbb{R}^{(k-d_{\text{sp}})\times d_{\text{sp}}}$ where $k - d_{\text{sp}} > d_{\text{sp}}$, then it is clear that the above mapping is surjective. Note that the derivatives along the curve are just a subset of the range of $DF(U)$, hence $DF(U)$ is also surjective at each point $U \in \mathbf{U}$. It follows from the submersion theorem that $\dim(\mathbf{M}^1_*(k, d_{\text{sp}})) = kd_{\text{sp}} - (k - d_{\text{sp}}) = k(d_{\text{sp}} - 1) + d_{\text{sp}}$ as desired for our result to hold. □

A.4 Regression Under Covariate Shift and Spurious Features

We now move on to the second scenario presented in the paper where the mechanism $P(Y \mid X)$ is invariant and the diagram depicting the data generating process is given in Figure S2. Here for each environment $i \in [k]$ we will have:

$$\begin{aligned} X_c &\sim \mathcal{N}(\mu_i^c, \Sigma_i^c) \\ Y &= \mathbf{w}_c^{*\top}\mathbf{x}_c + \xi,\ \xi \sim \mathcal{N}(0, \sigma_y^2) \\ X_{sp} &= y\mu_i + \eta,\ \eta \sim \mathcal{N}(\mathbf{0}, \Sigma_i). \end{aligned} \tag{11}$$

We consider a regressor $f : \mathcal{X} \to \mathbb{R}^2$, where the estimate of the mean is linear, i.e. $[f(\mathbf{x}; \mathbf{w})]_1 = \mathbf{w}^\top\mathbf{x}$, and the estimate of the variance is constant $[f(\mathbf{x}; \mathbf{w})]_2 = c$.[7] We decompose the weights $\mathbf{w}$ into their parts corresponding to causal and spurious features $[\mathbf{w}_c, \mathbf{w}_{sp}]$. Then our result regarding calibration and generalization to $\mathcal{E}$ is given below.

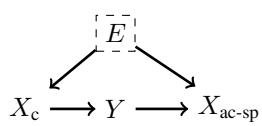


Figure S2: Diagram for data generating process in the covariate shift scenario.

Theorem 2. *Denote the dimensions of X_c, X_{sp} by d_c, d_{sp} accordingly. Assume we have k environments with parameters $\{\mu_i^c, \mu_i, \Sigma_i^c, \Sigma_i\}_{i=1}^k$. For any matrix A denote its i-th row by A^i, and define the matrices $M(\{\mu_i^c, \mu_i\}_{i=1}^k) \in \mathbb{R}^{k\times d_c+d_{sp}+1}$ and $M_2(\{\mu_i^c, \Sigma_i^c\}_{i=1}^k, \sigma_y^2, \mathbf{w}_c^*) \in \mathbb{R}^{k\times d_c+2}$ whose rows are given by:*

$$M(\{\mu_i^c, \mu_i\}_{i=1}^k) = \begin{bmatrix} \mu_i^{c\top} & \left(\mathbf{w}_c^{*\top}\mu_1^c\right)\mu_1^\top & 1 \\ & \vdots & \\ \mu_k^{c\top} & \left(\mathbf{w}_c^{*\top}\mu_k^c\right)\mu_k^\top & 1 \end{bmatrix},$$

$$M_2(\{\mu_i^c, \Sigma_i^c\}_{i=1}^k, \sigma_y^2, \mathbf{w}_c^*) = \begin{bmatrix} \mathbf{w}_c^{*\top}\Sigma_1^c + \left(\frac{\mathbf{w}_c^{*\top}\Sigma_1^c\mathbf{w}_c^* + \sigma_y^2}{\mathbf{w}_c^{*\top}\mu_1^c}\right)\mu_1^{c\top} & \frac{\mathbf{w}_c^{*\top}\Sigma_1^c\mathbf{w}_c^*}{\mathbf{w}_c^{*\top}\mu_1^c} & 1 \\ & \vdots & \\ \mathbf{w}_c^{*\top}\Sigma_k^c + \left(\frac{\mathbf{w}_c^{*\top}\Sigma_k^c\mathbf{w}_c^* + \sigma_y^2}{\mathbf{w}_c^{*\top}\mu_k^c}\right)\mu_k^{c\top} & \frac{\mathbf{w}_c^{*\top}\Sigma_k^c\mathbf{w}_c^*}{\mathbf{w}_c^{*\top}\mu_k^c} & 1 \end{bmatrix}.$$

Let $f(\mathbf{x}; \mathbf{w})$ be a calibrated regressor, assume $\mathbf{w}_c^{*\top}\mu_i^c \neq 0$ for all $i \in [k]$ and that there exists $i, j \in [k]$ such that $\mathbb{E}[Y \mid E = e_i] \neq \mathbb{E}[Y \mid E = e_j]$. Furthermore assume that one of the following conditions hold:

- *$k > \max\{d_c + 2, d_{sp}\}$, $M_2(\{\mu_i^c, \Sigma_i^c\}_{i=1}^k, \sigma_y^2, \mathbf{w}_c^*)$ has full rank and the means of spurious features $\{\mu_i\}_{i=1}^k$ span $\mathbb{R}^{d_{sp}}$.*

[7] Limiting the variance estimate to a constant does not make a difference for the purpose of our proof. The proof does not rely on the correctness of the variance estimate as imposed by Equation (4), but only on the variances being equal across environments when conditioned on $f(\mathbf{x})$. In other words it relies on the correctness of the mean estimate, and the distribution of Y conditioned on $f(X)$ being the same across environments.

- $k > d_c + d_{sp} + 1$ *and* $M(\{\mu_i^c, \mu_i\}_{i=1}^k)$ *has full rank.*

then the weights of f *must be* $\mathbf{w} = [\mathbf{w}_c^*, \mathbf{0}]$.

It is rather clear that rank-deficiency of M_2 would impose some highly non-trivial conditions on the relationships between $\mu_i^c, \mathbf{w}_c^{*\top}\Sigma_i^c$ and the conditions given above are satisfied for all settings of environments other than a measure zero under any absolutely continuous measure on the parameters $\mathbf{w}_c^*, \{\mu_i^c, \Sigma_i^c\}_{i=1}^k$. The proof proceeds by writing the conditional distribution of Y on $f(X)$, and showing that the conditions in the theorem are the direct result of the calibration constraints.

Proof. Since X_c, X_{sp}, Y are jointly Gaussian, we can write their distribution at environment $i \in [k]$ as:

$$\begin{bmatrix} X_c \\ X_{sp} \\ Y \end{bmatrix} \sim \mathcal{N}\Bigg(\begin{bmatrix} \mu_i^c \\ (\mathbf{w}_c^{*\top}\mu_i^c)\mu_i \\ \mathbf{w}_c^{*\top}\mu_i^c \end{bmatrix}, \begin{bmatrix} \Sigma_i^c & \Sigma_i^c\mathbf{w}_c^*\mu_i^\top & \Sigma_i^c\mathbf{w}_c^* \\ \mu_i\mathbf{w}_c^{*\top}\Sigma_i^c & \left(\mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2\right)\mu_i\mu_i^\top + \Sigma_i & \left(\mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2\right)\mu_i \\ \mathbf{w}_c^{*\top}\Sigma_i^c & (\mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2)\mu_i^\top & \mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2 \end{bmatrix} \Bigg).$$

The predictions $\mathbf{w}^\top X$ are then also normally distributed, and jointly with Y this can be written as:

$$\begin{bmatrix} \mathbf{w}^\top X \\ Y \end{bmatrix} \sim \mathcal{N}\Bigg(\begin{bmatrix} \mathbf{w}_c^\top\mu_i^c + (\mathbf{w}_{sp}^\top\mu_i)(\mathbf{w}_c^{*\top}\mu_i^c) \\ \mathbf{w}_c^{*\top}\mu_i^c \end{bmatrix}, \begin{bmatrix} \sigma_{f,i}^2 & \sigma_{f,y,i} \\ \sigma_{f,y,i} & \sigma_{y,i}^2 \end{bmatrix} \Bigg),$$

where we defined the items of the covariance matrix:

$$\begin{aligned} \sigma_{f,i}^2 &= \mathbf{w}_c^\top\Sigma_i^c\mathbf{w}_c + 2(\mathbf{w}_c^\top\Sigma_i^c\mathbf{w}_c^*)(\mu_i^\top\mathbf{w}_{sp}) + \mathbf{w}_{sp}^\top\left(\mu_i\mu_i^\top(\mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2) + \Sigma_i\right)\mathbf{w}_{sp}, \\ \sigma_{f,y,i} &= \mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c + (\mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2)\mu_i^\top\mathbf{w}_{sp}, \\ \sigma_{y,i}^2 &= \mathbf{w}_c^{*\top}\Sigma_i^c\mathbf{w}_c^* + \sigma_y^2. \end{aligned}$$

Now we can write the mean of the conditional distribution of Y on $f(X)_1 = \alpha$ as:

$$\mathbb{E}\left[Y \mid f(X)_1 = \alpha, E = e_i\right] = \mathbf{w}_c^{*\top}\mu_i^c + \frac{\sigma_{f,y,i}}{\sigma_{f,i}^2}(\alpha - \mathbf{w}_c^\top\mu_i^c - (\mathbf{w}_{sp}^\top\mu_i)(\mathbf{w}_c^{*\top}\mu_i^c)).$$

For each environment $i \in [k]$, the above is a linear function of α. Demanding $f(X)$ to be calibrated on all environments then imposes both the slopes and intercepts to be equal across environments. Writing this for the slope, we obtain that there must exist $t \in \mathbb{R}$ such that:

$$\frac{\sigma_{f,y,i}}{\sigma_{f,i}^2} = t \quad \forall i \in [k]. \tag{12}$$

We note that $t \neq 0$ since if it is zero then we have that $\mathbb{E}[Y \mid f(X)_1 = \alpha, E = i]$ does not depend on α, where calibration demands that it equals α. This can only happen if $\mathbf{w}_c = \mathbf{0}$, otherwise the range of $f(\mathbf{x})$ is $\mathbb{R}$ because we assumed in the definition of the environments that $\Sigma_c^i \succ 0$. Furthermore, $\mathbf{w}_c = \mathbf{0}$ cannot be calibrated if $\mathbb{E}[Y \mid E = e_i]$ is not constant across environments; which is also part of the non-degeneracy constraints we required. Next we demand the equality of the intercepts across environments. Taking these equations and replacing Equation (12) into each of them, we get:

$$\mathbf{w}_c^{*\top}\mu_i^c - t\left(\mathbf{w}_c^\top\mu_i^c + (\mathbf{w}_{sp}^\top\mu_i)(\mathbf{w}_c^{*\top}\mu_i^c)\right) = \mathbf{w}_c^{*\top}\mu_j^c - t\left(\mathbf{w}_c^\top\mu_j^c + (\mathbf{w}_{sp}^\top\mu_j)(\mathbf{w}_c^{*\top}\mu_j^c)\right) \ \forall i, j \in [k].$$

Dividing both sides by t and defining $\bar{\mathbf{w}}_c = \frac{\mathbf{w}_c^*}{t} - \mathbf{w}_c$, we can introduce another variable $t_2 \in \mathbb{R}$ and write this as a linear system of equations in variables $\mathbf{w}_{sp}, \bar{\mathbf{w}}_c, t_2$:

$$\bar{\mathbf{w}}_c^\top\mu_i^c - \mathbf{w}_{sp}^\top\mu_i(\mathbf{w}_c^{*\top}\mu_i^c) + t_2 = 0 \quad \forall i \in [k]. \tag{13}$$

We see that given $d_c + d_{sp} + 1$ environments, then with mild conditions on their non-degeneracy (i.e. the vectors containing the environment means and an extra entry of 1 span $\mathbb{R}^{d_c+d_{sp}+1}$), the only solution to the system is $\bar{\mathbf{w}}_c = 0, \mathbf{w}_{sp} = 0$, proving the last part of our statement.

Moving forward to demand multiple calibration on second moments $\mathbb{E}[Y^2 \mid f(X)_1 = \alpha, E = e_i] = \mathbb{E}[Y^2 \mid f(X)_1 = \alpha, E = e_j]$ for all $i, j \in [k]$, we may write this as:

$$\sigma_{y,i}^2 - \frac{\sigma_{f,y,i}^2}{\sigma_{f,i}^2} = \sigma_{y,j}^2 - \frac{\sigma_{f,y,j}^2}{\sigma_{f,j}^2} \quad \forall i, j \in [k].$$

Plugging Equation (12) into the above, a simplified expression is obtained:

$$\sigma_{y,i}^2 - t\sigma_{f,y,i} = \sigma_{y,j}^2 - t\sigma_{f,y,j} \quad \forall i, j \in [k].$$

Again we can divide by t and obtain an explicit expression using $\bar{\mathbf{w}}_c, \mathbf{w}_{sp}$:

$$\bar{\mathbf{w}}_c^\top \Sigma_i^c \mathbf{w}_c^* - (\mathbf{w}_c^{*\top} \Sigma_i^c \mathbf{w}_c^* + \sigma_y)\mathbf{w}_{sp}^\top \mu_i = \bar{\mathbf{w}}_c^\top \Sigma_j^c \mathbf{w}_c^* - (\mathbf{w}_c^{*\top} \Sigma_j^c \mathbf{w}_c^* + \sigma_y)\mathbf{w}_{sp}^\top \mu_j \quad \forall i, j \in [k].$$

Finally, we can plug in Equation (13) and introduce another variable $t_3 \in \mathbb{R}$ to turn the above equations into:

$$\bar{\mathbf{w}}_c^\top \left(\Sigma_i^c \mathbf{w}_c^* + \left(\frac{\mathbf{w}_c^{*\top} \Sigma_i^c \mathbf{w}_c^* + \sigma_y^2}{\mathbf{w}_c^{*\top} \mu_i^c} \right) \mu_i^c \right) + t_2 \left(\frac{\mathbf{w}_c^{*\top} \Sigma_i \mathbf{w}_c^*}{\mathbf{w}_c^{*\top} \mu_i^c} \right) + t_3 = 0.$$

It is now easy to see that if $k > d_c + 2$ and $\mathbf{M}_2(\{\mu_i, \Sigma_i\}_{i=1}^k, \sigma_y^2, \mathbf{w}_c^*)$ has full rank, the only solution to these equations satisfies $\bar{\mathbf{w}}_c = \mathbf{0}, t_2 = t_3 = 0$. When this is plugged into Equation (13), we find that if $k > d_{sp}$ and the spurious means span $\mathbb{R}^{d_{sp}}$ then the only possible solution is $\mathbf{w}_{sp} = \mathbf{0}$. Finally, $\bar{\mathbf{w}}_c = \mathbf{0}$ means $\mathbf{w}_c^* = t\mathbf{w}_c$, and if $f(\mathbf{x})$ is calibrated then we must have $t = 1$ since otherwise its estimate of the conditional mean is incorrect. Hence our proof is concluded. □

We note that even though the setting we considered is restricted to causal features, anti-causal non-spurious features as those in Figure S1 can also be treated (resulting in the graph given in Figure 1). This is since for a single environment, the distribution $P[X_{\text{c}}, X_{\text{ac-ns}}, X_{\text{ac-spurious}}, Y \mid E = e]$ (we shorten here to P^e for convenience) can always be written as follows, treating $X_{\text{ac-ns}}$ as causal features:

$$\begin{aligned} P^e[X_{\text{c}}, X_{\text{ac-ns}}, X_{\text{ac-sp}}, Y] &= P^e(X_{\text{c}}, X_{\text{ac-ns}}) P^e(Y \mid X_{\text{c}}, X_{\text{ac-ns}}) P^e(X_{\text{ac-sp}} \mid Y, X_{\text{ac-ns}}, X_{\text{c}}) \\ &= P^e(X_{\text{c}}, X_{\text{ac-ns}}) P^e(Y \mid X_{\text{c}}, X_{\text{ac-ns}}) P^e(X_{\text{ac-sp}} \mid X_{\text{ac-ns}}). \end{aligned}$$

The last equality is due to the separation properties of the graph, and since the joint distribution is a multivariate Gaussian, so are all the factors in the above product. Hence each environment can be described using a structural equation model of the same type as Equation (11) and Theorem 2 applies.

A.5 Causal Graphical Models

In order to answer queries about unseen distribution based on data from different, observed distributions, one must make certain assumptions about the data generating processes and the relationships between the seen and unseen distributions. One way of articulating such models of the world is by using causal graphs. In a causal graph, edges from a variable X to a variable Y mean that changing the value of X *may* change the distribution of Y. Causal graphs entail all statistical dependencies between variables, and we can read off such independence statements using the d-separation criterion [32]. We refer to background material to discuss how to identify and estimate causal effects with these causal graphical models in hand [33].

In Section 2 we illustrate our assumed causal graph for a general problem of distribution shift, and in Section 3 we illustrate assumed causal graph for the causal and anti-causal simplified examples described in equations 2 and 3, respectively.

B Dataset Statistics and Models

For each of the four WILDS experiments presented in Section 6, we briefly describe the data and report the splits we use for training, validation and test. In each experiment we train a model on the training set, and the calibrators on the validation set. The post-processing calibrators receive tuples of model predictions and labels as input, whereas fine tuning with CLOvE receives a latent

representation (values of the last hidden layer for *Camelyon17* and *FMoW*, and average of the representation of the cls token over the last 4 hidden layers in *CivilComments*). CLOvE is trained over a Multilayer Perceptron with 3 hidden layers, with batch size of 64 and the Adam optimizer. We then compare all alternatives (Original, Naive Calibration, Robust Calibration and CLOvE) on the held-out test set (OOD). Whenever an In-Domain (ID) test set is available (*PovertyMap* and *Camelyon17*), we evaluate the model on it as well. Throughout our experiments, we measure and report the Expected Calibration Error (ECE) using 10 bins, dividing the $[0, 1]$ interval into sub-intervals of equal length. The licenses to the datasets are CC0 for *Camelyon17* and *CivilComments*, *FMoW* is distributed under the FMoW Challenge Public License and *PovertyMap* is public domain. All model training is done on an infrastructure with 4 RTX 2080 Ti GPUs.

B.1 *PovertyMap*

Problem Setting *PovertyMap* is a regression task of poverty mapping across countries. Input $\mathbf{x}$ is a multispectral satellite image, output y is a real-valued asset wealth index and domain d is a country and whether the satellite image is of an urban or a rural area. The goal is to generalize across countries and demonstrate subpopulation performance across urban and rural areas.

Data *PovertyMap* is based on a dataset collected by [49], which organized satellite images and survey data from 23 African countries between 2009 and 2016. There are 23 countries, and every location is classified as either urban or rural. Each example includes the survey year, and its urban/rural classification.

1. Training: 10000 images from 13 countries.
2. Validation (OOD): 4000 images from 5 different countries (distinct from training and test (OOD) countries).
3. Test (OOD): 4000 images from 5 different countries (distinct from training and validation (OOD) countries).
4. Validation (ID): 1000 images from the same 13 countries in the training set.
5. Test (ID): 1000 images from the same 13 countries in the training set.

B.2 *Camelyon17*

Problem Setting *Camelyon17* is a tumor identification task across different hospitals. Input $\mathbf{x}$ is an histopathological image, label y is a binary indicator of whether the central region contains any tumor tissue and domain d is an integer identifying the hospital. The training and validation sets include the same four hospitals, and the goal is to generalize to an unseen fifth hospital. We note that in [23] they include data from three hospitals in the training set and validate on data from a fourth hospital. Our setting includes a validation set from multiple hospitals since our fine tuning methods requires multiple domains.

Data The dataset comprises 450000 patches extracted from 50 whole-slide images (WSIs) of breast cancer metastases in lymph node sections, with 10 WSIs from each of five hospitals in the Netherlands [2]. Each WSI was manually annotated with tumor regions by pathologists, and the resulting segmentation masks were used to determine the labels for each patch. Data is split according to the hospital from which patches were taken.

1. Training: 335996 patches taken from each of the 4 hospitals in the training set.
2. Validation: 60000 patches taken from each of the 4 hospitals in the training set (15000 patches from each hospital).
3. Test (OOD): 85054 patches taken from the 5th hospital, which was chosen because its patches were the most visually distinctive.

B.3 *CivilComments*

Problem Setting *CivilComments* is a toxicity classification task across different demographic identities. Input $\mathbf{x}$ is a comment on an online article, label y indicates if it is toxic, and domain d is a one-hot vector with 8 dimensions corresponding to whether the comment mentions either of the 8

demographic identities *male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religions*, *Black*, and *White*. The goal is to do well across all subpopulations, as computed through the average and worst case model performance.

Data *CivilComments* comprises 450000 comments, annotated for toxicity and demographic mentions by multiple crowdworkers, where toxicity classification is modeled as a binary task [4]. Each comment was originally made on an online article. Articles are randomly partitioned into disjoint training, validation, and test splits, and then formed the corresponding datasets by taking all comments on the articles in those splits.

1. Training: 269038 comments.
2. Validation: 45180 comments.
3. Test: 133782 comments.

B.4 *FMoW*

Problem Setting *FMoW* is a building and land multi-class classification task across regions and years. Input $\mathbf{x}$ is an RGB satellite image, label y is one of 62 building or land use categories, and domain d is the time the image was taken and the geographical region it captures. The goal is to generalize across time, and improve subpopulation performance across all regions.

Data *FMoW* is based on the Functional Map of the World dataset [6], which includes over 1 million high-resolution satellite images from over 200 countries, based on the functional purpose of the buildings or land in the image, over the years 2002–2018. We use a subset of this data introduced in [23], which is split into three time range domains, 2002–2013, 2013–2016, and 2016–2018, as well as five geographical regions as subpopulations: *Africa*, *Americas*, *Oceania*, *Asia* and *Europe*.

1. Training: 76863 images from the years 2002–2013.
2. Validation (OOD): 19915 images from the years from 2013–2016.
3. Test (OOD): 22108 images from the years from 2016–2018.
4. Validation (ID): 11483 images from the years from 2002–2013.
5. Test (ID): 11327 images from the years from 2002–2013.

Models In the following we briefly describe each of the models used in the experiments reported in Section 6.

- **BERT** - BERT is a 12-layer Transformer model [46] that represents textual inputs contextually and sequentially [9]. It is widely used in NLP, and is considered the standard benchmark for any state-of-the-art system. It was previously shown to be miscalibrated across its training and test environments [8]. In our *CivilComments* experiments, we use BERT-base-uncased, a smaller variant of BERT which has a layer size of 768
- **DenseNet** - Dense Convolutional Network (DenseNet), is a feed-forward neural network where for each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers [19]. DenseNets are widely used in computer vision, especially for image classification tasks . We use a DenseNet-121 model, a DenseNet variant with 121 layers, in the *Camelyon17* and *FMoW* experiments.
- **ResNet** - Residual Network (ResNet) is a feed-forward neural network where layers are reformulated to learning residual functions with reference to the layer inputs [15]. DenseNets where shown to be successful in multiple image recognition tasks. We use the 18-layer variant, ResNet-18, in the *PovertyMap* experiment.

We run our models using the default setting used in [23]. Each model is trained four times, using a different random seed at each run. We report performance averages and their standard deviation in Section 6.

Robustness to Model Architecture Choice For each of the five WILDS datasets we report results on (*PovertyMap*, *Camelyon17*, *CivilComments* and *FMoW*) also tested the robustness of our results to different model architectures. In the following we describe the architecture we tested for each dataset, and the relative results achieved.

- **BERT** - We used a pre-trained *BERT* in the *Civilcomments* experiments. On the *Civilcomments* dataset, we compared results on the BERT-base-uncased model with the cased and large versions. While we did find the performance increases with model size, perfromance drops on OOD examples remained consistent across models, with CLOvE outperforming *Robust Calibration* and *Naive Calibration* by an average of 1.4% and 3.1% (absolute), respectively.
- **DenseNet** - In the *FMoW* experiments, we tested the relative performance of the 121 layer version to the 169 and 201 layer alternatives available via `https://pytorch.org/hub/pytorch_vision_densenet/`. Differences between the three models were not statistically significant.
- **ResNet** - In the *PovertyMap* experiments, we compare *ResNet-18* to the 34 and 50 layers alternatives available via `https://pytorch.org/hub/pytorch_vision_resnet/`. We found that *ResNet-18* performs slightly on the OOD test set, with average gain of 0.01 in pearson correlation compared with *ResNet-34*. *Robust Calibration* remained better than *Naive Calibration* and the original model across runs.

Training Algorithms In the WILDS experiments, for each dataset we train our models using three out of these four alternatives:

- **ERM** - Empirical risk minimization (ERM) is a training algorithms the looks for models that minimize the average training loss, regardless of the training environment.
- **IRM** Invariant risk minimization (IRM) [1] is a training algorithm that penalizes feature distributions that have different optimal linear classifiers for each environment.
- **DeepCORAL** DeepCORAL [44] is an algorithm that penalizes differences in the means and covariances of the feature distributions for each training environment. It was originally proposed in the context of domain adaptation, and has been subsequently adapted for domain generalization [11].
- **GroupDRO** - Group DRO [18] uses distributionally robust optimization (DRO) to explicitly minimize the loss on the worst-case environment.

We do not perform any hyperparameter search, and use the default version available in [23].

C Experiments on Colored MNIST

For the colored MNIST[8] dataset we trained Multi-Layer Perceptrons (MLPs) with ERM, IRMv1 and CLOvE, based on the code provided in [20] with the following adjustments: we add CLOvE and optimize it using SGD with batches of size 512 from each training environment, for 5001 steps at each run (50 epochs). We used either the Adagrad optimizer [10] or Adam [22] (Adam was replaced with Adagrad in one environment where it produced highly unstable training metrics). All models were trained on a single NVidia Tesla P100 GPU virtual machine, on the Google Cloud Platform. Other algorithms were trained with Gradient Descent (i.e. without batching the dataset, which is infeasible for CLOvE since it is based on kernels) and Adam for 500 steps/epochs, exactly as done in the code provided by [1, 20]. For CLOvE, hyperparamters are drawn similarly to the rest of the algorithms, except when using Adagrad where we multiply the originally drawn learning rate by 5.

C.1 Performance of CLOvE

We will refer to environments with tuples (α, β) that denote correlation with digit and color respectively, as done in Section 6.1. For each setting of training and test environments we experiment with, 100 models are trained using each algorithm: ERM, IRM and CLOvE. To illustrate the failure case pointed out in [20] and Section 6 of the paper, we train the algorithms with training environments corresponding to $e_1 = (0.1, 0.05), e_2 = (0.2, 0.05)$ and use data from test environment $e_3 = (0.9, 0.05)$. Figure S3 which we produce using code provided in [20] shows the results, where each point corresponds to a model trained with some set of drawn hyperparameters. Most models trained by CLOvE achieve log-loss that is close to that of the optimal invariant classifier (marked by

[8]The MNIST dataset is available under the terms of the Creative Commons Attribution-Share Alike 3.0 license

dashed black line), while the models trained with IRMv1 are more scattered and specifically those that achieve lower log-loss are the ones that also obtain lower training objective. The bold colored lines mark the points that minimize $\sum_{e \in E_{\text{train}}} l^e(f_\theta) + \lambda \cdot r^e(f_\theta)$ with $\lambda = 10^6$ (expect for ERM where it's the point which minimizes the empirical loss), showing that out of the models trained with IRMv1, the one which minimizes the objective has loss close to that of the solution $\text{OPT}_{\text{IRMv1}}$ from Figure 3(a) in the paper (marked by dashed red line). That is while the CLOvE model with the lowest training objective is very close to the optimal invariant classifier in its test loss (marked by black dashed line). Note that in this case color is the invariant feature while the digit is spurious.

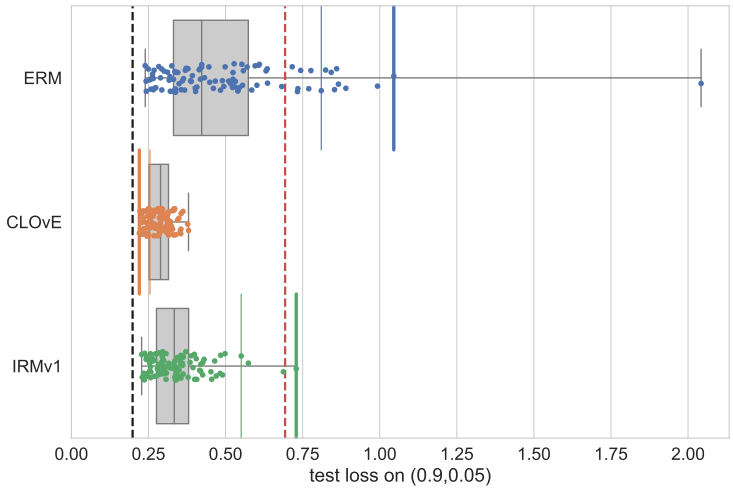


Figure S3: Log-loss on test environment $(0.9, 0.05)$ of classifiers trained with ERM, CLOvE and IRMv1 on training environments $(0.1, 0.05), (0.2, 0.05)$. Black dashed line marks the log-loss achieved by the optimal invariant classifier, while the red dashed line shows the loss achieved by $\text{OPT}_{\text{IRMv1}}$ from Figure 3(a). Bold colored lines mark the test loss achieved by the model which minimizes $\sum_{e \in E_{\text{train}}} l^e(f_\theta) + \lambda \cdot r^e(f_\theta)$ with $\lambda = 10^6$ out of all trained models.

For the opposite case, where the digit is invariant, the error incurred by MLPs in digit recognition makes it difficult to find the exact invariant classifier by optimizing CLOvE (since this error is close to the magnitude of the 0.05 correlation). Yet in Section C.2 the failure case of IRMv1 in these environments will be illustrated by average ECE (which CLOvE is a surrogate for) being a better measure of invariance than the IRMv1 objective.

The experiment presented in [1] used the training environments $e_1 = (0.25, 0.1), e_2 = (0.25, 0.2)$ with test environments $e_3 = (0.25, 0.9)$, where IRMv1 can in principle learn the optimal invariant classifier. We give the results on learning with these environments for completion. As can be observed in Figure S4, both CLOvE and IRMv1 learn models that are close to the optimal invariant one. While IRMv1 learned more of those models during the hyperparameter sweep[9], CLOvE still obtains some close-to-invariant models during the sweep. The rest of this section will be dedicated to studying model selection with the proposed average ECE criterion and the correlation between ID average ECE and OOD performance.

C.2 Model Selection Experiments

Let us recall and elaborate the selection procedure proposed in Section 5:

- Given a desired threshold for In-Domain accuracy Thr_{ID} and a set of models $f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})$ from which we would like to select a candidate, perform the following.

[9] This can be attributed to the choice of ranges for drawing hyperparameters which we did not carefully tune to accommodate CLOvE.

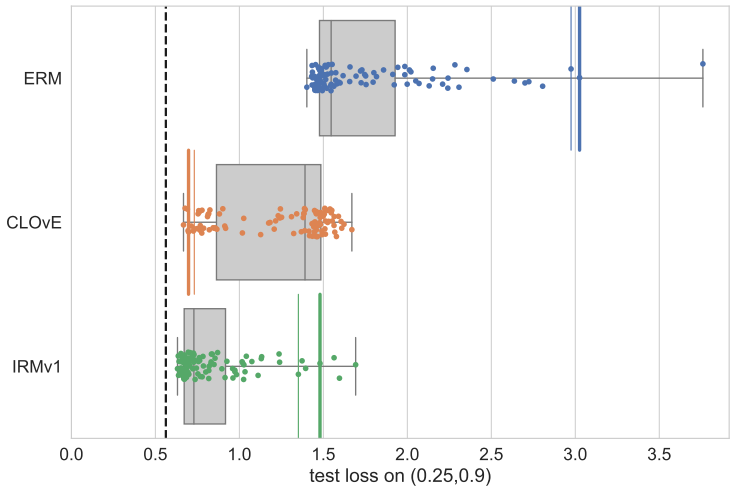


Figure S4: Log-loss on test environment $(0.25, 0.9)$ of classifiers trained with ERM, CLOvE and IRMv1 on training environments $(0.25, 0.1), (0.25, 0.2)$. Lines denote the same corresponding quantities in Figure S3, except we omit the red dashed line from that figure.

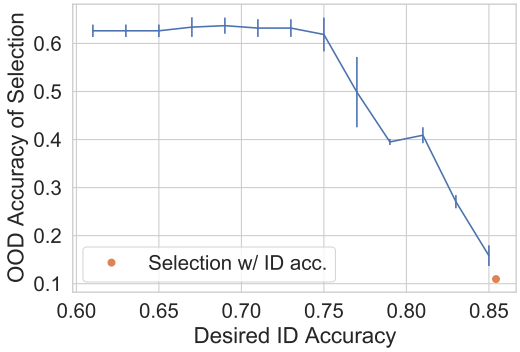
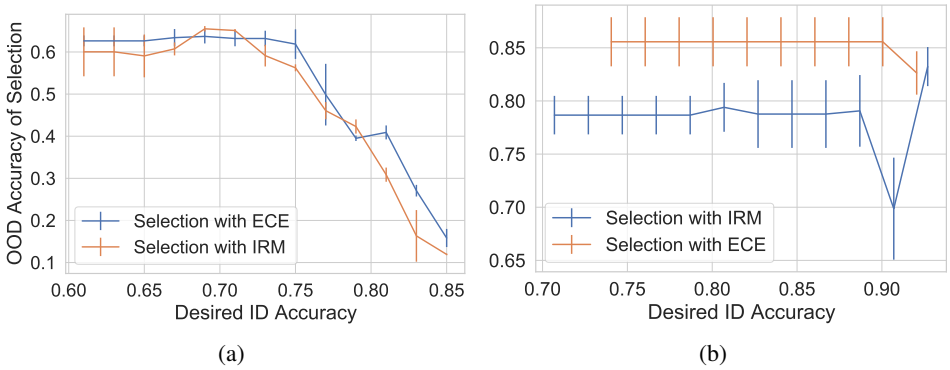


Figure S5: Model selection under a constraint on the ID accuracy. The OOD accuracy obtained by the proposed model selection method is plotted against the desired In-Domain accuracy, acc_{ID}, which is the minimal validation accuracy that we constrain the selected model to achieve. Red marker denotes the performance of the model achieved by selection based on ID validation accuracy alone.

- For each candidate model $\hat{f}$, recalibrate it with Isotonic Regression or some other preferred post-processing technique [10]. Calculate its ID validation error $\text{val}_{\text{ID}}(\hat{f})$ over a held-out dataset. For the held-out dataset from each environment $e \in E_{\text{train}}$ also calculate $ECE^e(\hat{f})$: the ECE of $\hat{f}$ over this dataset. Then take $ECE(\hat{f}) = \sum_{e \in E_{\text{train}}} ECE^e(\hat{f})$.
- Choose $\arg\min_{\hat{f}:\text{val}_{\text{ID}}(\hat{f}) \geq \text{Thr}_{\text{ID}}} ECE(\hat{f})$.

Selection with minimal ECE facilitates a tradeoff between ID accuracy and stability. We use the trained models from the last section (all models trained with either ERM, IRMv1 or CLOvE are pooled into a set of candidates), over environments $e_1 = (0.25, 0.1), e_2 = (0.25, 0.2)$. Selecting the model with minimal $\text{val}_{ID}(\hat{f})$ delivers a classifier with $10.96\%(\pm 0.81)$ accuracy on $e_{\text{test}} = (0.25, 0.9)$ and $85.43\%(\pm 0.13)$ accuracy on the training environments. The trade-off achieved by selection with the proposed criterion is shown in Figure S5. Demanding ID accuracy that is higher than 75% (the ID error obtained by an optimal invariant classifier) yields a relatively sharp drop towards the OOD accuracy obtained by a classifier that purely minimizes empirical error. Going below 75% retrieves a

[10] This is a crucial step, since models that are highly miscalibrated can become well-calibrated upon post-processing

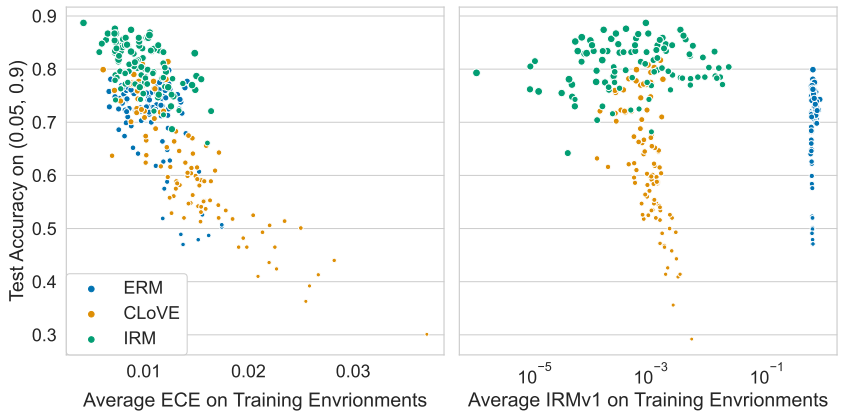


(a) (b)

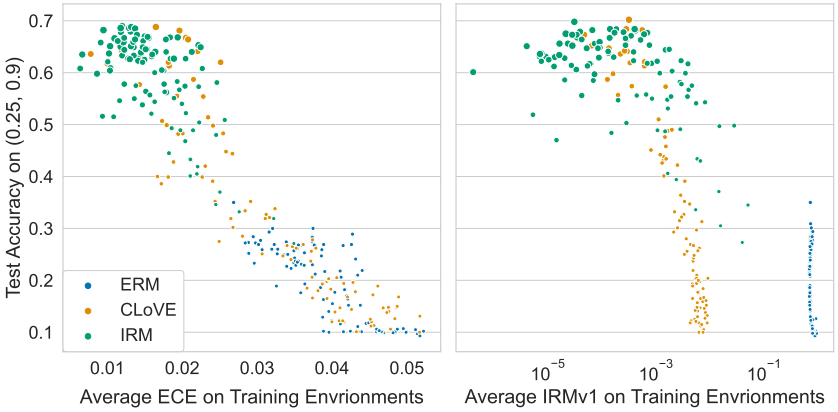
Figure S6: Comparison of proposed model selection procedure applied with IRMv1 objective and the average ECE over training environments in two settings. (a) $e_1 = (0.05, 0.1), e_2 = (0.05, 0.2), e_{\text{test}} = (0.05, 0.9)$ and (b) $e_1 = (0.25, 0.1), e_2 = (0.25, 0.2), e_{\text{test}} = (0.25, 0.9)$.

classifier that achieves $64.98\%(\pm 2.67)\%$ OOD accuracy.

Comparison with IRMv1 Penalty as Selection Criterion. As a baseline to the average ECE over training environments we compare it with using the value of the IRMv1 regularizer, also calculated with a validation set from each training environment. In Figure S6 we compare the curves obtained by the proposed model selection procedure, and that same procedure when replacing the ECE with the value of IRMv1. Figure S6(a) shows the result on the scenario where $e_1 = (0.25, 0.1), e_2 = (0.25, 0.2)$ and $e_{\text{test}} = (0.25, 0.9)$. In this case the two methods are quite comparable, expect for the tail of high desired ID accuracies, where the chosen models are trained with ERM and the IRMv1 criterion fails to rank them by their OOD accuracy. Figure S6(b) shows the same plot on the scenario where $e_1 = (0.05, 0.1), e_2 = (0.05, 0.2)$ and $e_{\text{test}} = (0.05, 0.9)$, which corresponds to the failure case of IRM in Figure 3(a). Due the observation of [20], we may expect the IRMv1 objective to fail at capturing invariance in this setting. Indeed, the model selection done using the IRMv1 penalty gives a worst model than the one selected by ECE in this case. In Figure S7 we also plot the correspondence between OOD accuracy and these quantities (namely ID average ECE, and IRMv1 penalty) as in Figure 3(b) for both settings depicted in Figure S6 showing the erratic behavior of the IRM penalty when considered on different training regimes.



(a)



(b)

Figure S7: Scatter plots of average ECE, and average IRMv1 penalty over training environments against the accuracy on test environments in settings (a) $e_1 = (0.05, 0.1), e_2 = (0.05, 0.2), e_{\text{test}} = (0.05, 0.9)$ and (b) $e_1 = (0.25, 0.1), e_2 = (0.25, 0.2), e_{\text{test}} = (0.25, 0.9)$. Size of marker is proportional to the ratio between OOD and ID accuracies.