ON CALIBRATION AND OUT-OF-DOMAIN GENERALIZATION

Yoav Wald* Hebrew University
Amir Feder* Technion
Daniel Greenfeld Jether Energy Research
Uri Shalit Technion

ABSTRACT

Out-of-domain (OOD) generalization is a significant challenge for machine learning models. To overcome it, many novel techniques have been proposed, 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 prove in a simplified setting that models which achieve multi-domain calibration are free of spurious correlations. Using datasets from the recently proposed WILDS OOD benchmark Koh et al. (2020) we demonstrate that re-calibrating models across multiple domains in a validation set improves performance on unseen test domains. We believe this connection between calibration and OOD generalization is promising from a practical point of view and deserves further research from a theoretical point of view.

1 INTRODUCTION

Recent improvements in the design of machine learning systems have led to impressive results in a plethora of fields Huang et al. (2017); Devlin et al. (2019); Senior et al. (2020). However, as models are typically only trained and tested on in-domain data, they often fail to generalize to out-of-domain (OOD) data Koh et al. (2020). The problem is amplified when deploying such systems in the wild, where they are required to perform well under conditions that might not have been observed during training. Many methods have been proposed to improve the OOD generalization of machine learning models. Specifically, there is a rapidly growing interest in learning models that are invariant to distribution shifts and do not rely on spurious correlations in the training data (Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2019). Such attempts highlight the need for learning robust models with invariance properties, but have demonstrated limited success in scaling to realistic high-dimensional data, and in learning invariant representations (Rosenfeld et al., 2020; Kamath et al., 2021).

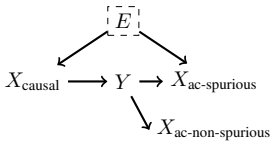

Figure 1: Learning in the presence of causal and anti-causal features. Anti-causal features can be either spurious ($X_{\text{ac-spurious}}$), or non-spurious ($X_{\text{ac-non-spurious}}$).

In this paper, we argue that an alternative approach for learning invariant representations could be achieved through model calibration (Platt et al., 1999; Niculescu-Mizil & Caruana, 2005). Concretely, let X be the features, Y a binary label and E the domains (or environments) (Peters et al., 2016) be random variables with sample spaces $\mathcal{X}, \mathcal{Y} = \{-1, 1\}, \mathcal{E}$ respectively. We assume the data generating process for E, X, Y follows the causal graph in Fig. 1, and that some parts of X are causal and others anti-causal. We differentiate between the anti-causal variables X which are affected and unaffected by E, denoted as $X_{\text{ac-spurious}}$ and $X_{\text{ac-non-spurious}}$, respectively. We do not assume

*Equal contribution, correspondence to yoav.wald@gmail.com, amirfeder@gmail.com

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. If such an arrow exists, it implies 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.

Generally, a representation $\Phi(X)$ contains a *spurious correlation* with respect to the environments E and label Y, if $Y \not\perp\!\!\!\perp E \mid \Phi(X)$. Similar observations have been made by Heinze-Deml et al. (2018); Arjovsky et al. (2019). Having a spurious correlation thus implies that the relation between Φ and the target variable depends on the environment – it is not transferable or stable across environments. Note that "spuriousness" is with respect to the environment random variable E, so the same representation might be spurious with respect to one set of environments and not another. For binary Y when the representation is simply the output of the classifier $f(X)$, we next show why discarding spurious correlations is equivalent, up-to a scalar transformation, to having f ***calibrated*** across all environments.

2 Calibration and Invariant Classifiers

Problem Setting. Let us further refine the definition of the formal setting we consider. We observe training data that has been collected from a certain finite subset of the possible environments $E_{\text{train}} \subset \mathcal{E}$ (note that there is no limitation on $|\mathcal{E}|$). 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 $P[X, Y \mid E = e_i] \quad \forall i \in [k]$. The scope of our formal discussion will be limited to the population setting, hence we do not introduce notation for the datasets collected from each environment. Our goal is to learn a predictor $f : \mathcal{X} \to [0, 1]$ and evaluate its risk with respect to a loss function $l : \mathcal{Y} \times [0, 1] \to \mathbb{R}$. We will refer to the range of f upon limiting its domain to the support of $P[X \mid E = e_i]$, as the range of f restricted to e_i. When taking expectations and conditional expectations over distributions, we will drop the subscripts whenever they are clear from context (e.g. write $\mathbb{E}[\cdot \mid E = e_i]$ instead of $\mathbb{E}_{\mathbf{x},y\sim P[X,Y|E=e_i]}[\cdot]$). All proofs are deferred to the supplementary material.

A classifier is calibrated if it delivers accurate uncertainty estimates for its prediction. The following definition states this for binary classification problems in the case of multiple training environments.

Definition 1. *Let $f : \mathcal{X} \to [0, 1]$, $f(\mathbf{x})$ is calibrated on E_{train} if for all $e_i \in E_{train}$ and α in the range of f restricted to e_i it holds that $\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \alpha$.*

To tie the notion of calibration with OOD generalization, we start by noting its correspondence with our definition of spurious correlations. Recall the definition that a representation $\Phi(X)$ does not contain spurious correlations if $Y \perp\!\!\!\perp E \mid \Phi(X)$. Treating the output of a classifier as a representation of the data, i.e. $\Phi(X) = f(X)$, and considering classifiers which satisfy this conditional independence with respect to training environments, we arrive at a definition of an invariant classifier.

Definition 2. *Let $f : \mathcal{X} \to [0, 1]$, it 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 $\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \mathbb{E}[Y \mid f(X) = \alpha, E = e_j]$.*

The following lemma gives a correspondence between invariant classifiers and calibration on multiple environments:

Lemma 1. *A binary classifier is invariant w.r.t E_{train} if and only if there exists a function $g : \mathbb{R} \to [0, 1]$ such that $g \circ f$ is calibrated on all training environments.*

The above notion of invariance is related to that of IRM (Arjovsky et al., 2019), where invariance of a *representation* $\Phi(X)$ is defined by the existence of classifier that is simultaneously optimal over all environments. When the loss function $l(\cdot)$ is either the squared or logistic loss and the representation is $\Phi(X) = f(X)$, it can be shown that their definition coincides with Definition 2. On the other hand, calibration does not impose risk minimization, which means that it does not depend on the choice of a loss function. Having established the connection between calibration on multiple environments and invariance, there are several interesting questions and points that arise.

Generalization. Suppose that $f(X)$ is calibrated on E_{train}. Under what conditions does this imply it is calibrated on $\mathcal{E}$? Furthermore, to justify our statement that calibration entails discarding spurious

correlations, we should show that $f(X)$ does not use $X_{\text{ac-spurious}}$ to produce its predictions.
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. Hence, calibration should be combined with some sort of guarantee on accuracy. In the calibration literature, this is often referred to as sharpness. Our work leaves much to be explored in this area; nonetheless, in the experimental part we will discuss simple ways to control the trade-off between calibration and sharpness when working with pre-trained models.

2.1 Motivation: a Linear-Gaussian Model

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: $\mathcal{E} = \{(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\}$. We will consider a special case of the setting depicted in figure 1 where all features are anti-causal, it also resembles the setting used in Rosenfeld et al. (2020). In the supplementary material we analyze another example with causal features that undergo covariate shift (along with the spurious features to be discarded).

Y is drawn from a Bernoulli distribution with parameter $\eta \in [0, 1]$, and observed features are generated conditionally on Y. $\mathbf{x}_{\text{ac-ns}} \in \mathbb{R}^{d_{\text{ns}}}$ features are invariant in the sense that their conditional distribution given Y is the same for all environments; $\mathbf{x}_{\text{ac-sp}} \in \mathbb{R}^{d_{\text{sp}}}$ features are spurious as their distribution may shift between environments, altering their correlation with Y. The data generating process for training environment $i \in [k]$ is thus given by:

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

We consider a linear classifier $f(\mathbf{x}; \mathbf{w}, b) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$, where the function $\sigma : \mathbb{R} \to [0, 1]$ is an invertible function (e.g. a sigmoid). $f(\mathbf{x})$ can absorb spurious correlations whenever the coefficients in $\mathbf{w}$ corresponding to $X_{\text{ac-sp}}$ are non-zero. Note that unless these coefficients are set to 0, a new environment can reverse and magnify the correlations observed on E_{train}, leading f to incur an arbitrarily high loss. Our result is as follows:

Theorem 1. *Given $k > 2d_{sp}$ training environments where data is generated according to Eq. (1) with parameters $\{\mu_i, \Sigma_i\}_{i=1}^k$, we say they lie in general position if for all non-zero $\mathbf{x}_{ac\text{-}sp} \in \mathbb{R}^d_{sp}$, $\dim\left(\text{span}\left\{\begin{bmatrix}\Sigma_i \mathbf{x}_{ac\text{-}sp} + \mu_i \\ 1\end{bmatrix}\right\}_{i\in[k]}\right) = d_{sp} + 1$. If a linear classifier is calibrated on k training environments which lie in general position then its coefficients for spurious features 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 $\mathcal{E}^k$.*

Overall, we see that when the number of environments is approximately linear in the number of spurious features then calibration on multiple domains generalizes to calibration on $\mathcal{E}$, and also entails discarding $X_{\text{ac-spurious}}$. The proof of this theorem is given in the supplementary material and uses similar tools to the proof of Theorem 10 in Arjovsky et al. (2019). We now turn to show that methods for model calibration often improve the OOD generalization of pre-trained models.

3 Experiments with Post Processing Calibration

The main part of our experimental evaluation examines post-processing calibration techniques. We examine whether the extremely simple and computationally cheap act of calibrating a model can have an effect on OOD accuracy. The other part, which we defer to the supplementary material, examines the use of calibration scores for model selection. As recently discussed by Gulrajani & Lopez-Paz (2020), model selection is non-trivial when the goal is OOD generalization.

Binary classifiers are often calibrated using Platt Scaling (Platt et al., 1999) or Isotonic Regression (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005)[1]. For our experiments we compared

[1] In multi-class or regression problems we use standard adaptations of these algorithms.

these and several other techniques using the Expected Calibration Error score (Naeini et al., 2015) they achieve. We wound up using Isotonic Regression as it usually outperformed the alternatives. The input to the method is a dataset $(f_i, y_i)_{i=1}^N$, where f_i is a shorthand for the prediction of the model $f(\mathbf{x})$ on the i-th datapoint and y_i is the label. Isotonic Regression learns a *monotonic* mapping $z : \mathbb{R} \to \mathbb{R}$ minimizing the Mean-Squared-Error of $z \circ f$. Unlike standard calibration problems, we have multiple domains to calibrate over. In our experiments we compare two methods for multi-domain calibration: 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. Motivated by the goal of simultaneous calibration, we offer the following alternative where we attempt to bound the worst-case misclaibration across 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, robust isotonic regression solves $z^* = \arg\min_z \max_{e \in E_{\text{train}}} \frac{1}{N_e} \sum_{i=1}^{N_e} (z(f_{e,i}) - y_i)^2$. Since Isotonic Regression can be cast as a Quadratic Program, and this problem minimizes a pointwise maximum on similar objectives, we can cast it as a convex program and solve with standard optimizers. Our implementation uses CVXPY (Diamond & Boyd, 2016). We then evaluate the OOD performance of $z^* \circ f$.

3.1 Results on WILDS Benchmarks

Dataset / Algorithm	PovertyMap				Camelyon17			
	ID	OOD			ID	OOD		
		Orig.	Naive Cal.	Rob. Cal.		Orig.	Naive Cal.	Rob. Cal.
ERM	0.828	0.832	0.827	**0.834**	94.81	66.66	**71.23**	70.93
	(0.018)	(0.011)	(0.014)	(0.006)	(0.023)	(0.144)	(0.089)	(0.086)
DeepCORAL	0.833	0.832	0.835	**0.837**	96.1	73.87	78.51	**79.96**
	(0.009)	(0.011)	(0.009)	(0.012)	(0.0006)	(0.041)	(0.025)	(0.019)
IRM	0.823	0.735	0.812	**0.815**	95.62	77.03	78.99	**79.31**
	(0.012)	(0.117)	(0.016)	(0.015)	(0.029	(0.059)	(0.058)	(0.06)

Table 1: Left: Pearson correlation r on in-domain (ID) and OOD (unseen countries) test sets in *PovertyMap*. Right: Accuracy on ID and OOD (unseen hospital) test sets in *Camelyon17*. Orig.: original algorithm with no changes applied. Best OOD result for each domain is highlighted in **bold**. Standard deviation across model runs in brackets, lowest OOD std. is underlined.

WILDS is a recently proposed benchmark of in-the-wild distribution shifts from several data modalities and applications. We experiment with four datasets, chosen to represent different OOD generalization scenarios. We use models and training algorithms proposed by Koh et al. (2020). To perform multi-domain calibration we modify the data to include a multi-domain validation set. [2]

Table 1 presents our results on the *Camelyon17* and *PovertyMap* datasets, showing that calibration consistently improves OOD performance. Specifically, the robust calibration approach outperforms alternatives in most cases. It also leads to more stable results, as can be seen in the standard deviation across different model runs. In the supplementary we give results on the *CivilComments* and *FMoW* datasets, which have several test environments. Interestingly, in both datasets robust calibration improves the accuracy most significantly on the test environment with lowest accuracy (i.e. the worst-case risk).

4 Conclusion

We highlight the connection between multi-domain calibration and OOD generalization, arguing that such calibration can be viewed as an invariant representation. In a simplified setting we prove that models calibrated on multiple domains are free of spurious correlations and therefore generalize better. We have demonstrated that actively tuning models to achieve multi-domain calibration improves model performance on unseen test domains. The supplementary material includes further experiments showing that in-domain calibration on a validation set is a useful criterion for model

[2] See supp. mat. for a detailed description of the datasets and train-validation splits.

selection. We hope that this work will spur more research, both empirical and theoretical, on the intriguing link between calibration and OOD generalization.

REFERENCES

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

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao 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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

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

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pp. 4171–4186, 2019.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Pritish Kamath, Akilesh Tangella, Danica J Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? *arXiv preprint arXiv:2101.01134*, 2021.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo 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*, pp. 9012–9020, 2019.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pp. 1–31. Springer, 2013.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

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

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nat.*, 577(7792):706–710, 2020. doi: 10.1038/s41586-019-1923-7. URL https://doi.org/10.1038/s41586-019-1923-7.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Carla E. Brodley and Andrea Pohoreckyj Danyluk (eds.), *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pp. 609–616. Morgan Kaufmann, 2001.

A PROOFS FOR THEORETICAL CLAIMS

We begin by supplementing the definition of multiple domain calibration, extending it for the case of regression. This will be required for the second linear-Gaussian model we analyze. 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:

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

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 Kuleshov et al. (2018), 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:

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

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

A.2 INVARIANCE AND MULTIPLE-DOMAIN CALIBRATION

We start by proving our statement regarding the connection between invariant classifiers and calibration w.r.t multiple environments.

Lemma S1. *A binary classifier is invariant w.r.t E_{train} if and only if there exists a function $g : \mathbb{R} \to [0, 1]$ such that $g \circ f$ is calibrated on all training environments.*

Proof. Assume that the classifier is invariant and let $\hat{\alpha} \in \mathbb{R}$ be any value in the range of f. Due to invariance, for all $e_i \in E_{\text{train}}$ where $\hat{\alpha}$ is in the range of f restricted to e_i, there must exist $\alpha \in [0, 1]$ such that $\mathbb{E}[Y \mid f(X) = \hat{\alpha}, E = e_i] = \alpha$. Setting $g(\hat{\alpha}) = \alpha$ results in a classifier $g \circ f$ that is calibrated on E_{train}. Conversely, if a classifier is calibrated on all E_{train}, then setting g to the identity function shows it is also invariant w.r.t E_{train}. □

Let us now turn to the analysis of Linear-Gaussian models.

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.

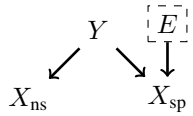

Figure S1: Diagram for data generating process in the invariant features scenario.

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:

$$\frac{\mathbf{w}_{ns}^\top \mu_{ns} + \mathbf{w}_{sp}^\top \mu_i}{\mathbf{w}_{ns}^\top \Sigma_{ns} \mathbf{w}_{ns} + \mathbf{w}_{sp}^\top \Sigma_i \mathbf{w}_{sp}} = t \quad \forall i \in [k]. \tag{3}$$

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 Eq. (3)), 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 Rosenfeld et al. (2020). 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 Arjovsky et al. (2019).

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 $\mathbf{x} \in \mathbb{R}^d_{sp}$:*

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

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 Arjovsky et al. (2019), 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 Eq. (3) 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:

$$\mathbf{w}_{sp}^\top\Sigma_i\mathbf{w}_{sp} - \mu_i^\top\mathbf{w}_{sp} - t = 0 \quad \forall i \in [k]. \tag{4}$$

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 Eq. (4) 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 Arjovsky et al. (2019).

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 Arjovsky et al. (2019), 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 Lee (2013), Example 5.30; our proof follows a similar line of reasoning). Now write A as a block matrix using $B \in \mathbb{R}^{d_{\text{sp}}\times d_{\text{sp}}}, C \in \mathbb{R}^{(k-d_{\text{sp}})\times d_{\text{sp}}}$:

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

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 next move to a second scenario 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:

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

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 $[f(\mathbf{x}; \mathbf{w})]_2 = c$.[3] We decompose the weights $\mathbf{w}$

[3]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 Eq. (2), 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.

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.

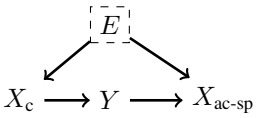

Figure S2: Diagram for data generating process in the covariate shift scenario.

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 $\mathbb{R}^{d_{sp}}$.*
- *$k > d_c + d_{sp} + 1$ 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:

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

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:

$$\frac{\sigma_{f,y,i}}{\sigma_{f,i}^2} = t \quad \forall i \in [k]. \tag{5}$$

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 Eq. (5) 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 $\mathbf{w}_{sp}, \bar{\mathbf{w}}_c, t_2$:

$$\bar{\mathbf{w}}_c^\top \mu_i^c - \mathbf{w}_{sp}^\top \mu_i (\mathbf{w}_c^{*\top} \mu_i^c) + t_2 = 0 \quad \forall i \in [k]. \tag{6}$$

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 Eq. (5) 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 Eq. (6) 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 Eq. (6), 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. □

B ADDITIONAL EXPERIMENTS

Next we move to present our experimental results on the remaining WILDS datasets (Koh et al., 2020) and on model selection with IRM (Arjovsky et al., 2019) and ColoredMNIST (Kim et al., 2019).

B.1 POST-PROCESSING CALIBRATION ON CIVILCOMMENTS AND FMOW

Table S1 presents results on the *FMoW* dataset, using the same post-processing calibration techniques described in the paper. As this is a multi-class classification task, where calibration methods were shown to be less effective, differences between all approaches are substantially smaller. While the average performance does not substantially increase following the calibration post-process, it does significantly improve the worst-case performance for all training algorithms.

	Algorithm	Orig.	Naive Cal.	Rob. Cal.
Average	ERM	52.98 (0.003)	**53.01** (0.004)	52.83 (0.004)
	DeepCORAL	50.61 (0.003)	**50.71** (0.003)	50.5 (0.003)
	IRM	50.72 (0.004)	**51.05** (0.004)	50.92 (0.003)
Worst	ERM	32.63 (0.016)	33.09 (0.021)	**37.19** (0.035)
	DeepCORAL	31.73 (0.01)	31.75 (0.01)	**33.86** (0.016)
	IRM	31.33 (0.012)	31.81 (0.016)	**34.41** (0.015)

Table S1: Average (top) and worst (bottom, region) accuracy on the OOD test set in *FMoW*.

Observing the *CivilComments* results (Table S2), ERM's performance is not improved substantially when calibrating it, but IRM is improved by 31.7% (absolute) after calibration. Interestingly, the very large variance observed in the IRM runs (0.199) is reduced by a factor of 16 after applying robust calibration (0.012). Looking at worst-case performance across demographic groups, we see calibration is very useful, improving performance by an average of 17.8%.

B.2 MODEL SELECTION METHOD

To use multiple domain calibration as an observable surrogate for OOD performance we use standard tools to post-process and assess calibration of a model: *calibration plots* and scalar summaries such as the Expected Calibration Error, known as the *ECE score*.

Calibration plots DeGroot & Fienberg (1983) 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

	Algorithm	Orig.	Naive Cal.	Rob. Cal.
Average	ERM	92.06 (0.004)	92.32 (0.004)	**92.63** (0.003)
	IRM	55.74 (0.199)	87.21 (0.016)	**87.47** (0.012)
	GroupDRO	89.35 (0.006)	92.18 (0.003)	**92.56** (0.001)
Worst	ERM	62.09 (0.026)	76.38 (0.005)	**79.93** (0.008)
	IRM	40.61 (0.16)	**69.09** (0.013)	68.76 (0.013)
	GroupDRO	72.02 (0.004)	76.69 (0.013)	**79.31** (0.007)

Table S2: Average (top) and worst-case (bottom) group accuracy on the test set in the *CivilComments* dataset.

against each other to form a curve, where deviations from a diagonal represent miscalibration.
ECE score is one common scalar summary of the calibration curve, calculated by averaging the deviation between accuracy and confidence:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|. \tag{7}$$

n_b is the number of predictions in bin b, N is the total number of data points.
Isotonic Regression. The input to the method is a dataset $(f_i, y_i)_{i=1}^N$, where f_i is a shorthand for the prediction of the model $f(\mathbf{x})$ on the i-th datapoint and y_i is the label. Isotonic Regression learns a *monotonic* mapping $z : \mathbb{R} \to \mathbb{R}$ solving:

$$(ISO) \quad \arg\min_z \frac{1}{N} \sum_{i=1}^{N} (z(f_i) - y_i)^2.$$

Our experiment examines the use of calibration scores for model selection. As recently discussed by Gulrajani & Lopez-Paz (2020), model selection is non-trivial when the goal is OOD generalization. They show that suboptimal model selection can eliminate the advantage that techniques designed for OOD generalization have over ERM with data augmentation. As most model selection methods are based on accuracy, we consider a method based on the ECE score instead.

Algorithm 1 Model Selection with Worst-Case ECE

Input: Validation sets $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^k$, Models $\{f_j(\mathbf{x})\}_{j=1}^L$
Output: Selected model $f^*(\mathbf{x})$
  Pool data from all environments to $(\mathbf{X}, \mathbf{y})$
  for $j \in [L]$ **do**
    $z_j \leftarrow ISO(f_j(\mathbf{X}), \mathbf{y})$
    for $i \in [k]$ **do**
      $ECE_{i,j} \leftarrow ECE$ of $z_j \circ f_j$ on $(\mathbf{X}_i, \mathbf{y}_i)$
  Return f_j for $j \in \arg\min_j \max_i ECE_{i,j}$

The model selection procedure we examine takes L pretrained models on k training environments, along with validation sets for each environment. As described in algorithm 1 it selects a model based on the best worst-case ECE over environments. We follow Naeini et al. (2015) and calibrate each model before calculating the ECE. This is important since ECE can be very high on models that actually become well calibrated upon simple post-processing.

B.3 Experimental Setup and Results for Model Selection

Colored MNIST is a variant of the classical MNIST dataset LeCun (1998), where a color bias was planted into each digit Kim et al. (2019). We use it to test the efficacy of using calibration measures for model selection, choosing hyperparameters for an IRM trained multilayer perceptron over 25 restarts. In their experiments Arjovsky et al. (2019) perform model selection by measuring error on data taken from the test environment, which is undesirable in practice. Our experiment serves as a proof of concept for calibration based model selection in such scenarios. We test the ability of doing model selection with in-domain calibration metrics on Colored MNIST. Using the validation set of two training environments we choose a model with algorithm 1, according to the best worst-case ECE. Colorings in Colored MNIST are drawn randomly, so we average the results over different instantiations. The average test environment accuracy of the model chosen by our method is $63.1\%(\pm 3.2)$. It is superior to selection based on the best worst-case validation accuracy over training domains, which chooses a model with $52\%(\pm 0.4)$ accuracy. The model with overall best test performance has 66.9% accuracy, though of course that requires evaluating on the test set. In figure S3 we show the test accuracy vs. average worst environment ECE.

C Dataset Statistics and Models

For each of the five 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

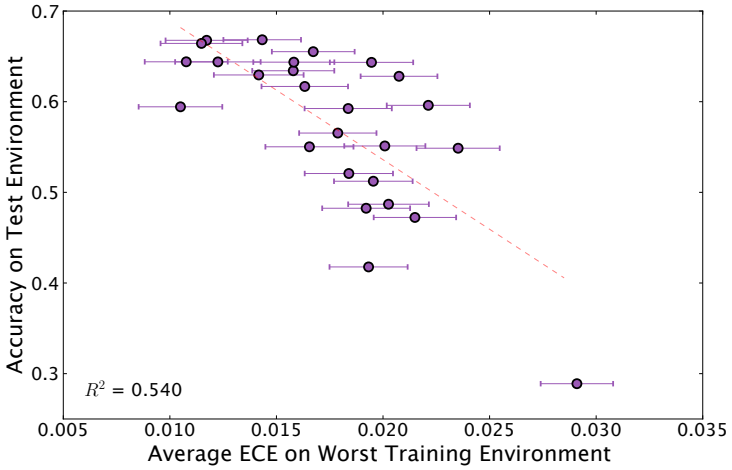

Figure S3: Accuracies on test environment of Colored MNIST for all trained models against their worst-case train ECEs. Dotted line marks a linear fit of the test accuracy from the observed ECE.

training set, and the calibrators on the validation set. We then compare all alternatives (Original, Naive Calibration and Robust Calibration) on the held-out test set (OOD). Whenever an In-Domain (ID) test set is available (*PovertyMap*, *Camelyon17* and *AmazonReviews*), we evaluate the model on it as well.

C.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 Yeh et al. (2020), 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.

C.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.

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 Bandi et al. (2018). 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.

C.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 Borkan et al. (2019). 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.

C.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 Christie et al. (2018), 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 Koh et al. (2020), 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 7.

- **BERT** - BERT is a 12-layer Transformer model Vaswani et al. (2017) that represents textual inputs contextually and sequentially Devlin et al. (2019). 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 Desai & Durrett (2020). 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 Huang et al. (2017). 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 He et al. (2016). 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 Koh et al. (2020). Each model is trained four times, using a different random seed at each run. We report performance averages and their standard deviation in Section 7.

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) Arjovsky et al. (2019) is a training algorithm that penalizes feature distributions that have different optimal linear classifiers for each environment.
- **DeepCORAL** DeepCORAL Sun & Saenko (2016) 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 Gulrajani & Lopez-Paz (2020).
- **GroupDRO** - Group DRO Hu et al. (2018) 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 Koh et al. (2020).