Environment Inference for Invariant Learning

Elliot Creager [1 2] **Jörn-Henrik Jacobsen** [1 2] **Richard Zemel** [1 2]

Abstract

Learning models that gracefully handle distribution shifts is central to research on domain generalization, robust optimization, and fairness. A promising formulation is domain-invariant learning, which identifies the key issue of learning which features are domain-specific versus domain-invariant. An important assumption in this area is that the training examples are partitioned into "domains" or "environments". Our focus is on the more common setting where such partitions are not provided. We propose EIIL, a general framework for domain-invariant learning that incorporates Environment Inference to directly infer partitions that are maximally informative for downstream Invariant Learning. We show that EIIL outperforms invariant learning methods on the CMNIST benchmark without using environment labels, and significantly outperforms ERM on worst-group performance in the Waterbirds and CivilComments datasets. Finally, we establish connections between EIIL and algorithmic fairness, which enables EIIL to improve accuracy and calibration in a fair prediction problem.

(a) **Inferred environment 1** *(mostly) landbirds on land, and waterbirds on water*

(b) **Inferred environment 2** *(mostly) landbirds on water, and waterbirds on land*

Figure 1. In the Waterbirds dataset (Sagawa et al., 2020), the two target labels (landbirds and waterbirds) are correlated with their respective typical background habitats (land and water). This spurious correlation causes sub-par performance on the smallest subgroups (e.g. waterbirds on land). Environment Inference for Invariant Learning (EIIL) organizes the training data into two environments that are maximally informative for use by a downstream invariant learner, enabling the use of invariant learning in situations where environment labels are not readily available. By grouping examples where class and background disagree into the same environment, EIIL encourages learning an invariance w.r.t. background features, which improves worst-group test accuracy by 18% relative to standard supervised learning.

1. Introduction

Machine learning achieves super-human performance on many tasks when the test data is drawn from the same distribution as the training data. However, when the two distributions differ, model performance can severely degrade, even to below-chance predictions (Geirhos et al., 2020). Tiny perturbations can derail classifiers, as shown by adversarial examples (Szegedy et al., 2014) and common image corruptions (Hendrycks & Dietterich, 2019). Even new test sets collected from the same data acquisition pipeline induce distribution shifts that significantly harm performance (Recht et al., 2019; Engstrom et al., 2020). Many approaches have been proposed to overcome the brittleness of supervised learning—e.g. Empirical Risk Minimization (ERM)—in the face of distribution shifts. Robust optimization aims to achieve good performance on any distribution close to the training distribution (Goodfellow et al., 2015; Duchi et al., 2021; Madry et al., 2018). Invariant learning on the other hand tries to go one step further, to generalize to distributions potentially far away from the training distribution.

However, common invariant learning methods typically come at a serious disadvantage: they require datasets to be partitioned into multiple domains or environments.[1] Environment assignments should implicitly define variation the algorithm should become invariant or robust to, but often such environment labels are unavailable at training time, either because they are difficult to obtain or due to privacy

[1]University of Toronto [2]Vector Institute. Correspondence to: Elliot Creager <creager@cs.toronto.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]We use "domains", "environments" and "groups"/"subgroups" interchangeably.

limitations. In some cases, relevant side-information or metadata, e.g., human annotations, or device ID used to take a medical image, hospital or department ID, etc., may be abundant, but it remains unclear how best to specify environments based on this information (Srivastava et al., 2020). A similar issue arises in mitigating algorithmic unfairness, where so-called sensitive attributes may be difficult to define in practice (Hanna et al., 2020), or their values may be impossible to collect. We aim to overcome the difficulty of manual environment specification by developing a new method inspired by fairness approaches for unknown group memberships (Kim et al., 2019; Lahoti et al., 2020).

The core idea is to leverage the bias of an ERM-trained reference model to discover useful environment partitions directly from the training data. We derive an environment inference objective that maximizes variability across environments, and is differentiable w.r.t. a distribution over environment assignments. After performing environment inference given a fixed reference classifier, we use the inferred environments to train an invariant learner from scratch.

Our method, **Environment Inference for Invariant Learning (EIIL)**, discovers environment labels that can then be used to train any off-the-shelf invariant learning algorithm in applications where environment labels are unavailable. This approach can outperform ERM in settings where standard learning tends to focus on spurious features or exhibit performance discrepancies between subgroups of the training data (which need not be specified ahead of time). EIIL discovers environments capturing spurious correlations hidden in the dataset (see Figure 1), making them readily available for invariant learning. Surprisingly, even when manual specification of environments is available (e.g. the CMNIST benchmark), inferring environments directly from aggregated data may *improve* the quality of invariant learning.

Our main contributions are as follows:

- We propose a general framework for inferring environments from data based on the bias of a reference classifier;
- we provide a theoretical characterization of the dependence on the reference classifier, and when we can expect the method to do well;
- we derive a specific instance of environment inference in this framework using gradients w.r.t. soft environment assignments, which outperforms invariant learning (using environment labels) on the CMNIST benchmark and outperforms ERM on Waterbirds;
- we establish a connection to similar themes in the fairness literature, and show that our method can improve accuracy and calibration in a fair prediction problem.

2. Invariant Learning

This section discusses the problem setting and presents background materials that will be used to formulate our proposed method. Our approach is primarily motivated by recent approaches to learning domain- or environment-invariant representations—which we simply refer to as "invariant learning"—that have been applied to domain adaptation and generalization tasks.

Notation Let $\mathcal{X}$ be the input space, $\mathcal{E}^{obs}$ the set of training environments (a.k.a. "domains"), $\mathcal{Y}$ the target space. Let $x, y, e \sim p^{obs}(x, y, e)$ be observational data, with $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $e \in \mathcal{E}^{obs}$. $\mathcal{H}$ denotes a representation space, from which a classifier $w \circ \Phi$ (that maps to the logit space of $\mathcal{Y}$ via a linear map w) can be applied. $\Phi : \mathcal{X} \to \mathcal{H}$ denotes the parameterized mapping or "model" that we optimize. We refer to $\Phi(x) \in \mathcal{H}$ as the "representation" of example x. $\hat{y} \in \mathcal{Y}$ denotes a hard prediction derived from the classifier by stochastic sampling or probability thresholding. $\ell : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}$ denotes a scalar loss, which guides the learning.

The empirical risk minimization (ERM) solution is found by minimizing the global risk, expressed as the expected loss over the observational distribution:

$$C^{ERM}(\Phi) = \mathbb{E}_{p^{obs}(x,y,e)}[\ell(\Phi(x), y)]. \tag{1}$$

Representation Learning with Environment Labels Domain generalization is concerned with achieving low error rates on unseen test distributions $p(x, y|e_{test})$ for $e_{test} \notin \mathcal{E}^{obs}$. Domain adaption is a related problem where model parameters can be adapted at test time using unlabeled data. Recently, *Invariant Learning* approaches such as Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) and Risk Extrapolation (REx) (Krueger et al., 2021) were proposed to overcome the limitations of adversarial domain-invariant representation learning (Zhao et al., 2019) by discovering invariant relationships between inputs and targets across domains. Invariance serves as a proxy for causality, as features representing "causes" of target labels rather than effects will generalize well under intervention. In IRM, a representation $\Phi(x)$ is learned that performs optimally within each environment—and is thus invariant to the choice of environment $e \in \mathcal{E}^{obs}$—with the ultimate goal of generalizing to an unknown test dataset $p(x, y|e_{test})$. Because optimal classifiers under standard loss functions can be realized via a conditional label distribution ($f^*(x) = \mathbb{E}[y|x]$), an invariant representation $\Phi(x)$ must satisfy the following *Environment Invariance Constraint*:

$$\mathbb{E}[y|\Phi(x) = h, e_1] = \mathbb{E}[y|\Phi(x) = h, e_2] \quad \forall\, h \in \mathcal{H} \;\; \forall\, e_1, e_2 \in \mathcal{E}^{obs}. \tag{EIC}$$

Intuitively, the representation $\Phi(x)$ encodes features of the input x that induce the same conditional distribution over

labels across each environment. This is closely related to the notion of "group sufficiency" studied in the fairness literature (Liu et al., 2019) (see Appendix C).

Because trivial representations such as mapping all x onto the same value may satisfy environment invariance, other objectives must be introduced to encourage the predictive utility of Φ. Arjovsky et al. (2019) propose IRM as a way to satisfy (EIC) while achieving a good overall risk. As a practical instantiation, the authors introduce IRMv1, a regularized objective enforcing simultaneous optimality of the same classifier $w \circ \Phi$ in all environments;[2] here w.l.o.g. $w = \bar{w}$ is a constant scalar multiplier of 1.0 for each output dimension. Denoting by $R^e = \mathbb{E}_{p^{obs}(x,y|e)}[\ell]$ the per-environment risk, the objective to be minimized is

$$C^{IRM}(\Phi) = \sum_{e \in \mathcal{E}^{obs}} R^e(\Phi) + \lambda ||\nabla_{\bar{w}} R^e(\bar{w} \circ \Phi)||. \tag{IRMv1}$$

Robust Optimization Another approach at generalizing beyond the training distribution is robust optimization (Ben-Tal et al., 2009), where one aims to minimize the worst-case loss for every subset of the training set, or other well-defined perturbation sets around the data (Duchi et al., 2021; Madry et al., 2018). Rather than optimizing a notion of invariance, Distributionally Robust Optimization (DRO) (Duchi et al., 2021) seeks good performance for all nearby distributions by minimizing the worst-case loss: $\max_q \mathbb{E}_q[\ell]$ s.t. $D(q||p) < \epsilon$, where D denotes similarity between two distributions (e.g., χ^2 divergence) and ϵ is a hyperparameter. The objective can be computed as an expectation over p via per-example importance weights $\gamma_i = \frac{q(x_i,y_i)}{p(x_i,y_i)}$. GroupDRO operationalizes this principle by sharing importance weights across training examples, using environment labels to define relevant groups for this parameter sharing. This can be expressed as an expected risk under a worst-case distribution over group proportions:

$$C^{GroupDRO}(\Phi) = \max_g \mathbb{E}_{g(e)}[R^e(\Phi)]$$

This is a promising approach towards tackling distribution shift with deep nets (Sagawa et al., 2020), and we show in our experiments how environment inference enables application of GroupDRO to improve over standard learning without requiring group labels.

Limitations of Invariant Learning While the use of invariant learning to tackle domain generalization is still relatively nascent, several known limitations merit discussion. IRM can provably find an invariant predictor that generalizes OOD, but only under restrictive assumptions, such as linearity of the data generative process and access to many environments (Arjovsky et al., 2019). However, most benchmark datasets are in the non-linear regime; Rosenfeld et al. (2021) demonstrated that for some non-linear datasets, the IRMv1 penalty term induces multiple optima, not all of which yield invariant predictors. Nevertheless, IRM has found empirical success in some high dimensional non-linear classification tasks (e.g. CMNIST) using just a few environments (Arjovsky et al., 2019; Koh et al., 2021). On the other hand, it was recently shown that, using careful and fair model selection strategies across a suite of image classification tasks, neither IRM nor other invariant learners consistently beat ERM in OOD generalization (Gulrajani & Lopez-Paz, 2021). This last study underscores the importance of *model selection* in any domain generalization approach, which we discuss further below.

[2] $w \circ \Phi$ yields a classification decision via linear weighting on the representation features.

3. Invariance Without Environment Labels

In this section we propose a novel invariant learning framework that does not require a priori domain/environment knowledge. This framework is useful in algorithmic fairness scenarios when demographic makeup is not directly observed; it is also applicable in standard machine learning settings when relevant environment information is either unavailable or not clearly identified. In both cases, a method that sorts training examples $\mathcal{D}$ into environments that maximally separate the spurious features—i.e. inferring populations $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$—can facilitate effective invariant learning.

3.1. Environment Inference for Invariant Learning

Our aim is to find environments that maximally violate the invariant learning principle. We can then evaluate the quality of these inferred environments by utilizing them in an invariant learning method. Our overall algorithm EIIL is a two-stage process: (1) Environment Inference (EI): infer the environment assignments; and (2) Invariant Learning (IL): run invariant learning given these assignments.

The primary goal of invariant-learning is to find features that are domain-invariant, i.e, that reliably predict the true class regardless of the domain. The EI phase aims to identify domains that help uncover these features. This phase depends on a reference classifier $\tilde{\Phi}$; which maps inputs X to outputs Y, and defines a putative set of invariant features. This model could be found using ERM on $p^{obs}(x, y)$, for example. Environments are then derived that partition the mapping of the reference model which maximally violate the invariance principle, i.e., where for the reference classifier the same feature vector is associated with examples of different classes. While any of the aforementioned invariant learning objectives can be incorporated into the EI phase, the invariance principle or group-sufficiency—as expressed

in (EIC)—is a natural fit, since it explicitly depends on learned feature representations Φ.

To realize an EI phase focused on the invariance principle, we utilize the IRM objective (IRMv1). We begin by noting that the per-environment risk R^e depends implicitly on the manual environment labels from the dataset. For a given environment e', we denote $\mathbb{1}(e_i = e')$ as an indicator that example i is assigned to that environment, and re-express the per-environment risk as:

$$R^e(\Phi) = \frac{1}{\sum_{i'} \mathbb{1}(e_{i'} = e)} \sum_i \mathbb{1}(e_i = e)\ell(\Phi(x_i), y_i) \quad (2)$$

Now we relax this risk measure to search over the space of environment assignments. We replace the manual assignment indicator $\mathbb{1}(e_i = e')$, with a probability distribution $\mathbf{q}_i(e') := q(e'|x_i, y_i)$, representing a soft assignment of the i-th example to the e'-th environment. To *infer* environments, we optimize $q(e|x_i, y_i)$ so that it captures the worst-case environments for a fixed classifier Φ. This corresponds to maximizing w.r.t. $\mathbf{q}$ the following soft relaxation of the regularizer[3] from C^{IRM}:

$$C^{EI}(\Phi, \mathbf{q}) = ||\nabla_{\bar{w}} \tilde{R}^e(\bar{w} \circ \Phi, \mathbf{q})||, \quad (3)$$

$$\tilde{R}^e(\Phi, \mathbf{q}) = \frac{1}{\sum_{i'} \mathbf{q}_{i'}(e)} \sum_i \mathbf{q}_i(e)\ell(\Phi(x_i), y_i) \quad (4)$$

where $\tilde{R}^e$ represents a soft per-environment risk that can pass gradients to the environment assignments $\mathbf{q}$. See Algorithm 1 in Appendix A for pseudocode.

To summarize, EIIL involves the following sequential[4] approach:

1. Input *reference model* $\tilde{\Phi}$;
2. Fix $\Phi \leftarrow \tilde{\Phi}$ and optimize the EI objective to infer environments: $\mathbf{q}^* = \arg\max_{\mathbf{q}} C^{EI}(\tilde{\Phi}, \mathbf{q})$;
3. Fix $\tilde{\mathbf{q}} \leftarrow \mathbf{q}^*$ and optimize the IL objective to yield the new model: $\Phi^* = \arg\min_{\Phi} C^{IL}(\Phi, \tilde{\mathbf{q}})$

In our experiments we consider binary environments and parameterize the $\mathbf{q}$ as a vector of probabilities for each example in the training data.[5] EIIL is applicable more broadly to any environment-based invariant learning objective through the choice of C^{IL} in Step 3. We present experiments using $C^{IL} \in \{C^{IRM}, C^{GroupDRO}\}$, and leave a more complete exploration to future work.

[3] We omit the average risk term as we are focused on maximally violating (EIC) regardless of the risk.

[4] We also tried jointly training Φ and $\mathbf{q}$ using alternating updates, as in GAN training, but did not find empirical benefits. This formulation introduces optimization and conceptual difficulties, e.g. ensuring that invariances apply to all environments discovered throughout learning.

[5] Note that under this parameterization, when optimizing the inner loop with fixed Φ the number of parameters equals the number of data points (which is small relative to standard neural net training). We leave amortization of q to future work.

3.2. Analyzing the Inferred Environments

To characterize the ability of EIIL to generalize to unseen test data, we now examine the inductive bias for generalization provided by the reference model $\tilde{\Phi}$. We state the main result here and defer the proofs to Appendix B. Consider a dataset with some feature(s) z which are spurious, and other(s) v which are valuable/invariant/causal w.r.t. the label y. Our proof considers binary features/labels and two environments, but the same argument extends to other cases. Our goal is to find a model Φ whose representation $\Phi(v, z)$ is invariant w.r.t. z and focuses solely on v.

Proposition 1 *Consider environments that differ in the degree to which the label y agrees with the spurious features z: $\mathbb{P}(\mathbb{1}(y = z)|e_1) \neq \mathbb{P}(\mathbb{1}(y = z)|e_2)$: then a reference model $\tilde{\Phi} = \Phi_{Spurious}$ that is invariant to valuable features v and solely focuses on spurious features z maximally violates the invariance principle (EIC). Likewise, consider the case with fixed representation Φ that focuses on the spurious features: then a choice of environments that maximally violates (EIC) is $e_1 = \{v, z, y|\mathbb{1}(y = z)\}$ and $e_2 = \{v, z, y|\mathbb{1}(y \neq z)\}$.*

If environments are split according to agreement of y and z, then the constraint from (EIC) is satisfied by a representation that ignores z: $\Phi(x) \perp z$. Unfortunately this requires a priori knowledge of either the spurious feature z or a reference model $\tilde{\Phi} = \Phi_{Spurious}$ that extracts it. When the suboptimal solution $\Phi_{Spurious}$ is not a priori known, it will sometimes be recovered directly from the training data; for example in CMNIST we find that Φ_{ERM} approximates Φ_{Color}. This allows EIIL to find environment partitions providing the starkest possible contrast for invariant learning.

Even if environment partitions are available, it may be possible to improve performance by inferring new partitions from scratch. It can be shown (see Appendix B.3) that the environments provided in the CMNIST dataset (Arjovsky et al., 2019) do not maximally violate (EIC) for a reference model $\tilde{\Phi} = \Phi_{Color}$, and are thus not maximally informative for learning to ignore color. Accordingly, EIIL improves test accuracy for IRM compared with the hand-crafted environments (Table 2).

If $\tilde{\Phi} = \Phi_{ERM}$ focuses on a mix of z and v, EIIL may still find environment partitions that enable effective invariant learning, as we find in the Waterbirds dataset, but they are not guaranteed to maximally violate (EIC).

3.3. Binned Environment Invariance

We can derive a heuristic algorithm for EI that maximizes violations of the invariance principle by stratifying examples into discrete bins (i.e. confidence bins for 1-D representations), then sorting them into environments within each bin. This algorithm provides insight into both the EI task and the relationship between the IRMv1 regularizer and the invariance principle. We define bins in the space of the learned representation $\Phi(x)$, indexed by b; s_{ib} indicates whether example i is in bin b. The intuition behind the algorithm is that a simple approach can separate the examples in a bin to achieve the maximal value of the (EIC).

The degree to which the environment assignments violate (EIC) can be expressed as follows, which can then be approximated in terms of the bins:

$$\begin{aligned}\Delta\text{EIC} &= (E[y|\Phi(x), e_1] - E[y|\Phi(x), e_2])^2 \\ &\approx \sum_b (\sum_i s_{ib} y_i \mathbf{q}_i(e = e_1) - \sum_i s_{ib} y_i \mathbf{q}_i(e = e_2))^2\end{aligned}$$

Inspection of this objective leads to a simple algorithm: assign all the $y = 1$ examples to one environment, and $y = -1$ examples to the other. This results in the expected values of y equal to ± 1, which achieves the maximum possible value of ΔEIC per bin. [6]

This binning leads to an important insight into the relationship between the IRMv1 regularizer and (EIC). Despite the analysis in Arjovsky et al. (2019), this link is not completely clear (Kamath et al., 2021; Rosenfeld et al., 2021). However, in the situation considered here, with binary classes, we can use this binning approach to show a tight link between the two objectives: finding an environment assignment that maximizes the violation of our softened IRMv1 regularizer (Equation 3) also maximizes the violation of the softened Environment Invariance Constraint (ΔEIC); see Appendix B.2 for the proof. This binning approach highlights the dependence on the reference model, as the bins are defined in its learned Φ space; the reference model also played a key role in the analysis above. We analyze it empirically in Section 5.3.

4. Related Work

Domain adaptation and generalization Beyond the methods discussed above, a variety of recent works have approached the domain generalization problem from the lens of learning invariances in the training data. Adversarial training is a popular approach for learning representations invariant (Zhang et al., 2017; Hoffman et al., 2018; Ganin et al., 2016) or conditionally invariant (Li et al., 2018) to the environment. However, this approach has limitations in settings where distribution shift affects the marginal distribution over labels (Zhao et al., 2019).

Arjovsky et al. (2019) proposed IRM to mitigate the effect of test-time label shift, which was inspired by applications of causal inference to select invariant features (Peters et al., 2016). Krueger et al. (2021) proposed the related Risk Extrapolation (REx) principle, which dictates a stronger preference to exactly equalize $R^e \; \forall \; e$ (e.g. by penalizing variance across e as in their practical algorithm V-REx), which is shown to improve generalization in several settings.[7]

Recently, Ahmed et al. (2021) proposed a new invariance regularizer based on matching class-conditioned average predictive distributions across environments, which we note is closely related to the equalized odds criterion commonly used in fair classification (Hardt et al., 2016). Moreover, they deploy this training on top of environments inferred by our EI method, showing that the overall EIIL approach can effectively handle "systematic" generalization (Bahdanau et al., 2019) on a semi-synthetic foreground/background task similar to the Waterbirds dataset that we study.

Several large-scale benchmarks have recently been proposed to highlight difficult open problems in this field, including the use of real-world data (Koh et al., 2021), handling subpopulation shift (Santurkar et al., 2021), and model selection (Gulrajani & Lopez-Paz, 2021).

Leveraging a reference classifier A number of methods have recently been proposed that improve performance by exploiting the mistakes of a pre-trained auxiliary model, as we do when inferring environments for the invariant learner using $\tilde{\Phi}$. Nam et al. (2020) jointly train a "biased" model f_B and a "debiased" model f_D, where the relative cross-entropy losses of f_B and f_D on each training example determine their importance weights in the overall training objective for f_D. Sohoni et al. (2020) infer a different set of "hidden subclasses" for each class label $y \in \mathcal{Y}$, Subclasses computed in this way are then used as group labels for training a GroupDRO model, so the overall two-step process corresponds to certain choices of EI and IL objectives.

Liu et al. (2021) and Dagaev et al. (2021) concurrently proposed to compute importance weights for the primary model using an ERM reference, which can be seen as a form of distributionally robust optimization where the worst-case distribution only updates once. Dagaev et al. (2021) use the confidence of the reference model to assign importance weights to each training example. Liu et al. (2021) split the training examples into two disjoint groups based on the errors of ERM, akin to our EI step, with the per-group

[6] Multiple global optima exist, this heuristic is not the only possible solution. For very confident reference models, where few confidence bins are populated, this relates to partitioning based on the error cases.

[7] Analogous to V-REx, Williamson & Menon (2019) adapt Conditional Variance at Risk (CVaR) (Rockafellar & Uryasev, 2002) to equalize risk across demographic groups.

Statistic to match/optimize	e known?	Dom-Gen method	Fairness method
match $\mathbb{E}[\ell\|e]$ $\forall e$	yes	REx (Krueger et al., 2021),	CVaR Fairness (Williamson & Menon, 2019)
$\min \max_e \mathbb{E}[\ell\|e]$	yes	Group DRO (Sagawa et al., 2020)	
$\min \max_q \mathbb{E}_q[\ell]$	no	DRO (Duchi et al., 2021)	Fairness without Demographics (Hashimoto et al., 2018; Lahoti et al., 2020)
match $\mathbb{E}[y\|\Phi(x), e]$ $\forall$ e	yes	IRM (Arjovsky et al., 2019)	Group Sufficiency (Chouldechova, 2017; Liu et al., 2019)
match $\mathbb{E}[y\|\Phi(x), e]$ $\forall$ e	no	**EIIL (ours)**	**EIIL (ours)**
match $\mathbb{E}[\hat{y}\|\Phi(x), e, y = y']$ $\forall$ e	yes	C-DANN (Li et al., 2018) PGI (Ahmed et al., 2021)	Equalized Odds (Hardt et al., 2016)

Table 1. Domain Generalization (Dom-Gen) and Fairness methods can be understood as matching or optimizing some statistic across conditioning variable e, representing "environment" or "domains" in Dom-Gen and "sensitive" group membership in the Fairness.

importance weights treated as a hyperparameter for model selection (which requires a subgroup-labeled validation set). We note that the implementation of EIIL using binning heuristic, discussed in 3.3, can also realize an error splitting behavior when the reference classifier is very confident. In this case, both methods use the same disjoint groups of training examples towards slightly different ends: we train an invariant learner, whereas Liu et al. (2021) train a cross-entropy classifier with fixed per-group importance weights.

Algorithmic fairness Our work draws inspiration from a rich body of recent work on learning fair classifiers in the absence of demographic labels (Hébert-Johnson et al., 2018; Kearns et al., 2018; Hashimoto et al., 2018; Kim et al., 2019; Lahoti et al., 2020). Generally speaking, these works seek a model that performs well for group assignments that are the worst case according to some fairness criterion. Table 1 enumerates several of these criteria, and draw analogies to domain generalization methods that match or optimize similar statistics.[8] Environment inference serves a similar purpose for our method, but with a slightly different motivation: rather than learn an fair model in an online way that provides favorable in-distribution predictions, we learn discrete data partitions as an intermediary step, which enables use of invariant learning methods to tackle distribution shift.

Adversarially Reweighted Learning (ARL) (Lahoti et al., 2020) is most closely related to ours, since they emphasize subpopulation shift as a key motivation. Whereas ARL uses a DRO objective that prioritizes stability in the loss space, we explore environment inference to encourage invariance in the learned representation space. We see these as complementary approaches that are each suited to different types of distribution shift, as we discuss in the experiments.

[8]We refer the interested reader to Appendix C for a more in-depth discussion of the relationships between domain generalization and fairness methods.

5. Experiments

For lack of space we defer a proof-of-concept synthetic regression experiment to Appendix F.1. We proceed by describing the remaining datasets under study in Section 5.1. We then present the main results measuring the ability of EIIL to handle distribution shift in Section 5.2, and offer a more detailed analysis of the EIIL solution and its dependence on the reference model in Section 5.3. See `https://github.com/ecreager/eiil` for code.

Model selection Tuning hyperparameters when train and test distributions differ is a difficult open problem (Krueger et al., 2021; Gulrajani & Lopez-Paz, 2021). Where possible, we reuse effective hyperparameters for IRM and GroupDRO found by previous authors. Because these works allowed limited validation samples for hyperparameter tuning (all baseline methods benefit fairly from this strategy), these results represent an optimistic view on the ability for invariant learning. As discussed above, the choice of reference classifier is of crucial importance when deploying EIIL; if worst-group performance can be measured on a validation set, this could be used to tune the hyperparameters of the reference model (i.e. model selection subsumes reference model selection). See Appendix E for further discussion.

5.1. Datasets

CMNIST CMNIST is a noisy digit recognition task[9] where color is a spurious feature that correlates with the label at train time but anti-correlates at test time, with the correlation strength at train time varying across environments (Arjovsky et al., 2019). In particular, the two training environments have $Corr(color, label) \in \{0.8, 0.9\}$ while the test environment has $Corr(color, label) = 0.1$. Cru-

[9]MNIST digits are grouped into $\{0, 1, 2, 3, 4\}$ and $\{5, 6, 7, 8, 9\}$ so the CMNIST target label y is binary.

cially, label noise is applied by flipping y with probability $\theta_y = 0.25$. This implies that shape (the invariant feature) is marginally less reliable than color in the training data, so ERM ignores shape to focus on color and suffers from below-chance test performance.

Waterbirds To evaluate whether EIIL can infer useful environments in a more challenging setting with high-dimensional images, we turn to the Waterbirds dataset (Sagawa et al., 2020). Waterbirds is a composite dataset that combines $4,795$ bird images from the CUB dataset (Welinder et al., 2010) with background images from the Places dataset (Zhou et al., 2017). It examines the proposition (which frequently motivates invariant learning approaches) that modern networks often learn spurious background features (e.g. green grass in pictures of cows) that are predictive of the label at train time but fail to generalize in new contexts (Beery et al., 2018; Geirhos et al., 2020). The target labels are two classes of birds—"landbirds" and "waterbirds" respectively coming from dry or wet habitats—superimposed on land and water backgrounds. At training time, landbirds and waterbirds are most frequently paired with land and water backgrounds, respectively, but at test time the 4 subgroup combinations are uniformly sampled. To mitigate failure under distribution shift, a robust representation should learn primarily features of the bird itself, since these are invariant, rather than features of the background. Beyond the increase in dimensionality, this task differs from CMNIST in that the ERM solution does not fail catastrophically at test time, and in fact can achieve 97.3% average accuracy. However, because ERM optimizes average loss, it suffers in performance on the worst-case subgroup (waterbirds on land, which has only 56 training examples).

Adult-Confounded To assess the ability of EIIL to address worst-case group performance without group labels, we construct a variant of the UCI Adult dataset,[10] which comprises $48,842$ census records collected from the USA in 1994. The task commonly used as an algorithmic fairness benchmark is to predict a binarized income indicator (thresholded at $\$50,000$) as the target label, possibly considering sensitive attributes such as age, sex, and race.

Lahoti et al. (2020) demonstrate the benefit of per-example loss reweighting on Adult using their method ARL to improve predictive performance for undersampled subgroups. Following Lahoti et al. (2020), we consider the effect of four sensitive *subgroups*—defined by composing binarized race and sex labels—on model performance, assuming the model does not know a priori which features are sensitive. However, we focus on a distinct generalization problem where a pernicious dataset bias confounds the training data, making subgroup membership predictive of the label on the training data. At test time these correlations are reversed, so a predictor that infers subgroup membership to make predictions will perform poorly at test time (see Appendix D for details). This large test-time distribution shift can be understood as a controlled *audit* to determine if the classifier uses subgroup information to predict the target label. We call our dataset variant Adult-Confounded.

CivilComments-WILDS We apply EIIL to a large and challenging comment toxicity prediction task with important fairness implications (Borkan et al., 2019), where ERM performs poorly on comments associated with certain identity groups. We follow the procedure and data splits of Koh et al. (2021) to finetune DistilBERT embeddings (Sanh et al., 2019). EIIL uses an ERM reference classifier and its inferred environments are fed to a GroupDRO invariant learner. Because the large training set ($N_{train} = 269,038$) increases the convergence time for gradient-based EI, we deploy the binning heuristic discussed in Section 3.3, which in this instance finds environments that correspond to the error and non-error cases of the reference classifier. While ERM and EIIL do not have access to the sensitive group labels, we note that worst-group validation accuracy is used to tune hyperparameters for all methods. See Appendix E for details. We also compare against a GroupDRO (oracle) learner that has access to group labels.

5.2. Results

Method	Handcrafted Environments	Train	Test
ERM	✗	**86.3 ± 0.1**	13.8 ± 0.6
IRM	✓	71.1 ± 0.8	65.5 ± 2.3
EIIL	✗	73.7 ± 0.5	**68.4 ± 2.7**

Table 2. Accuracy (%) on CMNIST, a digit classification task where color is a spurious feature correlated with the label during training but anti-correlated at test time. EIIL exceeds test-time performance of IRM *without* knowledge of pre-specified environment labels, by instead finding worst-case environments using aggregated data and a reference classifier.

CMNIST IRM was previously shown to learn an invariant representation on this dataset, allowing it to generalize relatively well to the held-out test set whereas ERM fails dramatically (Arjovsky et al., 2019). It is worth noting that label noise makes the problem challenging, so even an oracle classifier can achieve at most 75% test accuracy on this binary classification task. To realize EIIL in our experiments, we discard the environment labels, and run the procedure described in Section 3.1 with ERM as the reference model and IRM as the invariant learner used in the final stage. We find that EIIL's environment labels are

[10] https://archive.ics.uci.edu/ml/datasets/adult

very effective for invariant learning, ultimately *outperforming* standard IRM using the environment labels provided in the dataset (Table 2). This suggests that in this case the EIIL solution approaches the *maximally informative* set of environments discussed in Proposition 1.

Waterbirds Sagawa et al. (2020) demonstrated that ERM suffers from poor worst-group performance on this dataset, and that GroupDRO can mitigate this performance gap if group labels are available. In this dataset, group labels should be considered as oracle information, meaning that the relevant baseline for EIIL is standard ERM. The main contribution of Sagawa et al. (2020) was to show *how* deep nets can be optimized for the GroupDRO objective using their online algorithm that adaptively adjusts per-group importance weights. In our experiment, we combine this insight with our EIIL framework to show that distributionally robust neural nets can be realized without access to oracle information. We follow the same basic procedure as above,[11] in this case using GroupDRO as the downstream invariant learner for which EIIL's inferred labels will be used.

Method	Train (avg)	Test (avg)	Test (worst group)
ERM	100.0	**97.3**	60.3
EIIL	99.6	96.9	**78.7**
GroupDRO (oracle)	99.1	96.6	84.6

Table 3. Accuracy (%) on the Waterbirds dataset. EIIL strongly outperforms ERM on worst-group performance, approaching the performance of the GroupDRO algorithm proposed by Sagawa et al. (2020), which requires oracle access to group labels. In this experiment we feed environments inferred by EIIL into a GroupDRO learner.

EIIL is significantly more robust than the ERM baseline (Table 3), raising worst-group test accuracy by 18% with only a 1% drop in average accuracy. In Figure 2 we plot the distribution of subgroups for each inferred environment, showing that the minority subgroups (landbirds on water and waterbirds on land) are mostly organized in the same inferred environment. This suggests the possibility of leveraging environment inference for interpretability to automatically discover a model's performance discrepancies on subgroups, which we leave for future work.

Adult-Confounded Using EIIL—to first infer worst-case environments then ensure invariance across them—performs favorably on the audit test set, compared with ARL and an ERM baseline (Table 4). We also find that, without access to sensitive group information, the solution found by EIIL achieves significantly better calibration on the test distribution (Figure 3). Because the train and test distributions differ based on the correlation pattern of small subgroups to the target label, this suggests that EIIL has achieved favorable group sufficiency (Liu et al., 2019) in this setting. See Appendix F.3 for a discussion of this point, as well as an ablation showing that all components of the EIIL approach are needed to achieve the best performance.

[11] For this dataset, environment inference worked better with reference models that were not fully trained. We suspect this is because ERM focuses on the easy-to-compute features like background color early in training, precisely the type of bias EIIL can exploit to learn informative environments.

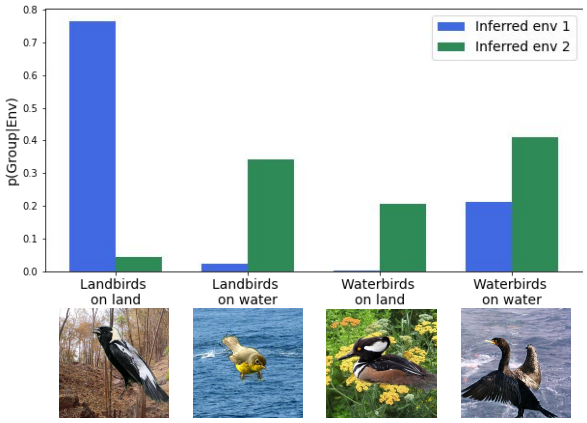


Figure 2. After using EIIL to directly infer two environments from the Waterbirds dataset, we examine the proportion of each subgroup (available in the original dataset but not used by EIIL) present in the inferred environment.

CivilComments-WILDS Without knowledge of which comments are associated with which groups, EIIL improves worst-group accuracy over ERM with only a modest cost in average accuracy, approaching the oracle GroupDRO solution (which requires group labels).

5.3. Influence of the reference model

As discussed in Section 3.2, the ability of EIIL to find useful environments—partitions yielding an invariant representation when used by an invariant learner—depends on its ability to exploit variation in the predictive distribution of a reference classifier. Here we study the influence of the reference classifier on the final EIIL solution. We return to

Method	Train accs	Test accs
ERM	**92.7 ± 0.5**	31.1 ± 4.4
ARL (Lahoti et al., 2020)	72.1 ± 3.6	61.3 ± 1.7
EIIL	69.7 ± 1.6	**78.8 ± 1.4**

Table 4. Accuracy on Adult-Confounded, a variant of the UCI Adult dataset where some sensitive subgroups correlate to the label at train time, and this correlation pattern is reversed at test time.

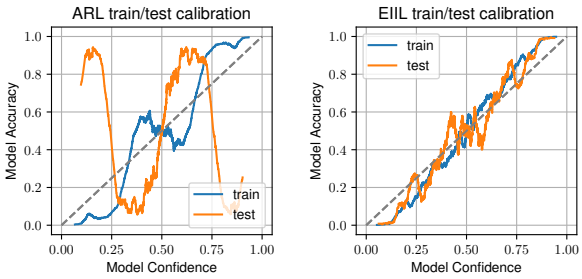
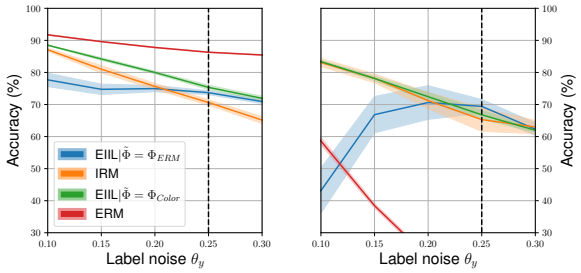


Figure 3. By inferring environments that maximally violate the invariance principle, and then applying invariant learning across the inferred environments, EIIL finds a solution that is well calibrated on the test set (right), compared with ARL (left).



(a) Train accuracy. (b) Test accuracy

Figure 4. CMNIST with varying label noise θ_y. Under high label noise ($\theta_y > .2$), where the spurious feature color correlates to label *more* than shape on the train data, EIIL matches or exceeds the test performance of IRM *without relying on hand-crafted environments*. Under medium label noise ($.1 < \theta_y < .2$), EIIL is worse than IRM but better than ERM, the logical approach if environments are not available. Under low label noise ($\theta_y < .1$), where color is *less* predictive than shape at train time, ERM performs well and EIIL fails. The vertical dashed black line indicates the default setting of $\theta_y = 0.25$, which we report in Table 2.

the CMNIST dataset, which provides a controlled sampling setup where particular ERM solutions can be induced to serve as reference for EIIL. In Appendix F.1, we discuss a similar experiment in a synthetic regression setting.

EIIL was shown to outperform IRM *without access to environment labels* in the standard CMNIST dataset (Table 2), which has label noise of $\theta_y = 0.25$. Because $Corr(color, label)$ is 0.85 (on average) for the train set, this amount of label noise implies that color is the most predictive feature on aggregated training set (although its predictive power varies across environments). ERM, even with access to infinite data, will focus on color given this amount of label noise to achieve an average train accuracy of 85%. However we can implicitly control the ERM solution Φ_{ERM} by tuning θ_y, an insight that we use to study the dependence of EIIL on the reference model $\tilde{\Phi} = \Phi_{ERM}$.

Figure 4 shows the results of our study. We find that EIIL generalizes better than IRM with sufficiently high label noise $\theta_y > .2$, but generalizes poorly under low label noise. This is precisely because ERM learns the color feature under high label noise, and the shape feature under low label noise. We verify this conclusion by evaluating EIIL when $\tilde{\Phi} = \Phi_{Color}$, i.e. a hand-coded color-based predictor as reference, which does well across all settings of θ_y.

We saw in the Waterbirds experiment that it is not a strict requirement that ERM fail completely in order for EIIL to succeed. However, this controlled study highlights the importance of the reference model in the ability of EIIL to find environments that emphasize the right invariances, which leaves open the question of how to effectively choose a reference model for EIIL in general. One possible way forward is by using validation data that captures the *kind* of distribution shift we expect at test time, without exactly producing the test distribution, e.g. as in the WILDS benchmark (Koh et al., 2021). In this case we could choose to run EIIL with a reference model that exhibits a large generalization gap between the training and validation distributions.

Method	Train (avg)	Test (avg)	Test (worst group)
ERM	96.0 ± 1.5	$\mathbf{92.0 \pm 0.4}$	61.6 ± 1.3
EIIL	97.0 ± 0.8	90.5 ± 0.2	$\mathbf{67.0 \pm 2.4}$
GroupDRO (oracle)	93.6 ± 1.3	89.0 ± 0.3	69.8 ± 2.4

Table 5. EIIL improves worst-group accuracy in the CivilComments-WILDS toxicity prediction task, without access to group labels.

6. Conclusion

We introduced EIIL, a new method that infers environment partitions of aggregated training data for invariant learning. Without access to environment labels, EIIL can outperform or approach invariant learning methods that require environment labels. EIIL has implications for domain generalization and fairness alike, because in both cases it can be hard to specify meaningful environments or sensitive subgroups.

Acknowledgements

We are grateful to David Madras, Robert Adragna, Silviu Pitis, Will Grathwohl, Jesse Bettencourt, and Eleni Triantafillou for their feedback on this manuscript. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (www.vectorinstitute.ai/partners).

References

Ahmed, F., Bengio, Y., van Seijen, H., and Courville, A. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

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

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. Systematic generalization: what is required and can it be learned? In *International Conference of Machine Learning*, 2019.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pp. 456–473, 2018.

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

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

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. 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.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. *Communications of the ACM*, 63(5): 82–89, 2020.

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., and Love, B. C. A too-good-to-be-true prior to reduce shortcut reliance. *arXiv preprint arXiv:2102.06406*, 2021.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Edwards, H. and Storkey, A. Censoring representations with an adversary. In *International Conference for Machine Learning*, 2016.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Steinhardt, J., and Madry, A. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, 2020.

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

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency*, pp. 501–512, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1989–1998, 2018.

Kamath, P., Tangella, A., Sutherland, D. J., and Srebro, N. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, 2021.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, 2021.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. Fairness without demographics through adversarially reweighted learning. In *Neural Information Processing Systems*, 2020.

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

Liu, E., Haghgoo, B., Chen, A., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, 2019.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. In *International Conference on Learning Representations*, 2016.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *Neural Information Processing Systems 2020*, 2020.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Rockafellar, R. T. and Uryasev, S. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, 2019.

Sohoni, N. S., Dunnmon, J. A., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Neural Information Processing Systems 2020*, 2020.

Srivastava, M., Hashimoto, T., and Liang, P. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pp. 9109–9119. PMLR, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

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

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Williamson, R. C. and Menon, A. K. Fairness risk measures. In *International Conference on Machine Learning*, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhang, Y., Barzilay, R., and Jaakkola, T. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 5:515–528, 2017.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Algorithm 1 Pseudocode for environment inference (EI) with the invariance principle (realized via relaxed IRMv1 penalty) as the EI objective.

Input: Reference model Φ, dataset $\mathcal{D} = \{x_i, y_i\}$, loss ℓ, duration N_{steps}
Output: Worst case data splits $\mathcal{D}_1$, $\mathcal{D}_2$ for use with an invariant learner.

def $\tilde{R}^e(\Phi, \mathbf{q})$:
  return $\frac{1}{\sum_{i'} \mathbf{q}_{i'}(e)} \sum_i \mathbf{q}_i(e)\ell(\Phi(x_i), y_i)$ {Equation 4}

Randomly init. $\mathbf{q} \in [0,1]^N$ environment posterior ($\mathbf{q}_i(e) := q(e|x_i, y_i)$)
Randomly init. $\mathbf{q} \in [0,1]^N$ environment posterior ($\mathbf{q}_i(e) := q(e|x_i, y_i)$)
for $n \in 1 \dots N_{steps}$ **do**
  $SoftVari = \sum_{e \in \{1,2\}} ||\nabla_{\bar{w}} \tilde{R}^e(\bar{w} \circ \Phi, \mathbf{q})||$ {Aggregate reference model variances across soft envs}
  $Loss = -1 \cdot SoftVari$ {Maximize the EI objective by minimizing this loss}
  $\mathbf{q} \leftarrow OptimUpdate(\mathbf{q}, \nabla_{\mathbf{q}} Loss)$
end for
$\hat{\mathbf{q}} \sim Bernoulli(\mathbf{q})$ {sample splits}
$\mathcal{D}_1 \leftarrow \{x_i, y_i | \hat{\mathbf{q}}_i = 1\}$, $\mathcal{D}_2 \leftarrow \{x_i, y_i | \hat{\mathbf{q}}_i = 0\}$ {split data}
return $\mathcal{D}_1, \mathcal{D}_2$

A. Environment Inference Psuedocode

Algorithm 1 provides pseudocode for the environment inference procedure used in our experiments.

B. Proofs

B.1. Proof of Proposition 1

Consider a dataset with some feature(s) z which are spurious, and other(s) v which are valuable/causal w.r.t. the label y. This includes data generated by models where $v \to y \to z$, such that $P(y|v, z) = P(y|v)$. Assume further that the observations x are functions of both spurious and valuable features: $x := f(v, z)$. The aim of invariant learning is to form a classifier that predicts y from x that focuses solely on the causal features, i.e., is invariant to z and focuses solely on v.

Consider a classifier that produces a score $S(x)$ for example x. In the binary classification setting S is analogous to the model Φ, while the score $S(x)$ is analogous to the representation $\Phi(x)$. To quantify the degree to which the constraint in the Invariant Principle (EIC) holds, we introduce a measure called the *group sufficiency gap*[12]:

$$\Delta(S, e) = \mathbb{E}[\mathbb{E}[(y|S(x), e_1)] - \mathbb{E}[(y|S(x), e_2)]]$$

Now consider the notion of an environment: some setting in which the $x \to y$ relationship varies (based on spurious features). Assume a single binary spurious feature z. We restate Proposition 1 as follows:

Claim: If environments are defined based on the agreement of the spurious feature z and the label y, then a classifier that predicts based on z alone maximizes the group-sufficiency gap (and vice versa – if a classifier predicts y directly by predicting z, then defining two environments based on agreement of label and spurious feature—$e_1 = \{v, z, y | \mathbb{1}(y = z)\}$ and $e_2 = \{v, z, y | \mathbb{1}(y \neq z)\}$—maximizes the gap).

We can show this by first noting that if the environment is based on spurious feature-label agreement, then with $e \in \{0, 1\}$ we have $e = \mathbb{1}(y = z)$. If the classifier predicts z, i.e. $S(x) = z$, then we have

$$\Delta(S, e) = \mathbb{E}[\mathbb{E}[y|z(x), \mathbb{1}(y = z)] - \mathbb{E}[y|z(x), \mathbb{1}(y \neq z)]]$$

For each instance of x either $z = 0$ or $z = 1$. Now we note that when $z = 1$ we have $\mathbb{E}(y|z, \mathbb{1}(y = z)) = 1$ and $\mathbb{E}(y|z, \mathbb{1}(y \neq z)) = 0$, while when $z = 0$ $E(y|z, \mathbb{1}(y = z)]) = 0$ and $\mathbb{E}[y|z, \mathbb{1}(y \neq z)] = 1$. Therefore for each example $|E(y|z(x), \mathbb{1}(y = z)) - E(y|z(x), \mathbb{1}(y \neq z)| = 1$, contributing to an overall $\Delta(S, e) = 1$, which is the maximum value for the sufficiency gap.

[12] This was previously used in a fairness setting by Liu et al. (2019) to measure differing calibration curves across groups.

B.2. Heuristic for soft environment assignment based on binning violates the invariance principle

Here we analyze the heuristic discussed in Section 3.3. We want to show that finding environment assignments in this way both maximizes the violation of the softened version of the regularizer (Equation 3), and also also maximally violates the invariance principle (EIC).

Because the invariance principle $\mathbb{E}[Y|\Phi(X), e] = \mathbb{E}[Y|\Phi(X), e']\forall e, e'$ is difficult to quantify for continuous $\Phi(X)$, we consider a binned version of the representation, with b denoting the discrete index of the bin in representation space. Let $q_i \in [0, 1]$ denote the soft assignment of example i to environment 1, and $1 - q_i$ denote its converse, the assignment of example i to environment 2. Denote by $y_i \in \{0, 1\}$ the binary target for example i, and $\hat{y} \in [0, 1]$ as the model prediction on this example. Assume that ℓ represents a cross entropy or squared error loss so that $\nabla_w \ell(\hat{y}, y) = (\hat{y} - y)\Phi(x)$.

Consider the IRMv1 regularizer with soft assignment, expressed as

$$\begin{aligned}
D(q) &= \sum_e ||\nabla_{w|w=1.0} \frac{1}{N_e} \sum_i q_i(e)\ell(w \circ \Phi(x_i), y_i)||^2 \\
&= \sum_e ||\frac{1}{N_e} \sum_i q_i(e)(\hat{y}_i - y_i)\Phi(x_i)||^2 \\
&= ||\frac{1}{\sum_i' q_i'} \sum_i q_i(\hat{y}_i - y_i)\Phi(x_i)||^2 + ||\frac{1}{\sum_i' 1 - q_i'} \sum_i (1 - q_i)(\hat{y}_i - y_i)\Phi(x_i)||^2 \\
&= ||\frac{\sum_i q_i \hat{y}_i \Phi(x_i)}{\sum_{i'} q_{i'}} - \frac{\sum_i q_i y_i \Phi(x_i)}{\sum_{i'} q_{i'}}||^2 + ||\frac{\sum_i (1 - q_i)\hat{y}_i \Phi(x_i)}{\sum_{i'} 1 - q_{i'}} - \frac{\sum_i (1 - q_i) y_i \Phi(x_i)}{\sum_{i'} 1 - q_{i'}}||^2. \qquad (5)
\end{aligned}$$

Now consider that the space of $\Phi(X)$ is discretized into disjoint bins b over its support, using $z_{i,b} \in \{0, 1\}$ to indicate whether example i falls into bin b according to its mapping $\Phi(x_i)$. Thus we have

$$\begin{aligned}
D(q) = \sum_b (&||\frac{\sum_i z_{i,b} q_i \hat{y}_i \Phi(x_i)}{\sum_{i'} z_{i,b} q_{i'}} - \frac{\sum_i z_{i,b} q_i y_i \Phi(x_i)}{\sum_{i'} z_{i,b} q_{i'}}||^2 \\
&+ ||\frac{\sum_i z_{i,b}(1 - q_i)\hat{y}_i \Phi(x_i)}{\sum_{i'} z_{i,b}(1 - q_{i'})} - \frac{\sum_i z_{i,b}(1 - q_i) y_i \Phi(x_i)}{\sum_{i'} z_{i,b}(1 - q_{i'})}||^2) \qquad (6)
\end{aligned}$$

The important point is that within a bin, all examples have roughly the same $\Phi(x_i)$ value, and the same value for $\hat{y}_i$ as well. So denoting $K_b^{(1)} := \frac{\sum_i z_{i,b} q_i \hat{y}_i \Phi(x_i)}{\sum_{i'} z_{i,b} q_{i'}}$ and $K_b^{(2)} := \frac{\sum_i z_{i,b}(1-q_i)\hat{y}_i \Phi(x_i)}{\sum_{i'} z_{i,b}(1-q_{i'})}$ as the relevant constant within-bin summations, we have the following objective to be maximized by EIIL:

$$D(q) = \sum_b (||K_b^{(1)} - \frac{\sum_i z_{i,b} q_i y_i \Phi(x_i)}{\sum_{i'} z_{i,b} q_{i'}}||^2 + ||K_b^{(2)} - \frac{\sum_i z_{i,b}(1 - q_i) y_i \Phi(x_i)}{\sum_{i'} z_{i,b}(1 - q_{i'})}||^2.$$

One way to maximize this is to assign all $y_i = 1$ values to environment 1 ($q_i = 1$ for these examples) and all $y_i = 0$ to the other environment ($q_i = 0$). We can show this is maximized by considering all of the examples except the i-th one have been assigned this way, and then that the loss is maximized by assigning the i-th example according to this rule.

Now we want to show that the same assignment maximally violates the invariance principle (showing that this soft EIIL solution provides maximal non-invariance). Intuitively within each bin the difference between $\mathbb{E}[y|e = 1]$ and $\mathbb{E}[y|e = 2]$ is maximized (within the bin) if one of these expected label distributions is 1 while the other is 0. This can be achieved by assigning all the $y_i = 1$ values to the first environment and the $y_i = 0$ values to the second.

Thus a global optimum for the relaxed version of EIIL (using the IRMv1 regularizer) also maximally violates the invariance principle.

B.3. Given CMNIST environments are suboptimal w.r.t. sufficiency gap

The regularizer from IRMv1 encourages a representation for which sufficiency gap is minimized between the available environments. Therefore when faced with a new task it is natural to measure the natural sufficiency gap between these

environments, mediated through a naive or baseline method. Here we show that for CMNIST, when considering a naive color-based classifier as the reference model, the given environment splits are actually *suboptimal* w.r.t. sufficiency gap, which motivates the inference of environments via EIIL that have a higher sufficiency gap for the reference model.

We begin by computing $\Delta(S, e)$, the sufficiency gap for color-based classifier g over the given train environments $\{e_1, e_2\}$. We introduce an auxiliary color variable z, which is not observed but can be sampled from via the color based classifier g:

$$p(y|g(x) = x', e) = \mathbb{E}_{p(z|x')}\left[p(y|z, e, x').\right]$$

Denote by GREEN and RED the set of green and red images, respectively. I.e. we have $z \in G$ iff $z = 1$ and $x \in$ GREEN iff $z(x) = 1$. The the sufficiency gap is expressed as

$$\begin{aligned}
\Delta(S, e) &= \mathbb{E}_{p(x,e)}\left[\left|\mathbb{E}_{p(y|x,e_1)}[y|g(x), e_1] - \mathbb{E}_{p(y|x,e_2)}[y|g(x), e_2]\right|\right] \\
&= \mathbb{E}_{p(z,e)}\left[\left|\mathbb{E}_{p(y|z,e_1)}[y|z, e_1] - \mathbb{E}_{p(y|z,e_2)}[y|z, e_2]\right|\right] \\
&= \frac{1}{2}\sum_{z\in\{\text{GREEN},\text{RED}\}}\left[\left|\mathbb{E}_{p(y|z,e_1)}[y|z, e_1] - \mathbb{E}_{p(y|z,e_2)}[y|z, e_2]\right|\right] \\
&= \frac{1}{2}(|\mathbb{E}[y|z = \text{GREEN}, e_1] - \mathbb{E}[y|z = \text{GREEN}, e_2]| + |\mathbb{E}[y|z = \text{RED}, e_1] - \mathbb{E}[y|z = \text{RED}, e_2]|) \\
&= \frac{1}{2}(|0.1 - 0.2| - |0.9 - 0.8|) = \frac{1}{10}.
\end{aligned}$$

The regularizer in IRMv1 is trying to reduce the sufficiency gap, so in some sense we can think about this gap as a learning signal for the IRM learner. A natural question would be whether a different set of environment partition $\{e\}$ can be found such that this learning signal is stronger, i.e. the sufficiency gap is increased. We find the answer is yes. Consider an environment distribution $q(e|x, y, z)$ that assigns each data point to one of two environments. Any assignment suffices so far as its marginal matches the observed data: $\int_z \int_e q(x, y, z, e) = p^{\text{obs}}(x, y)$.

We can now express the sufficiency gap (given a color-based classifier g) as a function of the environment assignment q:

$$\begin{aligned}
\Delta(S, e \sim q) &= \mathbb{E}_{q(x,e)}[|\mathbb{E}_{q(y|x,e,x)}[y|g(x), e_1] - \mathbb{E}_{q(y|x,e,x)}[y|g(x), e_2]|] \\
&= \mathbb{E}_{q(x,e)}[|\mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z, e_1] - \mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z, e_2]|]
\end{aligned}$$

Where we use the same change of variables trick as above to replace $g(x)$ with samples from $p(z|x)$ (note that this is the color factor from the generative process p according with our assumption that g matches this distribution).

We want to show that there exists a q yielding a higher sufficiency gap than the given environments. Consider q that yields the conditional label distribution

$$q(y|x, e, z) := q(y|e, z) = \begin{cases} \mathbb{1}(y = z) \text{ if } e = e_1, \\ \mathbb{1}(y \neq z) \text{ if } e = e_2. \end{cases}$$

This can be realized by an encoder/auditor $q(e|x, y, z)$ that ignores image features in x and partitions the example based on whether or not the label y and color z agree. We also note that z is deterministically the color of the image in the generative process: $p(z|x) = \mathbb{1}(x = \text{RED})$

Now we can compute the sufficiency gap:

$$
\begin{aligned}
\Delta(S, e \sim q) &= \mathbb{E}_{q(x,e)}[|\mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_1] - \mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_2]|] \\
&= \frac{1}{2}\mathbb{E}_{x\in\textsc{Red}}|\mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_1] - \mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_2]| \\
&\quad + \frac{1}{2}\mathbb{E}_{x\in\textsc{Green}}|\mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_1] - \mathbb{E}_{q(y|z,e,x)p(z|x)}[y|z,e_2]| \\
&= \frac{1}{2}\mathbb{E}_{x\in\textsc{Red}}(|\sum_y\sum_z(y * \mathbb{1}(y=z) * \mathbb{1}(g(x)=z)) - \sum_y\sum_z(y * \mathbb{1}(y\neq z) * \mathbb{1}(g(x)=z))|) \\
&\quad + \mathbb{E}_{x\in\textsc{Green}}\frac{1}{2}(|\sum_y\sum_z(y * \mathbb{1}(y=z) * \mathbb{1}(g(x)=z)) - \sum_y\sum_z(y * \mathbb{1}(y\neq z) * \mathbb{1}(g(x)=z))|) \\
&= \frac{1}{2}\mathbb{E}_{x\in\textsc{Red}}(|\sum_y(y * \mathbb{1}(y=1) * \mathbb{1}(x\in\textsc{Red})) - \sum_y(y * \mathbb{1}(y\neq 1) * \mathbb{1}(x\in\textsc{Red}))|) \\
&\quad + \mathbb{E}_{x\in\textsc{Green}}\frac{1}{2}(|\sum_y\sum_z(y * \mathbb{1}(y=0) * \mathbb{1}(x\in\textsc{Green})) - \sum_y\sum_z(y * \mathbb{1}(y\neq 0) * \mathbb{1}(x\in\textsc{Green}))|) \\
&= \frac{1}{2}\mathbb{E}_{x\in\textsc{Red}}[|1-0|] + \mathbb{E}_{x\in\textsc{Green}}[\frac{1}{2}|0-1|] = \frac{1}{2} + \frac{1}{2} = 1.
\end{aligned}
$$

Note that 1 is the maximal sufficiency gap, meaning that the described environment partition maximizes the sufficiency gap w.r.t. the color-based classifier g.

C. Connections Between Invariant Learning and Algorithmic Fairness

Here we lay out some connections to algorithmic fairness, where demographic information, which is often considered "sensitive", is used to inform learning. Table 1 from the main paper provides a high-level comparison of the objectives and assumptions of several relevant methods. Loosely speaking, recent approaches from both areas share the goal of matching some chosen statistic across a conditioning variable e, representing sensitive group membership in algorithmic fairness or an environment/domain indicator in domain generalization. The statistic in question informs the *learning objective* for the resulting model, and is motivated differently in each case. In domain generalization, learning is informed by the properties of the test distribution where good generalization should be achieved. In algorithmic fairness the choice of statistic is motivated by a context-specific *fairness notion*, that likewise encourages a particular solution that achieves "fair" outcomes (Chouldechova & Roth, 2020).

Early approaches to learning fair representations (Zemel et al., 2013; Edwards & Storkey, 2016; Louizos et al., 2016; Zhang et al., 2018; Madras et al., 2018) leveraged statistical independence regularizers from domain adaptation[13] (Ben-David et al., 2010; Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018), noting that marginal or conditional independence from domain to prediction relates to the fairness notions of demographic parity $\hat{y} \perp e$ (Dwork et al., 2012) and equal opportunity $\hat{y} \perp e|y$ (Hardt et al., 2016).

Recall that (EIC) involves an environment-specific conditional label expectation given a data representation $\mathbb{E}[y|\Phi(x) = h, e]$. Objects of this type have been closely studied in the fair machine learning literature, where e now denotes a "sensitive" attribute indicating membership in a protected demographic group (age, race, gender, etc.), and the vector representation $\Phi(x)$ is typically replaced by a scalar score[14] $S(x) \in \mathbb{R}$. Noting that $\sigma(S(x))$ represents the probability of the model prediction, $\mathbb{E}[y|S(x), e]$ can now be interpreted as a *calibration curve* that must be regulated according to some fairness constraint. Chouldechova (2017) showed that equalizing this calibration curve across groups is often incompatible with a common fairness constraint, demographic parity, while Liu et al. (2019) studied "group sufficiency" of classifiers with strongly convex losses, concluding that ERM naturally finds group sufficient solutions without fairness constraints.

Because Liu et al. (2019) consider convex losses, their theoretical results do not hold for neural network representations. However, by noting the link between group sufficiency and the constraint from (EIC), we observe that the IRMv1 regularizer (applicable to neural nets) in fact minimizes the group sufficiency gap in the case of a scalar representation $\Phi(x) \subseteq \mathbb{R}$, and when e indicates sensitive group membership. It is worth noting that Arjovsky et al. (2019) briefly discuss using groups as

[13] Whereas domain generalization requires model predictions on entirely novel domains at test time, domain adaptation assumes a set of target domain examples are available at test time to guide model adaptation.

[14] For binary classification, score-based and representation-based approaches are closely related since scores are commonly implemented as (or can be interpreted as) as the linear mapping of a data representation: $S(x) = w \circ \Phi(x)$.

environments, but without specifying a particular fairness criterion. We leave an empirical study of these methods for future work.

Our approach in searching for worst-case data partitions in EIIL was inspired by recent work on fair prediction without sensitive labels (Kearns et al., 2018; Hébert-Johnson et al., 2018; Hashimoto et al., 2018; Lahoti et al., 2020). Reliance on sensitive demographic information is cumbersome since it often cannot be collected without legal or ethical repercussions. Hébert-Johnson et al. (2018) discussed the problem of mitigating subgroup unfairness when group labels are unknown, and proposed *Multicalibration* as a way of ensuring a classifier's calibration curve is invariant to efficiently computable environment splits. Since the proposed algorithm requires brute force enumeration over all possible environments/groups, Kim et al. (2019) suggested a more practical algorithm by relaxing the calibration constraint to an accuracy constraint, yielding a *Multiaccurate* classifier.[15] The goal here is to boost the predictions of a pre-trained classifier through multiple rounds of auditing (searching for worst-case subgroups using an auxiliary model) rather than learning an invariant representation.

A related line of work also leverages inferred subgroup information to improve worst-case model performance using the framework of DRO. Hashimoto et al. (2018) applied DRO to encourage long-term fairness in a dynamical setting where the average loss for a subpopulation influences their propensity to continue engaging with the model. Lahoti et al. (2020) proposed Adversarially Reweighted Learning (ARL), which extends DRO using an auxiliary model to compute the importance weights γ_i mentioned above. Amortizing this computation mitigates the tendency of DRO to overfit its reweighting strategy to noisy outliers.

Limitations of generalization-first fairness One exciting direction for future work is to apply methods developed in the domain generalization literature to tasks where distribution shift is related to some societal harm that should be mitigated. However, researchers should be wary of blind "solutionism", which can be ineffectual or harmful when the societal context surrounding the machine learning system is ignored (Selbst et al., 2019). Moreover, many aspects of algorithmic discrimination are not simply a matter of achieving few errors on unseen distributions. Unfairness due to task definition or dataset collection, as discussed in the study of target variable selection by Obermeyer et al. (2019), may not be reversible by novel algorithmic developments.

D. Dataset details

CMNIST This dataset was provided by Arjovsky et al. (2019)[16]. The two training environments comprise $25,000$ images each, with $Corr(color, label) = 0.8$ for the firs training environment and $Corr(color, label) = 0.8$ for the second. A held-out test set with $Corr(color, label) = 0.1$ is used for evaluation. Label noise is applied by flipping the binary target y with probability $\theta_y = 0.25$, with color correlation applied w.r.t. the noisy label. Given that only two color channels are used, we follow Arjovsky et al. (2019) in downsampling the digit images to 14×14 pixels and 2 channels.

Waterbirds We follow the procedure outlined by Sagawa et al. (2020) to reproduce the Waterbirds dataset. As noted by the authors, due to random seed differences our version of the dataset may differ slightly from the one originally used by the paper. The train/validation/test splits are of size $4,795/1,200/5,794$. As noted in the Appendix of (Sagawa et al., 2020), the validation and test distributions represent upweight the minority groups so that the number of examples coming from each habitat is equal (although there are still marginally more landbirds than waterbirds). For example on train set the subgroup sizes are $3,498/184/56/1,057$ while on the test set the sizes are $467/466/133/133$.

CivilComments-WILDS We use the train/validation/test splits from Koh et al. (2021); we refer the interested reader the Appendix of their paper for a detailed description of this version of the dataset, including how it differs from the original dataset (Borkan et al., 2019).

Constructing the Adult-Confounded dataset To create our semi-synthetic dataset, called Adult-Confounded, we start by observing that the conditional distribution over labels varies across the subgroups, and in some cases subgroup membership is very predictive of the target label. We construct a test set (a.k.a. the audit set) where this relationship between subgroups and target label is reversed.

The four sensitive subgroups are defined following the procedure of Lahoti et al. (2020), with sex (recorded as binary:

[15] Kearns et al. (2018) also proposed a boosting procedure to equalize subgroup errors without sensitive attributes.

[16] https://github.com/facebookresearch/InvariantRiskMinimization

Male/Female) and binarized race (Black/non-Black) attributes compose to make four possible subgroups: Non-Black Males (SG1), Non-Black Females (S2), Black Males (SG3), and Black Females (SG4).

We start with the observation that each subgroup has a different correlation strength with the target label, and in some cases subgroup membership alone can be used to achieve relatively low error rates in prediction. As these correlations should be considered "spurious" to mitigate unequal treatment across groups, we create a semi-synthetic variant of the UCI Adult dataset, which we call Adult-Confounded, where these spurious correlations are exaggerated. Table 6 shows various conditional label distributions for the original dataset and our proposed variant. The test set for Adult-Confounded reverses the correlation strengths, which can be thought of as a worst-case audit to ensure the model is not relying on subgroup membership alone in its predictions. We generate samples for Adult-Confounded using importance sampling, keeping the original train/test splits from UCI Adult as well as the subgroup sizes, but sampling individual examples under/over-sampled according to importance weights $\frac{p^{Adult-Confounded}}{p^{UCIAdult}}$.

Subgroup (SG)	$p(y=1\|SG)$			
	UCIAdult		Adult-Confounded	
	Train	Test	Train	Test
1	0.31	0.30	0.94	0.06
2	0.11	0.12	0.06	0.94
3	0.19	0.16	0.94	0.06
4	0.06	0.04	0.06	0.94

Table 6. Adult-Confounded is a variant of the UCI Adult dataset that emphasizes test-time distribution shift.

E. Experimental details

Model selection Krueger et al. (2021) discussed the pitfalls of achieving good test performance on CMNIST by using test data to tune hyperparameters. Because our primary interest is in the properties of the inferred environment rather than the final test performance, we sidestep this issue in the Synthetic Regression and CMNIST experiments by using the default parameters of IRM without further tuning. However for Adult-Confounded a specific strategy for model selection is needed.

We refer the interested reader to Gulrajani & Lopez-Paz (2021) for an extensive discussion of possible model selection strategies. They also provide a large empirical study showing that ERM is difficult baseline to beat when all methods are put on equal footing w.r.t. model selection.

In our case, we use the most relaxed model selection method proposed by Gulrajani & Lopez-Paz (2021), which amounts to allowing each method a 20 test evaluations using hyperparameter chosen at random from a reasonable range, with the best hyperparameter setting selected for each method. While none of the methods is given an unfair advantage in the search over hyperparameters, the basic model selection premise does not translate to real-world applications, since information about the test-time distribution is required to select hyperparameters. Thus these results can be understood as being overly optimistic for each method, although the relative ordering between the methods can still be compared.

Training times Because EIIL requires a pre-trained reference model and optimization of the EI objective, overall training time is longer than standard invariant learning. It depends primarily on the number of steps used to train the reference model and number of steps used in EI optimization. The extra training time incurred is manageable and varies from dataset to dataset.

In CMNIST, we train the ERM reference model for $1,000$ steps, which is the same duration as the downstream invariant learner that eventually uses the inferred environments. In this setting the $10,000$ steps required to optimize the EI objective is actually more than used for representation learning. The overall EIIL train time is 6.6 minutes to run 10 restarts on a NVIDIA Tesla P100, compared with 2.18 minutes for ERM and 2.20 minutes for IRM.

However, as the problem size scales, the relative overhead cost of EIIL becomes progressively discounted. On Waterbirds, training GroupDRO takes 4.716 hours on a NVIDIA Tesla P100. Our reference model trains for 1 epoch, so taking this into account along with the $20,000$ steps of EI optimization, EIIL runs at 4.737 hours. This is a relative increase of 0.4%.

Batch environment inference As mentioned in the main paper, we aggregate logits for the entire training set and optimize the EI objective using the entire training batches. This can be done by cycling through the train set once in minibatches, computing logits per minibatch, and aggregating the logits only (discarding network activations) prior to EI. We leave minibatched environment inference and amortization of the soft environment assignments to future work.

Experimental infrastructure Our experiments were run on a cluster of NVIDIA Tesla P100 machines.

CMNIST IRM is trained on the two training environments and tested on a holdout environment constructed from $10,000$ test images in the same way as the training environments, where colour is predictive of the noisy label 10% of the time. So using color as a feature to predict the label will lead to an accuracy of roughly 10% on the test environment, while it yields 80% and 90% accuracy respectively on the training environments.

To evaluate EIIL we remove the environment identifier from the training set and thus have one training set comprised of $50,000$ images from both original training environments. We then train an MLP with binary cross-entropy loss on the training environments, freeze its weights and use the obtained model to learn environment splits that maximally violate the IRM penalty. When optimizing the inner loop of EIIL, we use Adam with learning rate 0.001 for $10,000$ steps with full data batches used to computed gradients.

The obtained environment partitions are then used to train a new model from scratch with IRM. Following Arjovsky et al. (2019), we allow the representation to train for several hundred annealing steps before applying the IRMv1 penalty.

We used the default architecture—an MLP with two hidden layers of 390 neurons—and hyperparameter values[17]—-learning rate, weight decay, and penalty strength—from (Arjovsky et al., 2019). We do not use minibatches as the entire dataset fits into memory.

Waterbirds Following Sagawa et al. (2020), we use the default `torchvision` ResNet50 models, using the pre-trained weights as the initial model parameters, and train without any data augmentation using the For GroupDRO and ERM, we use hyperparameters reported by the authors[18], and note that the authors make use of the the validation set (whose distribution contains less group imbalance than the training data), to select hyperparameters in their experiments (all methods benefit equally from this strategy). We train for 300 epochs without any early stopping (to avoid any further influence from the validation data). For EIIL, we optimize the EI objective of EIIL with learning rate 0.01 for $20,000$ steps using the Adam optimizer, and use GroupDRO (using the same hyperparameters as the GroupDRO baseline) as the invariant learner. An ERM model trained for 1 epoch was used as the reference model. We also tried using reference modeled trained for longer, but found that EIIL did not perform as well in this case. We hypothesize that this is because the reference ERM model focuses on background features early in training, leading to stark performance discrepancies across subgroups, which in turn provides a strong learning signal for EIIL to infer effective environments. While subgroup disparities are present for more well-trained models, the learning signal in the EI phase will weaken.

Adult-Confounded Following Lahoti et al. (2020), we use a two-hidden-layer MLP architecture for all methods, with 64 and 32 hidden units respectively, and a linear adversary for ARL. We use IRM as the invariant learner in the final stage of EIIL. We optimize all methods using Adagrad; learning rates, number of steps, and batch sizes chosen by the model selection strategy described above (with 20 test evaluations per method), as are penalty weights for IRMv1 regularizer and standard weight decay. For the inner loop of EIIL (inferring the environments), we use the same settings as in CMNIST. We find that the performance of EIIL is somewhat sensitive to the number of steps taken with the IRMv1 penalty applied. To limit the number of test queries needed during model selection, we use an early stopping heuristic by enforcing the IRMv1 penalty only during the final 500 steps of training, with the previous steps serving as annealing period to learn a baseline representation to be regularized.

CivilComments-WILDS Following (Koh et al., 2021), we finetune DistilBERT embeddings (Sanh et al., 2019) using the default HuggingFace implementation and default weights (Wolf et al., 2019). EIIL uses an ERM reference classifier and its inferred environments are fed to a GroupDRO invariant learner. During prototyping the EI step, we noticed that the binning heuristic described in Section 3.3 consistently split the training examples into environments according to the error cases of

[17] https://github.com/facebookresearch/InvariantRiskMinimization/blob/master/code/colored_mnist/reproduce_paper_results.sh

[18] https://worksheets.codalab.org/worksheets/0x621811fe446b49bb818293bae2ef88c0

the reference classifier. Because error splitting is even simpler to implement than confidence binning, we used this heuristic for the EI step; we believe this is a promising approach for scaling EI to large datasets, and note its equivalence to the first stage of the method independently proposed by Liu et al. (2021), which is published concurrently to ours. We experimented with gradient-based EI on this dataset, but did not find any improvement over the (faster) heuristic EI.

On this dataset, we treat reference model selection as part of the overall model selection process, meaning that the hyperparameters of the ERM reference model are treated as a subset of the overall hyperparameters tuned during model selection. Specifically we used a grid search to tune the reference model learning rate (1e-5, 1e-4), optimizer type (Adam, SGD) and scheduler (linear, plateau), and gradient norm clamping (off, clamped at 1.0), as well as the invariant learner (GroupDRO) learning rate (1e-5, 1e-4). Moreover, we allow all methods to evaluate worst-group validation accuracy to tune these hyperparameters; such validation data will not be available in most settings, so this result can be seen as an optimistic view of the performance of all methods, including EIIL. We train all methods (including the reference model) for 5 epochs, with the best epoch chosen according to validation performance. Interestingly, the reference model chosen in this way was a constant classifier, so the overall EIIL solution is equivalent to GroupDRO using the class label as the environment label.

The oracle GroupDRO method trains on two environments, with one containing comments where *any* of the 8 sensitive groups was mentioned, and other environment containing the remaining comments. We experimented with allowing the oracle method access to more fine-grained environment labels by evaluating all 2^8 combinations of binary group labels, but did not find any significant performance boost (consistent with observations from Koh et al. (2021)).

F. Additional Empirical Results

F.1. Synthetic Data

	Causal MSE	Noncausal MSE
ERM	0.827 ± 0.185	0.824 ± 0.013
ICP	1.000 ± 0.000	0.756 ± 0.378
IRM	0.666 ± 0.073	0.644 ± 0.061
EIIL	$\mathbf{0.148 \pm 0.185}$	$\mathbf{0.145 \pm 0.177}$

Table 7. IRM using EIIL-discovered environments (e_{EIIL}) outperforms IRM in a synthetic regression setting without the need for hand-crafted environments (e_{HC}). This is because the reference representation $\tilde{\Phi} = \Phi_{\text{ERM}}$ uses the spurious feature for prediction. MSE + standard deviation across 5 runs reported.

We begin with a regression setting originally used as a toy dataset for evaluating IRM (Arjovsky et al., 2019). The features $\mathbf{x} \in \mathbb{R}^N$ comprise a "causal" feature $\mathbf{v} \in \mathbb{R}^{N/2}$ concatenated with a "non-causal" feature $\mathbf{z} \in \mathbb{R}^{N/2}$: $\mathbf{x} = [\mathbf{v}, \mathbf{z}]$. Noise varies across hand-crafted environments e:

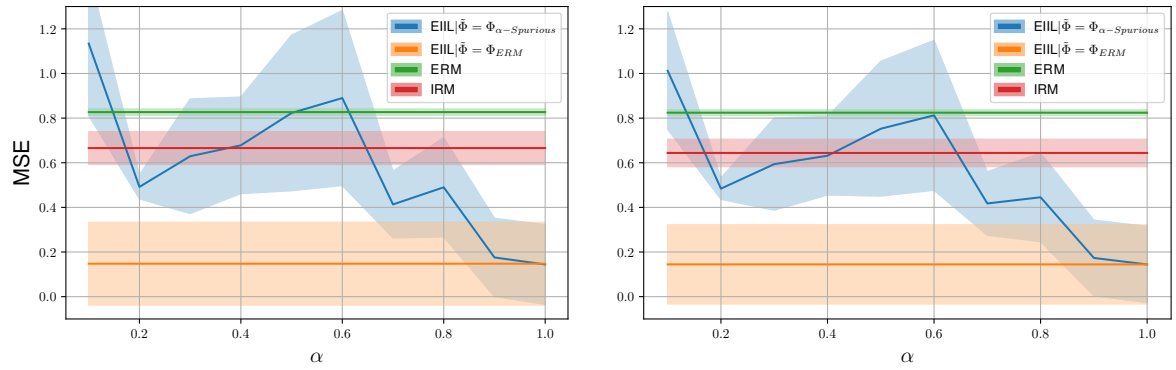
$$\begin{aligned} \mathbf{v} &= \epsilon_{\mathbf{v}} & \epsilon_{\mathbf{v}} &\sim \mathcal{N}(0, 25) \\ \mathbf{y} &= \mathbf{v} + \epsilon_{\mathbf{y}} & \epsilon_{\mathbf{y}} &\sim \mathcal{N}(0, e^2) \\ \mathbf{z} &= \mathbf{y} + \epsilon_{\mathbf{z}} & \epsilon_{\mathbf{z}} &\sim \mathcal{N}(0, 1). \end{aligned}$$

We evaluated the performance of the following methods:

- **ERM:** A naive regressor that does not make use of environment labels e, but instead optimizes the average loss on the aggregated environments;
- **IRM:** the method of Arjovsky et al. (2019) using hand-crafted environment labels;
- **ICP:** the method of Peters et al. (2016) using hand-crafted environment labels;
- **EIIL:** our proposed method (which does use hand-crafted environment labels) that infers useful environments based on the naive ERM, then applies IRM to the inferred environments.

The regression methods fit a scalar target $y = \mathbf{1}^T\mathbf{y}$ via a regression model $\hat{y} \approx \mathbf{w}^T\mathbf{x}$ to minimize $||y - \hat{y}||$ w.r.t. $\mathbf{w}$, plus an invariance penalty as needed. The optimal (causally correct) solution is $\mathbf{w}^* = [\mathbf{1}, \mathbf{0}]$ Given a solution $[\hat{\mathbf{w}}_v, \hat{\mathbf{w}}_z]$ from one of

the methods, we report the mean squared error for the causal and non-causal dimensions as $||\hat{\mathbf{w}}_v - \mathbf{1}||_2^2$ and $||\hat{\mathbf{w}}_z - \mathbf{0}||_2^2$ (Table 7). Because $\mathbf{v}$ is marginally noisier than $\mathbf{z}$, ERM focuses on the spurious $\mathbf{z}$. IRM using hand-crafted environments, denoted IRM, exploits variability in noise level in the non-causal feature (which depends on the variability of $\sigma_{\mathbf{y}}$) to achieve lower error. Using EIIL instead of hand crafted environments yields an improvement on the resulting IRM solution by learning worst-case environments for invariant training.



(a) Causal error (b) Non-causal error

Figure 5. MSE of the causal feature $\mathbf{v}$ and non-causal feature $\mathbf{z}$. EIIL applied to the ERM solution (Black) out-performs IRM based on the hand-crafted environment (Green vs. Blue). To examine the inductive bias of the reference model $\tilde{\Phi}$, we hard code a model $\tilde{\Phi}_{\alpha-\text{SPURIOUS}}$ where α controls the degree of spurious feature representation in the reference classifier; EIIL outperforms IRM when the reference $\tilde{\Phi}$ focuses on the spurious feature, e.g. with $\tilde{\Phi}$ as ERM or α-SPURIOUS for high α.

We show in a follow-up experiment that the EIIL solution is indeed sensitive to the choice of reference representation, and in fact, can only discover useful environments (environments that allow EIIL to learn the correct causal representation) when the reference representation encodes the *incorrect* inductive bias by focusing on the spurious feature. We can explore this dependence of EIIL on the mix of spurious and non-spurious features in the reference model by constructing a $\tilde{\Phi}$ that varies in the degree it focuses on the spurious feature, according to convex mixing parameter $\alpha \in [0, 1]$. $\alpha = 0$ indicates focusing entirely on the correct causal feature, while $\alpha = 1$ indicates focusing on the spurious feature. We refer to this variant as EIIL$|\tilde{\Phi} = \Phi_{\alpha-\text{SPURIOUS}}$, and measure its performance as a function of α (Figure 5). Environment inference only yields good test-time performance for high values of α, where the reference model captures the *incorrect* inductive bias.

F.2. ColorMNIST

	Train accs	Test accs
Grayscale (oracle)	75.3 ± 0.1	72.6 ± 0.6
IRM (oracle envs)	71.1 ± 0.8	65.5 ± 2.3
ERM	86.3 ± 0.1	13.8 ± 0.6
EIIL	73.7 ± 0.5	68.4 ± 2.7
Binned EI heuristic (Sec. 3.3)	73.9 ± 0.5	69.0 ± 1.5
Φ_{Color}	85.0 ± 0.1	10.1 ± 0.2
EIIL$\|\tilde{\Phi} = \Phi_{Color}$	75.9 ± 0.4	68.0 ± 1.2
ARL	88.9 ± 0.2	20.7 ± 0.9
GEORGE	84.6 ± 0.3	12.8 ± 2.0
LFF;$\mathcal{L}_{bias} = \text{GCE}_{q\to 0}$	96.6 ± 1.3	30.6 ± 1.0
LFF;$\mathcal{L}_{bias} = \text{GCE}_{q=0.7}$	15.0 ± 0.1	90.0 ± 0.3

Table 8. Additional baselines for the CMNIST experiment reported in Table 2. The mean and standard deviation of accuracy across ten runs ($\theta_y = 0.25$) are reported. See text for description of the baseline methods.

Table 8 expands on the results from Table 2 by adding the following baselines that do not require environment labels:

- Grayscale: a classifier that removes color via pre-processing, which represents an oracle solution
- EIIL$|\tilde{\Phi} = \Phi_{ERM}$ (reported as EIIL in Table 2
- Binned EI heuristic: the binning heuristic for environment inference described in Section 3.3.
- Φ_{Color}: a hard-coded classifier that predicts *only* based on the digit color
- EIIL$|\tilde{\Phi} = \Phi_{Color}$: EIIL using color-based classifier (rather than Φ_{ERM}) as reference.
- GEORGE (Sohoni et al., 2020): This two-stage method seeks to learn the "hidden subclasses" by fitting a latent cluster model to the (per-class) distribution of logits of a reference model. The inferred hidden subclasses are fed to a GroupDRO learner, so this approach can be seen as an instance of EIIL under particular choices of (unsupervised) EI and (robust optimization) IL objectives.
- ARL (Lahoti et al., 2020): A variant of DRO that uses an adversary/auxiliary model to learn worst-case per-example importance weights. Unlike with EIIL, the auxiliary model and main model are trained jointly.
- LFF (Nam et al., 2020) jointly trains a "biased" model f_B and "debiased" model f_D. f_B is similar to our ERM reference model, but is trained with $\text{GCE}_q(p(x;\theta), y) = \frac{1-p_y(x;\theta)^q}{q}$ with hyperparameter $q \in (0, 1]$,[19] and its per-example losses determine importance weights for f_D.

When expanding this study we find that, unlike EIIL, the new baselines fail to find an invariant classifier that predicts based on shape rather than color. Given that GEORGE does a type of unsupervised EI, it is perhaps surprising that it cannot uncover optimal environments for use with its GroupDRO learner. We hypothesize that this is due to assumption of the relevant latent environment labels being "hidden subclasses", meaning that all examples in an optimal environment must share the same class label value. In the CMNIST dataset, this assumption does not hold due to label noise.

We find that, on this dataset, LFF is very sensitive to the hyperparameter q, which shapes the GCE loss of f_B. Interestingly, using the default value of $q = 0.7$, LFF performs optimally on the test set, but this is *not* because the method has learned an invariant classifier based on the digit shape. The below-chance train set performance reveals that LFF has learned an *anti-color* classifier, exactly the opposite of what ERM does. When q approaches zero (GCE approaches standard cross entropy), LFF fails to generalizes to the OOD test distribution.

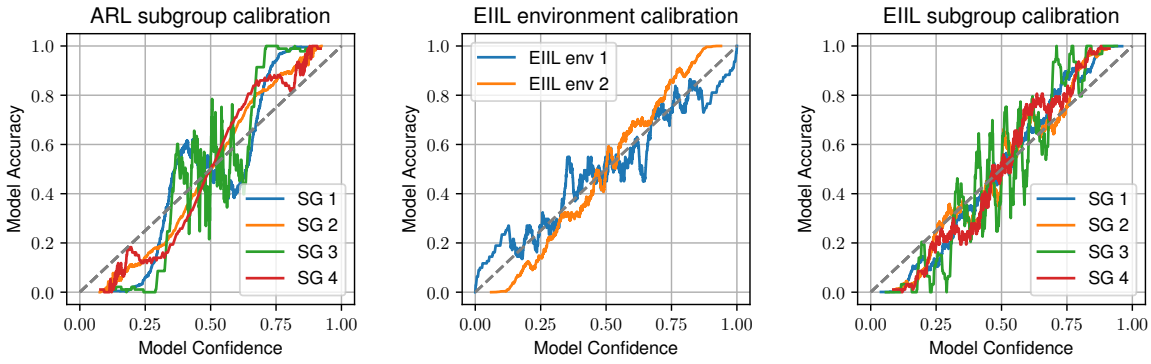
Finally, we found that because the reference classifier predicts with high confidence on the training set, there are only two populated bins in practice. Consequentially, the binned EI heuristic is equivalent to splitting errors into one environment and correct predictions into the other.

F.3. Adult-Confounded

Subgroup sufficiency In the main result we showed that EIIL improves test calibration and accuracy our variant of the UCIAdult dataset. Because the test set is subject to a drastic distribution shift where the correlation pattern between subgroup membership and label is reversed relative to the training set, we can say that this robustness in performance suggests that EIIL does not rely on subgroup membership to make its predictions.

Beyond the global calibration profile, we can also examine calibration curves for the various subgroups, noting again that subgroup labels were not used to train EIIL or the ARL baseline. Figure 6 shows the calibration profiles on the training data. We find that ARL contains noticeable discrepancies in the calibration curves across groups indicating that subgroup sufficiency has not been achieved. EIIL infers environments during the EI phase, which are then implicitly regularized to have roughly the same calibration profile during invariant learning. This can be seen by examining the calibration plots for the training data when it is stratified into the two inferred environments. Finally, looking at calibration curves for the subgroups themselves suggests that EIIL has improved on subgroup sufficiency relative to ARL by better matching the calibration curves across subgroups. These curves still exhibit some noise, indicating that further progress on subgroup sufficiency could be made by changing the invariant learner, possibly by using a different regularize (besides IRMv1) that better enforces the invariance principle.

[19] as $q \to 0$ GCE becomes standard cross entropy



(a) (b) (c)

Figure 6. We examine *subgroup sufficiency*—whether calibration curves match across demographic subgroups—on the Adult-Confounded dataset. Whereas ARL is not subgroup-sufficient (a), EIIL infers worst-case environments and regularizes their calibration to be similar (b), ultimately improving subgroup sufficiency (c). Note that neither method uses sensitive group information during learning.

Ablation Here we provide an ablation study extending Adult-Confounded experiments to demonstrate that both ingredients in the EIIL solution—finding worst-case environment splits and regularizing using the IRMv1 penalty—are necessary to achieve good test-time performance on the Adult-Confounded dataset.

	Train accs	Test accs
EIIL	68.7 ± 1.7	$\mathbf{79.8 \pm 1.1}$
EIIL (no regularizer)	78.6 ± 2.0	69.2 ± 2.8
IRM (random environments)	$\mathbf{94.7 \pm 0.1}$	17.6 ± 1.6

Table 9. Our ablation study shows that both ingredients of EIIL (finding worst-case environments and regularizing invariance across them) are required to achieve good test-time performance on the Adult-Confounded dataset.

From Lahoti et al. (2020) we see that ARL can perform favorably compared with DRO (Hashimoto et al., 2018) in adaptively computing how much each example should contribute to the overall loss, i.e. computing the per-example γ_i in $C = \mathbb{E}_{x_i,y_i \sim p}[\gamma_i \ell(\Phi(x_i), y_i)]$. Because all per-environment risks in IRM are weighted equally (see Equation IRMv1), and each per-environment risk comprises an average across per-example losses within the environment, each example contributes its loss to the overall objective in accordance with the size of its assigned environment. For example with two environments e_1 and e_2 of sizes $|e_1|$ and $|e_2|$, we implicitly have the per-example weights of $\gamma_i = \frac{1}{|e_1|}$ for $i \in e_1$ and $\gamma_i = \frac{1}{|e_2|}$ for $i \in e_2$, indicating that examples in the smaller environment count more towards the overall objective. Because EIIL can discover worst-case environments of unequal sizes, we measure the performance of EIIL using only this reweighting, without adding the gradient-norm penalty typically used in IRM (i.e. setting $\lambda = 0$). To determine the benefit of worst-case environment discovery, we also measure IRM with random assignment of environments. Table 9 confirms that both ingredients are required to attain good performance using EIIL.