Environment Inference for Invariant Learning

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

Abstract

Learning models that perform well under changes to the data distribution is central to research on domain- or out-of-distribution generalization, robust optimization and fairness. Domain-invariant learning has enabled exciting progress in this direction but has an important drawback: the reliance on implicit definition of which features are considered stable or spurious by manually partitioning the training set into "domains" or "environments". Our focus is on the more realistic setting where environment partitions are not provided. We propose an environment-discovery algorithm that leverages Invariant Risk Minimization to discover maximally informative environment partitions automatically, and connect it to the fairness notion of group-sufficiency. We then show theoretically and empirically how different partitioning schemes can lead to *increased* or *decreased* generalization performance, enabling us to outperform IRM with handcrafted environments in multiple cases. Our method outperforms IRM on the ColorMNIST dataset *without using the provided environment splits*. However, we also identify cases where partitioning schemes lead to models that ignore essential features and hence fail to generalize completely.

1. Introduction

Machine learning achieves super-human performance on many tasks when the test data is drawn form the same distribution as the training data. However, when the two distributions differ, model performance can severely degrade to even below chance predictions (Geirhos et al., 2020). Tiny perturbations can derail classifiers, as shown by adversarial examples (Szegedy et al., 2013) and common-corruptions in image classification (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 model brittleness in the face of input distribution changes. Robust optimization aims to achieve good performance on any distribution close to the training distribution (Goodfellow et al., 2014; Madry et al., 2017). Domain-generalization on the other hand tries to go one step further, to generalize to distributions potentially far away from the training distribution. Invariant learning is a successful approach to achieve domain generalization that takes inspiration from causal discovery and encourages invariance across training environments or domains (Peters et al., 2016; Arjovsky et al., 2019; Krueger et al., 2020). The goal is to learn predictors invariant to attributes of the data that change across environments, allowing the model to generalize to different and unobserved configurations of said attributes. However, these methods rely on pre-specified environment partitions implicitly defining the attributes one wishes to be invariant to. Making well-informed choices about useful environment partitions can require extensive expert knowledge or may even be impossible in practice, where only a large *observational* dataset is typically available, which begs the question of how to leverage it to make it useful for Invariant Learning.

We propose a new method to *infer* environment partitions from observational training data. The core idea is to identify environments that maximally violate an Invariant Learning objective, using the predictive distribution of a (potentially

Method	Handcrafted Environments	Train accs	Test accs
ERM	✗	86.3 ± 0.1	13.8 ± 0.6
IRM	✓	71.1 ± 0.8	65.5 ± 2.3
EIIL+IRM	✗	73.7 ± 0.5	$\mathbf{68.4} \pm \mathbf{2.7}$

Table 1. Results on CMNIST, a digit classification task where color is a spurious feature that correlates with the label at train time but anti-correlates with it at test time. Our method **Environment Inference for Invariant Learning (EIIL)** augments IRM to improve test set performance without knowledge of pre-specified environment labels, by instead finding worst-case environments using aggregated observational training data and a reference classifier.

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

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

biased) classifier trained with standard empirical risk minimization (ERM) on the whole dataset. We call this method **Environment Inference for Invariant Learning (EIIL)** and find that in some settings inferring environments directly from observational data *improves* Invariant Learning relative to using the hand-crafted environments (Table 1).

To characterize when EIIL will succeed or fail, we study the effect of encoding inductive biases into environment specification more generally. It turns out that it is possible to partition into environments in ways that *hurt* generalization performance of IRM when compared to a naive ERM baseline, highlighting the difficulty or even impossibility of handcrafting good environments for any given problem. We also give a sufficient condition for EIIL to succeed and discover environments that will have maximal utility for IRM to learn invariant representations. Finally, we point out some very close relationships between the invariant learning and fairness learning formulations, including work on fairness analogous to our environment discovery approach.

2. Methods

2.1. Environment Inference for Invariant Learning

We now derive a principle for inferring environments from observational data. Our exposition extends IRM, but we emphasize that EIIL is applicable more broadly to any environment-based learning objective.

Let $\mathcal{X}$ be the input space, $\mathcal{E}$ the set of environments, $\mathcal{Y}$ the target space.Let $x, y, e \sim p^{obs}(x, y, e)$ be observational data, perhaps realized by conditionally sampling from hand-crafted environments: $p^{obs}(x, y, e) = p^{obs}(e)p^{obs}(x, y|e)$. $\mathcal{H}$ denotes a representation space , e.g. the vector space of logits. $\Phi : \mathcal{X} \to \mathcal{H}$ denotes the parameterized mapping or "model" that we optimize. We refer to $\Phi(x) \in \mathcal{H}$ as the "representation" of example x. $\ell : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}$ denotes a scalar loss to be computed per example, and denote by ℓ^e an environment-dependent loss function. The empirical risk minimization (ERM) solution is found by minimizing the global risk, expressed as the expected loss over the observational distribution:

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

In *Invariant Learning*, we learn representation $\Phi(x)$ that encodes some notion of invariance across the environments $e \in \mathcal{E}$, with the ultimate goal of generalizing to an unknown test dataset $p(x, y|e_{test})$. Arjovsky et al. (2019) proposed IRM, an instantiation of the Invariant Learning principle:

$$\mathbb{E}[Y|\Phi(x) = h, e_1] = \mathbb{E}[Y|\Phi(x) = h, e_2] \quad \forall\, h \in \mathcal{H}\ \forall\, e_1, e_2 \in \mathcal{E}. \tag{1}$$

Intuitively, IRM aims to learn a predictor that is simultaneously Bayes optimal in all environments. As a practical instantiation of this principle, Arjovsky et al. (2019) introduce IRMv1, a gradient-penalty regularized objective enforcing simultaneous optimality of the same classifier $w \circ \Phi$ in all environments.[1] Denoting by $R^e = \mathbb{E}_{p^{obs}(x,y|e)}[\ell^e]$ the per-environment risk, the objective to be minimized is

$$C^{IRM}(\Phi) = \sum_{e \in \mathcal{E}} R^e(\Phi) + \lambda ||\nabla_{w|w=1.0} R^e(w \circ \Phi)||].$$

IRM assumes pre-defined environment indicators are given, but we are interested in understanding whether hand-crafted environments are necessary, and whether they could be improved upon. We first make a simplifying assumption[2] that $R^e = R\,\forall\, e$, i.e. $\ell^e = \ell\,\forall\, e$. Then by introducing $\mathbf{u}_i(e') = p^{obs}(e|x_i, y_i) = \mathbb{1}(e_i = e')$ as an indicator of the hand-crafted environment assignment per-example, and noting that $N_e := \sum_i \mathbf{u}_i(e)$ represents the number of examples in environment e, we can re-express this objective to make its dependence on environment labels explicit

$$C^{IRM}(\Phi, \mathbf{u}) = \sum_{e \in \mathcal{E}} \frac{1}{N_e} \sum_i \mathbf{u}_i(e)\ell(\Phi(x_i), y_i) + \sum_{e \in \mathcal{E}} \lambda ||\nabla_{w|w=1.0} \frac{1}{N_e} \sum_i \mathbf{u}_i(e)\ell(w \circ \Phi(x_i), y_i)||_2.$$

Our general strategy is to replace the binary indicator $\mathbf{u}_i(e)$, with a probability distribution $q(e|x_i, y_i)$, representing a soft assignment of the i-th example to the e-th environment. We would like $q(e|x_i, y_i)$ to capture worst-case environments w.r.t the invariant learning objective; rewriting $q(e|x_i, y_i)$ as $\mathbf{q}_i(e)$ for consistency with the above expression, we arrive at the following bi-level optimization:

$$\min_{\Phi} \max_{\mathbf{q}} C^{IRM}(\Phi, \mathbf{q}). \tag{EIIL}$$

We leave the full exploration of this bi-level optimization to future work, but for now propose the following practical sequential, which we call EIILv1 (cf. See Appendix A):

1. Input *reference model* $\tilde{\Phi}$
2. Fix $\Phi \leftarrow \tilde{\Phi}$ and fully optimize the inner loop of (EIIL) to infer environments $\tilde{\mathbf{q}}_i(e) = \tilde{q}(e|x_i, y_i)$
3. Fix $\mathbf{q} \leftarrow \tilde{\mathbf{q}}$ and fully optimize the outer loop to yield the new model Φ.

Instead of requiring hand-crafted environments, we instead require a trained reference model $\tilde{\Phi}$, which is arguably easier to produce and could be found using ERM on $p^{obs}(x, y)$,

[1] $w \circ \Phi$ yields a classification decicion via linear weighting on the representation features

[2] This holds for all experiments carried out by Arjovsky et al. (2019). Because we would like to handle the setting without pre-specified environments, we accordingly assume that environment-dependent risk functions are unknown so a global risk is used.

for example. In our experiments we consider binary environments and explicitly parameterize the $q(e|x, y)$ as a vector of probabilities for each example in the training data.[3]

2.2. The inductive bias of the reference representation

To characterize the ability of EIILv1 to generalize to unseen test data, we now examine the inductive bias for generalization provided by the reference model $\tilde{\Phi}$. For lack of space, we state the main results 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/causal w.r.t. the label Y. Our proof considers binary features and labels, but the same argument extends to other cases. The aim of invariant learning is to form a model Φ whose representation $\Phi(V, Z)$ is invariant w.r.t. Z and focuses solely on V. We consider the case with two environments. We discuss how satisfying the Invariance Principle (Equation 1) depends jointly on the model Φ and choice of environments $\{e\}$.

Theorem 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 $\tilde{\Phi}_{Spurious}$ invariant to valuable features V and solely focusing on spurious features Z maximally violates the Invariance Principle (Equation 1). And vice versa, if Φ focuses on the spurious features then a choice of environments that maximally violates the Invariance Principle is $e_1 = \{V, Z, Y|\mathbb{1}(Y = Z)\}$ and $e_2 = \{V, Z, Y|\mathbb{1}(Y \neq Z)\}$.*

Corollary 1.1 *The environment labels provided in the CMNIST dataset (Arjovsky et al., 2019) do not maximally violate the Invariance Principle in Equation 1 for a reference model invariant to V and solely focusing on Z, and are thus not maximally informative for learning a model ignoring color.*

Remark 1 *In practice we find that $\tilde{\Phi}_{ERM}$ on CMNIST approximates $\tilde{\Phi}_{Color}$. Thus we can automatically find environment partitions that provide a starker contrast w.r.t. the relationship between the spurious feature and label, compared with hand-crafted environments.*

Remark 2 *If environments are split according to agreement of Y and Z, then the constraint from Equation 1 (the Invariance Principle) is satisfied under a representation that ignores Z: $\Phi(X) \perp Z$.*

Remark 3 *On CMNIST, using EIIL to find environments based on the ERM solution $\tilde{\Phi}_{ERM}$ then running IRM (denoted IRM(e_{EIIL})) leads to better test set generalization than IRM on the hand-crafted environments (denoted IRM(e_{HC})). See Table 1.*

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

Table 2. 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.

2.3. Connections to Fairness

Equation 1 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 the environment e is replaced by a "sensitive" attribute a denoting membership in a protected demographic group (age, race, gender, etc.), and the vector representation $\Phi(X)$ is typically replaced by a scalar score $S(x) \in \mathbb{R}$. $\mathbb{E}[Y|S(X), a]$ 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. (2018) studied "group sufficiency" of classifiers with convex loss, concluding that ERM naturally finds calibrated (and thus group sufficient) solutions without fairness constraints[4].

In this work we are interested in whether worst-case environments can be automatically discovered. Closely related to our approach, Hébert-Johnson et al. (2017) proposed *Multicalibration* as a way of ensuring a classifier's calibration curve is invariant to efficiently computable environment splits. However we note that the proposed algorithm requires brute force enumeration over all possible environments[5]; we consider a more practical approach based on finding the worst-case environments.

[3]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.

[4]Because of the convexity assumption, these results do not apply to the neural net representations considered by IRM.

[5]A more practical algorithm was considered in Kim et al. (2019) by relaxing the calibration constraint to an accuracy constraint.

3. Experiments

3.1. Synthetic Data

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(e_{HC}):** the method of Arjovsky et al. (2019) using hand-crafted environment labels $\{e\}$. **ICP(e_{HC}):** the method of Peters et al. (2016) using hand-crafted environment labels $\{e\}$. **IRM(e_{EIIL}):** our proposed method (which does use environment labels $\{e\}$) 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 2). Because $\mathbf{v}$ is marginally noisier than $\mathbf{z}$, ERM focuses on the spurious $\mathbf{z}$. IRM using hand-crafted environments, denoted IRM(e_{HC}), 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 (denoted IRM(e_{EIIL})) by learning worst-case environments for invariant training.

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 capture the correct causal solution when the reference representation encodes the *incorrect* inductive bias by focusing on the spurious feature. Similarly, IRM will only improve over the ERM solution when ERM does not successfully focus on the causal feature $\mathbf{v}$, while the environments vary in their dependence on the non-causal feature $\mathbf{z}$. 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 IRM($e_{\text{EIIL}}|\tilde{\Phi} = \Phi_{\alpha-\text{SPURIOUS}}$) , and measure its performance as a function of α (Figure 1).

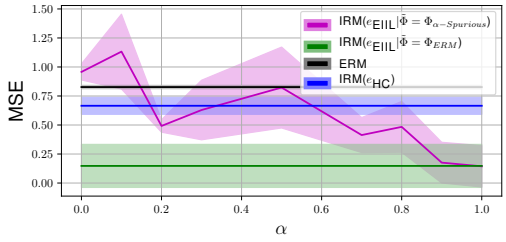


Figure 1. MSE of the causal feature $\mathbf{v}$. IRM(e_{EIIL})applied to the ERM solution (Black) out-performs IRM based on the hand-crafted environment (Green vs. Blue). To examine the inductive bias of the reference model $\tilde{\Phi}$, we we hard code a model $\tilde{\Phi}_{\alpha-\text{SPURIOUS}}$ where α controls the degree of spurious feature representation in the reference classifier; IRM(e_{EIIL}) outperforms IRM(e_{HC}) when the reference $\tilde{\Phi}$ focuses on the spurious feature, e.g. with $\tilde{\Phi}$ as ERM or α-SPURIOUS for high α.

3.2. ColorMNIST

ColorMNIST (CMNIST) is a noisy digit recognition task[6] 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 two pre-specified environments (Arjovsky et al., 2019). Crucially, label noise is applied by flipping y with probability θ_y; the default setting ($\theta_y = .0.25$) implies that shape (the correct feature) is marginally less reliable than color in the train set, so naive ERM ignores shape to focus on color and suffers from below-chance performance at test time.

After noting that EIIL outperforms IRM *without access to environment labels* in the default setting (See Tables 1 and 3), we examine how the various methods perform as a function of θ_y. This parameter influences the ERM solution since low θ_y implies shape is more reliable than color in the aggregated training data (thus ERM generalizes poorly), while the opposite trend holds for high θ_y. Because EIIL relies on a reference model $\tilde{\Phi}$, its performance is also affected when $\tilde{\Phi} = \text{ERM}$ (see Figure 3 in Appendix D.2). We find that IRM(e_{EIIL}) generalizes better than IRM(e_{HC}) with sufficiently high label noise $\theta_y > .2$, but generalizes poorly under low label noise. This is precisely due to the success of ERM in this setting, where shape is a more reliable feature in the training data than color. We verify this conclusion by evaluating IRM(e_{EIIL}) when $\tilde{\Phi} = \Phi_{Color}$, which uses a hand-coded color-based predictor as reference representation, which succeeds across all settings of θ_y.

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

References

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

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

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Steinhardt, J., and Madry, A. Identifying statistical bias in dataset replication. *arXiv preprint arXiv:2005.09619*, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

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

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

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

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.

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

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. *arXiv preprint arXiv:1808.10013*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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? *arXiv preprint arXiv:1902.10811*, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

A. EIIL Psuedocode

Algorithm 1 Example applying EIIL to infer two environments that maximally violate the IRM objective.

Input: Reference model $\tilde{\Phi}$, dataset $\mathcal{D} = \{x_i, y_i\}$ with $x_i, y_i \sim p^{obs}$ iid, loss function ℓ, duration N_{steps}
Output: Worst case data splits $\mathcal{D}_1, \mathcal{D}_2$ for use with IRM.
Randomly initialize $\mathbf{e} \in \mathbb{R}^{|\mathcal{D}|}$ as vectorized logit of posterior with $\sigma(\mathbf{e}_i) := q(e|x_i, y_i)$. **for** $n \in 1 \dots N_{steps}$ **do**

$R^1 = \frac{1}{\sum_{i'} \sigma(\mathbf{e}_{i'})} \sum_i \sigma(\mathbf{e}_i)\ell(\tilde{\Phi}(x_i), y_i)$; `// D1 risk`

$G^1 = \nabla_{\mathbf{w}|\mathbf{w}=1} || \frac{1}{\sum_{i'} \sigma(\mathbf{e}_{i'})} \sum_i \sigma(\mathbf{e}_i)\ell(\mathbf{w} \circ \tilde{\Phi}(x_i), y_i)||_2$; `// D1 invariance regularizer`

$R^2 = \frac{1}{\sum_{i'} 1-\sigma(\mathbf{e}_{i'})} \sum_i (1-\sigma(\mathbf{e}_i))\ell(\tilde{\Phi}(x_i), y_i)$; `// D2 risk`

$G^2 = \nabla_{\mathbf{w}|\mathbf{w}=1} || \frac{1}{\sum_{i'} 1-\sigma(\mathbf{e}_i)} \sum_i (1-\sigma(\mathbf{e}_i))\ell(\mathbf{w} \circ \tilde{\Phi}(x_i), y_i)||_2$; `// D2 invariance regularizer`

$L = \frac{1}{2} \sum_{e \in \{1,2\}} R^e + \lambda G^e$

$\mathbf{e} \leftarrow OptimUpdate(\mathbf{e}, \nabla_{\mathbf{e}} L)$

end

$\mathbf{e} \sim Bernoulli(\sigma(\mathbf{e}))$; `// sample splits`

$\mathcal{D}_1 \leftarrow \{x_i, y_i | \hat{\mathbf{e}}_i = 1\}, \mathcal{D}_2 \leftarrow \{x_i, y_i | \hat{\mathbf{e}}_i = 0\}$; `// split data`

B. Proof of Theorem 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 (Equation 1) holds, we introduce a measure called the *group sufficiency gap*[7]:

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

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 Theorem 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 $Z = 0$ $E(Y|Z, I[Y == Z]) = 0$ 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 $\Delta(S, X, Y, e) = 1$, which is the maximum value for the sufficiency gap.

[7] This was previously used in a fairness setting by Liu et al. (2018) to measure differing calibration curves across groups.

	Train accs	Test accs
ERM	86.3 ± 0.1	13.8 ± 0.6
IRM(e_{HC})	71.1 ± 0.8	65.5 ± 2.3
IRM($e_{\text{EIIL}}|\tilde{\Phi} = \Phi_{ERM}$)	73.7 ± 0.5	68.4 ± 2.7
IRM($e_{\text{EIIL}}|\tilde{\Phi} = \Phi_{Color}$)	75.9 ± 0.4	68.0 ± 1.2
Φ_{Color}	85.0 ± 0.1	10.1 ± 0.2
GRAYSCALE	75.3 ± 0.1	72.6 ± 0.6

Table 3. Accuracy across ten runs with label noise $\theta_y = 0.25$ GRAYSCALE hard codes out the color feature and thus represents an oracle solution to CMNIST.

C. Training details

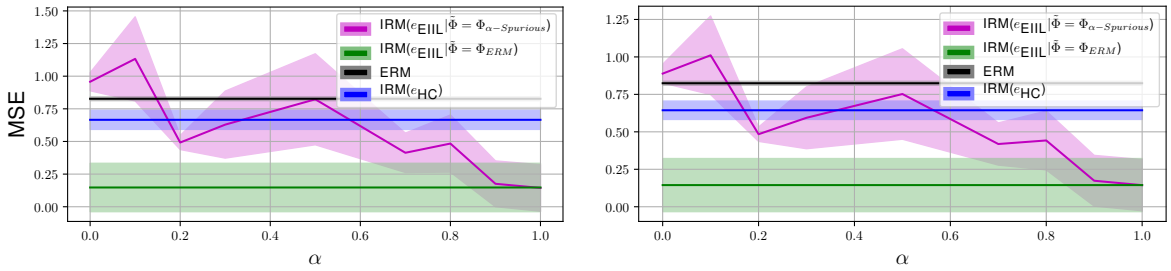
IRM is trained on these two environments and tested on a holdout environment constructed from 10.0000 test images in the same way as there training environments, where colour is predictive of the noisy label 10% of the time. So using color as a feature to predict the label will lead to an accuracy of roughly 10% on the test environment, while it yields 80% and 90% accuracy respectively on the training environments.

To evaluate IRM(e_{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. The obtained environment partitions are then used to train a new model from scratch with IRM.

Krueger et al. (2020) already discussed the problems of this dataset when the test data can be used 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 by using the default parameters of IRM without further tuning.

D. Additional Results

D.1. Synthetic data



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

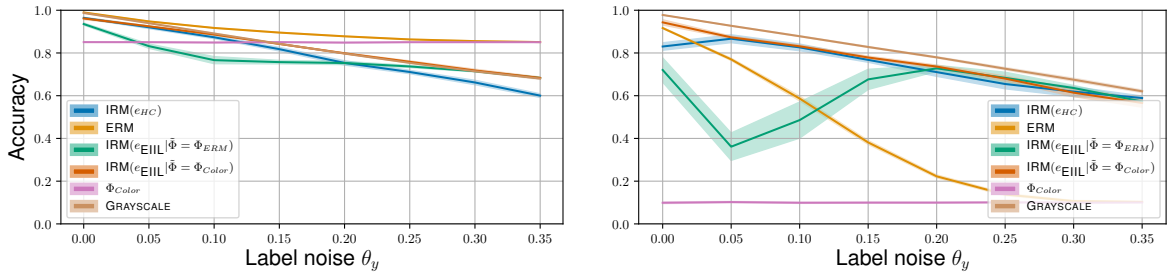
Figure 2. Examining the EIIL solution as a function of hard-coded spuriousness parameter α in the reference classifier. IRM(e_{EIIL})outperforms IRM(e_{HC})when the reference $\tilde{\Phi}$ focuses on the spurious feature, e.g. with $\tilde{\Phi}$ as ERMor α-SPURIOUS for high α.

Table 2 shows complete version of the result from Table 1, including the MSE to the non-causal feature.

D.2. ColorMNIST

Figure 3 shows the full results from the label noise sweep experiment.

Table 3 expands on the results from Table 1 by adding additional methods discussed in Section 2.



(a) Train accuracy.

(b) Test accuracy

Figure 3. CMNIST results for 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, IRM(e_{EIIL}) matches or exceeds the performance of IRM(e_{HC}) on the shifted (spurious correlation reversed) test set *without relying on hand-crafted environments*. Under medium label noise ($.1 < \theta_y < .2$), IRM(e_{EIIL}) under-performs relative to IRM(e_{HC}), but outperforms ERM, the logical approach if environments are not available. Under low label noise ($\theta_y < .1$), where the spurious feature color correlates to label *less* than shape on the train data, ERM performs well and IRM(e_{EIIL}) fails. GRAYSCALE indicates an oracle classifier that predicts based on the causal feature (shape), while Φ_{Color} represents an oracle classifier that predicts based on the spurious feature (color).