Repeated Environment Inference for Invariant Learning

Aayush Mishra [1] **Anqi Liu** [1]

Abstract

We study the problem of invariant learning when the environment labels are unknown. We focus on the Invariant representation notion when the Bayes optimal conditional label distribution is the same across different environments. Previous work conducts the Environment Inference (EI) by maximizing the penalty term in the Invariant Risk Minimization (IRM) framework. The EI step uses a reference model which focuses on spurious correlations to efficiently reach a good environment partition. However, it is not clear how to find such a reference model. In this work, we propose to repeat the EI process and retrain an ERM model on the *majority* environment inferred by the EI step in the previous step. Under mild assumptions, we find that this iterative process helps learn a representation capturing the spurious correlation better than the single step. This results in better Environment Inference and better Invariant Learning. We show that this method outperforms baselines on both synthetic and real-world datasets.

1. Introduction

In conventional machine learning, training data is assumed to be independently and identically distributed (*iid*) as the test data. This assumption is usually violated in the real world. Samples used in training might not be representative of the whole data distribution in many applications. Therefore, performance suffers when machine learning algorithms are deployed in new *domains* or *environments* where spurious correlations learned from the training data do not hold. To solve this problem, recent literature advocates focusing on extracting causal relationships (Peters et al., 2017; Pearl, 2019) from data. However, causal discovery and inference in high-dimensional data remain a big challenge in practice (Schölkopf et al., 2021).

[1]Department of Computer Science, Johns Hopkins University. Correspondence to: Aayush Mishra <amishr24@jh.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

Recent works (Arjovsky et al., 2019; Ilse et al., 2020; Chuang et al.) have proposed to learn stable correlations that are invariant across different domains because they have been shown to strongly relates with Causality (Heinze-Deml et al., 2018; Peters et al., 2016). However, Invariant Learning methods usually require predefined environment labels, which are not always available. For example, fair machine learning can be regarded as an Invariant Learning problem where different sub-populations correspond to different environments, and the sensitive features are usually unavailable or even hard to define (Corbett-Davies & Goel, 2018). On the other hand, environment labels play an important role in invariant learning. Figure 1 shows how the performance of IRM (Arjovsky et al., 2019) increases when training samples are grouped differently (Experiment B) from the ground-truth group labels (Experiment A). To create two training environments, we randomly *shuffle* the background for digits with different probabilities. IRM benefits from the regrouping as the new grouping reflects a better differentiation between the two environments. See Figure 1.

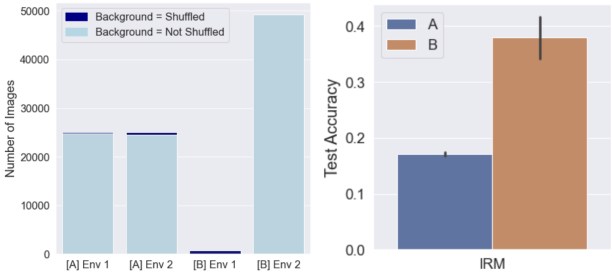


Figure 1. **Left:** Data distributions in two environments in experiment A and B; **Right:** IRM test accuracy in experiment A and B; Experiment **A**: CBMNIST [I] samples with *shuffled* background w.p. 1% and 2% in both environments, respectively; Experiment **B**: Samples with *shuffled* background are all grouped in one environment. The test environment has completely different backgrounds from training in both cases. IRM's performance is significantly higher in Experiment B. More details in Section 4.1.

Creager et al. (2021) proposed EIIL, enabling invariant learning when environment labels are unavailable. It conducts Environment Inference (EI) using the representation learned by a reference model by maximizing the penalty term in the Invariant Risk Minimization (IRM) framework.

Moreover, it is shown that EIIL usually works when the reference model captures "spurious" correlations. However, it is unclear how to find such a model besides learning an ERM model from all the training data.

In this paper, we find that environment partition obtained in a single EI step after ERM is often sub-optimal for invariant learning. We propose a simple repeated EI strategy to improve the quality of Environment Inference and, consequently, the performance of Invariant Learning. **Our contribution can be summarized as follows:**

1. We propose a novel method that generates environment labels by repeating the EI step and training the ERM model on the *majority* environment. We call it Repeated Environment Inference for Invariant Learning (REIIL).
2. Under mild assumptions, we show that our strategy better facilitates downstream invariant learning.
3. Our approach outperforms previous methods, including EIIL on previously used and a newly introduced dataset.

2. Related Work

IRM and Environment Inference Arjovsky et al. (2019) proposes a training objective for learning the invariant representations, under the assumption that the Bayes optimal conditional $P(y|\phi(x))$ remains invariant across domains. However, recent works (Rosenfeld et al., 2020; Kamath et al., 2021) have shown that IRM fails when target domains are not sufficiently similar to training domains. Besides, it also requires good environment labels beforehand. Several methods have been proposed to infer the environment labels before or along with the invariant learning. EIIL (Creager et al., 2021) takes a two-step strategy and first partitions the training data into environments using a biased reference model before using environment-based invariant learning. However, it is unclear how to get a perfectly biased reference model to obtain a good environment partition. Adversarial Invariant Learning (Ye et al., 2021) uses a minimax game to robustly train the predictors while inferring the worst-case environment partition. In this paper, we follow the EIIL setup and study how to conduct environment inference better for downstream invariant learning.

Other Types of Invariance in Domain Generalization Besides IRM, other definitions of *invariance* exist in the domain generalization literature. Dated back to early works in domain adaptation (Gretton et al., 2009; Sun & Saenko, 2016), marginal feature alignment motivates the learning of the domain-invariant features, that is $P(\phi(x))$ being invariant across domains (Ganin et al., 2016). However, the shortcoming of these methods is also obvious: it ignores the relationship between the features and the labels, so it still suffers under label shift (Zhao et al., 2019; 2020). Other works utilize tools like Information Bottleneck and propose learning objectives based on the mutual information (Li et al., 2021). Li et al. (2018) focus on conditional domain invariant representation and make sure that the $P(\phi(x)|y)$ remains invariant across domains. Focusing on only one type of invariance is inadequate (Shui et al., 2022), and choosing the right type of invariant learning algorithm can be tricky.

Some works also propose using domain-specific features along with domain-invariant features to improve performance in domain generalization tasks (Ding & Fu, 2017; Chattopadhyay et al., 2020; Bui et al., 2021). However, it is hard to say whether exploiting domain-specific features will always help in unseen domains without theoretical guarantees. Our paper aims to learn the invariant $P(y|\phi(x))$ distribution when environment labels are unavailable.

Representation Learning under Domain Shift Other representation learning techniques have also been found helpful under domain shift. For example, Liu et al. (2021) proposes that upweighting underperforming samples in a second iteration of training is sufficient to improve worst-case group performance. Kirichenko et al. (2022) suggest that last layer retraining of Neural Networks with interesting samples might be sufficient for robustness against spurious correlations. Contrastive Learning approaches (Wang et al., 2022; Kim et al., 2021) have also shown promise for domain adaptation and generalization. However, these methods do not explicitly learn any invariance and are usually hard to analyze. In our paper, we focus on the more principled IRM setting and overcome the difficulty of obtaining biased models for environment inference.

3. Method

3.1. IRM method and Environment Inference

The practical version of IRM (called IRMv1) has the following objective,

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\|^2 \quad (1)$$

where samples are taken from multiple environments and this information is used to regularize the training.
As multiple environments are not always available, EIIL proposes an Environment Inference step to produce the required environment labels. It uses a fixed reference classifier Φ and the IRM penalty term to do this. The **EI step** maximizes the following with respect to soft environment assignment $\boldsymbol{q}$,

$$C^{EI}(\Phi, \boldsymbol{q}) = \|\nabla_{w|w=1.0} \tilde{R}^e(w\cdot\Phi, \boldsymbol{q})\|^2, \quad (2)$$

where $R^e(\Phi, \boldsymbol{q}) = \frac{1}{N}\sum_i \boldsymbol{q}_i(e) l(\Phi(x_i), y_i)$.

This soft assignment is used to produce hard environment labels which are then used for invariant learning.

In summary, EIIL takes the following steps:

1. Takes an input reference model $\tilde{\Phi}$.
2. Maximizes C^{EI} to get environment partition. $\boldsymbol{q}^* = \arg\max_q C^{EI}(\tilde{\Phi}, \boldsymbol{q})$.
3. Minimizes C^{IL} (any Invariant Learning objective of choice) to get the final model. $\Phi^* = \arg\min_\Phi C^{IL}(\Phi, \boldsymbol{q}^*)$.

Note that here $\boldsymbol{q}$ is used to produce a **binary partition**, and we assume this binary partitioning throughout this paper.

EIIL works under the assumption that the reference model ($\tilde{\Phi}$) focuses only on the spurious features. It is only when $\tilde{\Phi} = \Phi_{Spurious}$, does the environment inference (EI) process maximally violates the Invariance Principle (EIC):

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

The authors use Φ_{ERM} as an approximation for $\Phi_{Spurious}$, but this does not always work. As underlined in the paper explicitly via various experiments, it remains unexplored to find better reference models (which would be worse than ERM in terms of generalization performance) to fully exploit the EI step in EIIL.

3.2. Proposed method

We present an assumption about the training distribution and a conjecture about the resulting EI step using a model trained from such a training distribution. We then reason how finding a model that satisfies the conjecture under the assumption would help the downstream invariant learning. We finally propose a strategy to approximate such a model.

Assumption 3.1. Assume a data generating graph $X \to Y \to Z$, where X and Z are observed features, Y is the target feature and the anti-causal mechanism is unstable. A training dataset is sampled following this graph. Assume that in the training dataset, anti-causal features are more informative than the causal features, i.e.,

$$I_{tr}(Y; Z) > I_{tr}(Y; X) > 0 \tag{3}$$

where I measures *mutual information* between features.

Conjecture 3.2. *There exists a model trained on such training datasets (tr) such that when it is used as a reference model for the EI step (2), the resulting majority environment (e_{maj}) satisfies the following:*

$$I_{e_{maj}}(Y; Z) - I_{e_{maj}}(Y; X) \geq I_{tr}(Y; Z) - I_{tr}(Y; X) \tag{4}$$

Remarks: By induction, if we find such a model after one EI step, there exists another such model in the next step, when the majority environment is the training dataset. This process repeats until a point of diminishing returns.

Therefore, we propose a simple repetition of reference model training on a biased subset of samples to get a more biased reference model better approximating $\Phi_{Spurious}$. We use ERM to approximate such a model in each step. The strategy can be defined in the following steps:

1. Perform EI step from EIIL using the ERM reference model trained on the whole dataset.
2. Instead of training an Invariant Learning algorithm directly on the obtained partition, retrain an ERM model on the *majority* environment obtained from the previous step. Repeat this step for n iterations.
3. Use the finally obtained partition for downstream Invariant Learning methods.

The repetition of the EI step is based on the proposed conjecture that majority subsets obtained would show even stronger correlation between spurious and target labels and, the ERM model trained on it would iteratively approximate $\Phi_{Spurious}$.

4. Experimental Results

We show the effectiveness of our proposed method on various datasets under different challenging conditions and find that it recovers *good* environment splits and provides boosted performance in invariant learning. We also provide empirical evidence to support our conjecture.

4.1. Datasets

Colored MNIST (CMNIST) was originally introduced in the IRM paper (Arjovsky et al., 2019). It has a synthetic binary classification task where color is introduced as an anti-causal spurious correlation.

CIFAR-Background MNIST (CBMNIST) We introduce this new dataset, which is based on the simple concept of putting MNIST (LeCun et al., 1998) digits on CIFAR (Krizhevsky et al., 2009) backgrounds which would act as spurious correlations. The classification task for the resultant images remains same as the MNIST digit classification task.

CBMNIST [I] was created in the following manner:

- One random CIFAR image from each class is chosen to act as the background for MNIST images. This results in 10 different backgrounds with different visual properties.
- Each MNIST class is mapped with a randomly selected CIFAR class. This mapping between classes acts as the anti-causal spurious correlation in the training set. For the test set, a completely different random mapping of classes is chosen such that there is no overlap of digit-background combinations with the training set.
- Training data is split into two environments with 25000

samples each (from 60000 MNIST training samples). Test data has the remaining 10000 samples.

- All training samples are filled with the mapped CIFAR backgrounds. In Environment 1, backgrounds of $\sim$1% of samples are *shuffled*. *Shuffled* here means a random background which is not equal to the one in the train or test mapping of that class (one of the remaining 8), is applied to this sample. Similarly, in Environment 2, $\sim$2% samples have their background *shuffled*.

The resulting dataset has $\sim$97% samples exhibiting a strong anti-causal relationship between the target class label and the background. The rest $\sim$3% samples have *shuffled* backgrounds so the only *invariant* features in them are the actual digit shapes. This ensures that the dataset satisfies the assumption (3.1) for our conjecture (3.2) to hold.

CBMNIST [II] has a slight variation of the concept of *shuffled*. Here *shuffled* samples have their background switched randomly to exactly one other pre-decided background rather than 8 in CBMNIST [I]. In this way, even the $\sim$3% *shuffled* samples exhibit a spurious correlation with the background and a particular digit is only ever seen with at most 2 different backgrounds in the training set. The test set remains the same.

CBMNIST [III] has the percentage of samples having *shuffled* backgrounds change from CBMNIST [I]. We changed the numbers 1% and 2% to 10% and 20% respectively, simulating a case where even simple ERM can focus on invariant causal features because of lots of background variations. It is unclear whether our assumption (3.1) holds in this case.

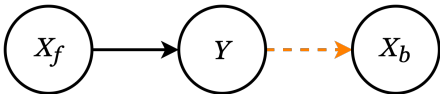


Figure 2. CBMNIST Data generating process. Here X_f is the foreground or digit shape (from MNIST) that directly causes the class label Y of a sample. X_b is the spuriously correlated background (from CIFAR) which is caused by the label Y but the mechanism generating this distribution $P(X_b|Y)$ is unstable and is represented with a colored dashed edge inspired from (Subbaswamy et al., 2022). Note that these variables are abstract and need to be extracted from images.

4.2. Implementation details

We reuse the code [github link], model architectures and hyperparameters from EIIL, which follows IRM for CMNIST experiments. For CMNIST, we increase the number of training steps to 900. For CBMNIST, we use an MLP with a similar structure but instead of down sampling as in CMNIST, we use all the $32 \times 32 \times 3$ input dimensions to help the model learn any complex background features. For REIIL, we found that in $n = 9$ iterations, the performance usually saturates. Note that for EIIL and REIIL, the environment labels are not used in the training. We also used L2 regularization on the parameters for all methods.

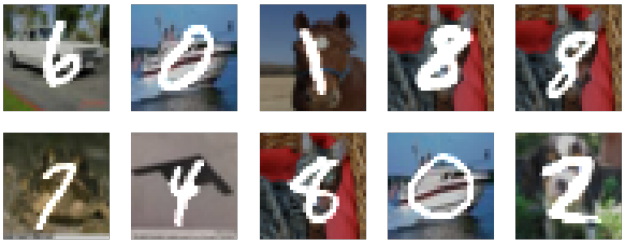


Figure 3. CBMNIST [I] samples. We use CIFAR images as the background for MNIST digits. Figure 2 shows the data generating procss.

4.3. Result Analysis

4.3.1. Better Invariant Learning

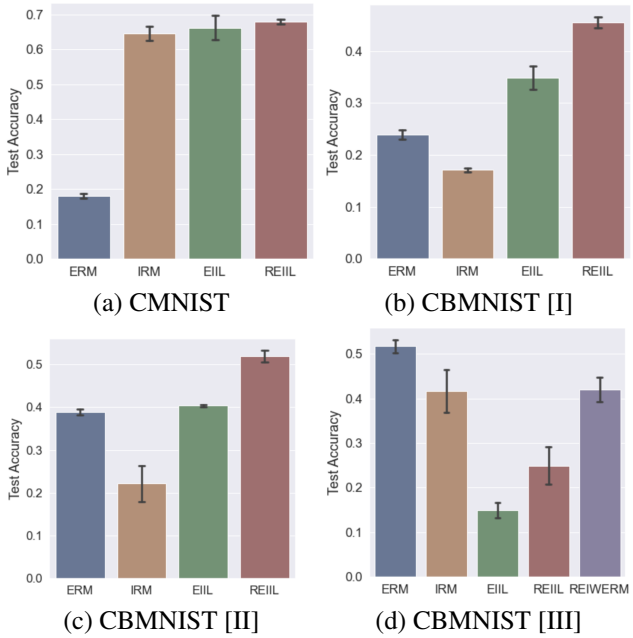
Test performance on CMNIST and CBMNIST variants can be seen in Figure 4. In both datasets, REIIL outperforms other methods. Now we analyze each experiment as follows.

In CMNIST, the performance gains are insignificant because Φ_{ERM} approximates $\Phi_{Spurious}$ quite well in the first EIIL step itself. However, REIIL's performance is more stable than EIIL's across five runs.

In CBMNIST [I], the performance gains are quite substantial. Because the *shuffled* samples are scarce and poorly split, IRM fails to beat even ERM. But REIIL consistently performs better than other methods. For CBMNIST [II], the results were quite similar. Interestingly, in both these cases, ERM outperforms the native IRM, and using the EI step proves critical for performance gains.

ERM emerges as the best-performing model in CBMNIST [III], probably because the higher percentage of *shuffled* background samples can regularize the ERM model's training. The initial reference model is also very far from $\Phi_{Spurious}$, making the performance of EIIL and even REIIL worse than IRM. REIIL still beats EIIL, but it cannot reach the performance of even IRM due to the poor starting point. It highlights the importance of a good reference model for the success of Environment Inference.

One thing to note is the superior performance of the reweighting method to Invariant Learning methods. Inspired from (Liu et al., 2021), we trained a Weighted ERM (WERM) model on the EI obtained environment splits (REI-WERM). WERM weighs samples within each environment equally rather than weighing each sample equally in the whole dataset as in ERM. This model surprisingly matched the performance of IRM. We defer investigating this further for future work.



(a) CMNIST (b) CBMNIST [I]

(c) CBMNIST [II] (d) CBMNIST [III]

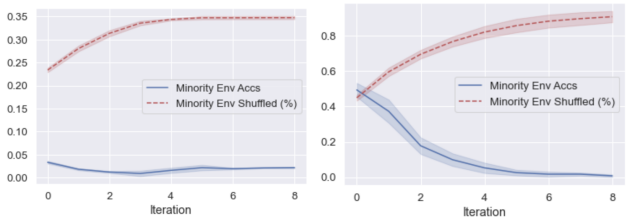
Figure 4. Comparison with baseline methods on CMNIST and CBMNIST. In all datasets, especially CBMNIST, REIIL outperforms other methods.

4.3.2. Dynamics of Repeated EI

If datasets follow our assumption (3.1) and we find a model satisfying Conjecture (3.2) using ERM, we claim that repeating the EI step produces better environment splits such that the spurious features are more informative about the labels in the majority environment (for example, corresponding to the *non-shuffled* images in CBMNIST [I]). This would imply that the minority environment's test accuracy would be low under the reference model trained using the majority environment. It would also imply that most of the *causally informative* samples (in the case of CBMNIST, these would be *shuffled* samples) would fall in the minority environment.

Figure 5 shows the accuracies of the reference models in each step for the samples in the *minority* environment and the percentage of *shuffled* samples assigned in the minority environment with each EI step. We see that accuracies keep decreasing while the proportion of *shuffled* samples increases in the minority environment. Note that the training accuracies in the *majority* environment always reach close to 1, implying that the ERM model successfully trains to convergence in each step. For CMNIST, *shuffled* samples would correspond to those which do not exhibit the color-based spurious correlation. This effect is not as pronounced as in CBMNIST [I], but follows a similar pattern.

Therefore, we can conclude that an ERM model trained on the majority environment obtained from the previous EI step better captures spurious features (background or color) and produces a better environment partition.



(a) CMNIST (b) CBMNIST [I]

Figure 5. Accuracy of reference model drops for the minority environment and percentage of *shuffled* samples increase in the minority environment with REIIL iterations.

5. Discussion

What if our assumption is violated? It was shown by (Gulrajani & Lopez-Paz, 2020) that carefully trained ERM models to work quite well and often outperform Domain Generalization methods. We also see that ERM outperforms other methods in CBMNIST [III]. Our assumption is supposed to hold in more challenging situations where ERM usually fails. However, in practice, it may not be easy to test whether our assumption is valid, especially when the variables in the causal graph are hard to extract directly from data as in the image data. We then conduct experiments on the synthetic data (Appendix A) where there is no clear *optimal* data split due to the significant noise level. We find that REIIL usually achieves the best test error. However, it may not recover the models close to the true ones. There also exists a model selection problem as the parameter tuning according to the validation data may make the model farther from the ground truth one. We refer more discussion to Appendix A and leave further exploration to future work.

What if there are more than two environments? For simplicity, we only investigate environment inference for two environments in the paper. We are aware that a good environment split is hard to define with multiple types of spurious correlations. It is unclear whether a binary partition is sufficient. We defer further investigation of these more complicated scenarios in invariant learning for the future.

6. Conclusion

In this paper, we propose a novel strategy to find a *more biased* reference model for environment inference. By repeating the EI step iteratively and training the reference model on the majority environment, our method helps find a better environment partition for downstream Invariant Learning tasks. Our assumption often holds in practice when most training samples exhibit strong spurious correlations and simple ERM methods fail to generalize. We conduct experiments on both CMNIST and our newly designed CBMNIST to demonstrate the effectiveness of our method.

References

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

Bui, M.-H., Tran, T., Tran, A., and Phung, D. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Chattopadhyay, P., Balaji, Y., and Hoffman, J. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pp. 301–318. Springer, 2020.

Chuang, C.-Y., Torralba, A., and Jegelka, S. Estimating generalization under distribution shifts via domain-invariant representations. *arXiv preprint arXiv:2007.03511*.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Ding, Z. and Fu, Y. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

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

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

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

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

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348. PMLR, 2020.

Kamath, P., Tangella, A., Sutherland, D., and Srebro, N. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2021.

Kim, D., Yoo, Y., Park, S., Kim, J., and Lee, J. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, B., Shen, Y., Wang, Y., Zhu, W., Reed, C. J., Zhang, J., Li, D., Keutzer, K., and Zhao, H. Invariant information bottleneck for domain generalization. *arXiv preprint arXiv:2106.06333*, 2021.

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

Liu, E. Z., Haghgoo, B., Chen, A. S., 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*, pp. 6781–6792. PMLR, 2021.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62 (3):54–60, 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.

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

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

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Shui, C., Wang, B., and Gagné, C. On the benefits of representation regularization in invariance based domain generalization. *Machine Learning*, pp. 1–21, 2022.

Subbaswamy, A., Chen, B., and Saria, S. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022.

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

Wang, R., Wu, Z., Weng, Z., Chen, J., Qi, G.-J., and Jiang, Y.-G. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022.

Ye, N., Tang, J., Deng, H., Zhou, X.-Y., Li, Q., Li, Z., Yang, G.-Z., and Zhu, Z. Adversarial invariant learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12441–12449. IEEE, 2021.

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.

Zhao, H., Dan, C., Aragam, B., Jaakkola, T. S., Gordon, G. J., and Ravikumar, P. Fundamental limits and tradeoffs in invariant representation learning. *arXiv preprint arXiv:2012.10713*, 2020.

A. Synthetic Data Experiments

We use all variants of the synthetic data defined in IRM (Arjovsky et al., 2019) as well as the synthetic data defined in (Creager et al., 2021) to test REIIL. The data is initially split into 3 environments with noises $0.2, 2$, and 5 respectively. In the previous implementations, they trained ERM on all three environments, trained IRM/EIIL only on the first two, and used the last as a validation set to pick the IRM regularizer. We change this setup by introducing a validation set explicitly using noise $= 3.5$ and treating the last environment with noise $= 5$ as the test set.

The IRM data experiment results can be summarized in Table 1. In almost all experiments, IRM finds the best solutions. Interestingly, REIIL almost always finds models with the lowest validation and test errors (always lower than EIIL), but that does not translate to the corresponding solutions having lower Causal and Non-Causal Errors compared with the ground truth solutions.

We find that the validation set in both previous and current settings had high noise (5 and 3.5), which is not representative of the real causal graph (which would have no noise). Therefore, we infer that selecting a model based on a lower validation error does not necessarily yield the best solutions. But in a learning setting with a fixed number of samples, finding a model with lower validation error is usually the best we can do. REIIL often achieves the lowest validation error, which is always lower than EIIL, suggesting that repeating the EI step improves performance. We found similar results with the synthetic data defined in EIIL, where REIIL usually found models with the lowest validation and test errors.

Discussion This questions the traditional methods of model and hyperparameter selection based on validation sets. In this case, the validation set was biased, and it is clear that the model which performs best for this set does not yield the causal solution. Even if the validation set was unbiased, it might not *generalize* to biased sets because the solution does not account for the noise explicitly.

Method/Type	FOU	FOS	FEU	FES	POU	POS	PEU	PES
ERM val	9.50 (0.26)	9.29 (0.23)	10.02 (0.40)	9.34 (0.47)	10.89 (0.46)	12.61 (1.28)	12.29 (2.23)	**11.68** (1.06)
ERM test	16.39 (0.76)	16.28 (1.14)	16.95 (1.13)	16.01 (1.04)	20.09 (0.95)	22.42 (2.00)	22.47 (5.43)	**20.45** (2.06)
IRM val	**5.11** (0.09)	7.66 (0.17)	14.84 (0.48)	13.96 (0.68)	**6.50** (0.57)	**8.74** (0.43)	16.16 (1.62)	16.26 (1.38)
IRM test	**5.50** (0.19)	12.52 (0.81)	27.12 (1.52)	26.18 (2.09)	**9.06** (1.06)	**13.67** (1.33)	30.23 (3.69)	30.22 (2.75)
EIIL val	10.38 (0.26)	7.57 (0.35)	14.01 (3.30)	18.32 (1.39)	10.07 (0.98)	12.38 (2.58)	19.01 (5.68)	18.78 (5.10)
EIIL test	17.67 (0.64)	9.51 (1.51)	26.06 (8.6)	32.75 (3.43)	18.06 (2.50)	19.89 (4.91)	35.37 (11.46)	35.92 (11.33)
REIIL val	7.17 (0.45)	**6.58** (0.88)	**8.97** (0.51)	**8.96** (0.33)	8.75 (0.49)	10.40 (1.91)	**11.81** (2.51)	11.72 (1.42)
REIIL test	10.77 (0.72)	**8.31** (1.96)	**14.72** (1.29)	**15.63** (1.36)	14.06 (1.50)	17.09 (4.13)	**21.41** (6.00)	20.86 (2.97)
ERM CE	0.11 (0.01)	0.12 (0.00)	0.44 (0.02)	0.45 (0.00)	0.12 (0.02)	0.14 (0.00)	0.44 (0.01)	0.47 (0.01)
ERM NCE	0.11 (0.00)	-	0.44 (0.01)	-	0.11 (0.01)	-	0.42 (0.02)	-
IRM CE	**0.01** (0.00)	0.09 (0.01)	**0.25** (0.01)	**0.29** (0.01)	**0.01** (0.00)	**0.04** (0.01)	**0.22** (0.02)	**0.29** (0.01)
IRM NCE	**0.01** (0.00)	-	**0.32** (0.00)	-	**0.01** (0.00)	-	**0.33** (0.01)	-
EIIL CE	0.12 (0.02)	**0.02** (0.00)	0.40 (0.05)	0.46 (0.02)	0.10 (0.03)	0.09 (0.03)	0.31 (0.10)	0.42 (0.05)
EIIL NCE	0.10 (0.02)	-	0.45 (0.01)	-	0.05 (0.04)	-	0.39 (0.08)	-
REIIL CE	0.05 (0.00)	**0.02** (0.00)	0.49 (0.04)	0.45 (0.02)	0.05 (0.04)	0.09 (0.02)	0.42 (0.06)	0.47 (0.02)
REIIL NCE	0.04 (0.01)	-	0.51 (0.02)	-	0.04 (0.03)	-	0.47 (0.05)	-

Table 1. Results of synthetic data experiments. val and test denote validation and test set errors of the model after training. CE and NCE denote Causal and Non-Causal errors. (standard deviation across 3 runs in brackets)