Invariant and Transportable Representations for Anti-Causal Domain Shifts

Anonymous Authors[1]

Abstract

Real-world classification problems must contend with domain shift, the (potential) mismatch between the domain where a model is deployed and the domain(s) where the training data was gathered. Methods to handle such problems must specify what structure is held in common between the domains and what is allowed to vary. A natural assumption is that causal (structural) relationships are invariant in all domains. Then, it is tempting to learn a predictor for label Y that depends only on its causal parents. However, many real-world problems are "anti-causal" in the sense that Y is a cause of the covariates X—in this case, Y has no causal parents and the naive causal invariance is useless. In this paper, we study representation learning under a particular notion of domain shift that both respects causal invariance and that naturally handles the "anti-causal" structure. We show how to leverage the shared causal structure of the domains to learn a representation that both admits an invariant predictor and that also allows fast adaptation in new domains. The key is to translate causal assumptions into learning principles that disentangle "invariant" and "non-stable" features. Experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed learning algorithm.

1. Introduction

This paper concerns the problem of domain shift in supervised learning, the phenomenon where a predictor with good performance in some (training) domains may have poor performance when deployed in a novel (test) domain. There are two goals when faced with domain shifts. First, we would like to learn a fixed predictor that is *domain-invariant* in the sense that it has good performance in all domains. Note, however, that even a good domain-invariant predictor may still be far from optimal in any given target domain. In such cases, we'd like to learn an optimal domain-specific predictor as quickly as possible. Then, the second goal is to learn a representation for our data that is *transportable* in the sense that, when given data from a new domain, we can use the representation to learn a domain-specific predictor using only a small number of examples.

Domain shifts plague real-world applications of machine learning and there is a large and active literature aimed at mitigating the problem (e.g., Arjovsky et al., 2019; Veitch et al., 2021; Peters et al., 2016; Rothenhäusler et al., 2021; Wang et al., 2021; Zhou et al., 2021; Koh et al., 2021; Zhuang et al., 2020; Cai et al., 2021; Shi et al., 2021; Sagawa et al., 2019; Bai et al., 2020; Subbaswamy et al., 2019; Zheng et al., 2021; Liu et al., 2021; Lu et al., 2021). Empirically, when domain shift methods are applied to wide-ranging benchmarks, there is no single dominant method—indeed, it's common for methods that work well in one context to do worse than naive empirical risk minimization (i.e., ignore the shift problem) in another context (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021). This problem is fundamental: it is impossible to build predictors that are robust to all possible kinds of shifts.[1] Accordingly, it is necessary to specify the manner in which the training and test domains are related to each other; that is, what structure is common to all domains, and what structure can vary across them. Then, to make progress on the domain shift problem, the task is to identify structural assumptions that are well matched to real-world problems and then find methods that can achieve domain-invariance and transportability under this structure.

In this paper, we rely on a particular variant of the assumption that structural causal relationships are invariant across domains, but certain "non-causal" relationships may vary. The motivation is that relationships fixed by the underlying dynamics of a system are the same regardless of the domain (Peters et al., 2017). A similar causal domain-structure assumption is already well-studied in the domain-shift literature (e.g., Peters et al., 2016; Arjovsky et al., 2019; Rothenhäusler et al., 2021; Rojas-Carulla et al., 2018; Muandet et al., 2013). There, the goal is to predict the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

[1]For any given predictor, it's possible to adversarially construct a domain where that predictor does poorly.

target label Y using only its causal parents (reconstructed from observed features X). In particular, the aim is to build predictors that do not rely on any part of the features that is causally affected by Y. However, in many problems, it can happen that the observed covariates X are all caused by Y, so that the causal parents of Y are the empty set. In this case, the naive causally invariant predictor is vacuous.

The purpose of this paper is to study an alternative causal notion of domain shift that handles such "anti-causal" (Y causes X) problems, and that maintains the interpretation that structural causal relationships are held fixed across all domains. Specifically,

1. We formalize the anti-causal domain shift assumption by adapting the notion of counterfactual invariance (Veitch et al., 2021).

2. We show how the causal domain shift assumption can be leveraged to identify an invariant predictor and transportable representation.

3. We use this identification as the basis of a concrete learning procedure for domain-invariant and domain-adaptive predictors.

4. We conduct several empirical studies, finding that the procedure can effectively learn invariant structures, learn fast-adapting structures, and disentangle the factors of variation that vary across domains.

2. Causal Setup

The first step is to make precise what structure is preserved across domains, and how they are allowed to vary. Once we have this, we'll make the notions of invariant and fast-adapting predictor precise.

In each domain e, we have observed data $(X_i, Y_i) \overset{\text{iid}}{\sim} P^e$, where P^e is a domain-specific data-generating distribution. We will mainly consider problems where we have (finite) datasets sampled from multiple distinct domains at training time, and wish to make predictions on data sampled from some additional domains not observed during training. This paper discusses classification problems where Y_i is discrete.

To formalize the causal structure assumption, we'll introduce two additional latent variables. First, Z, a (subset of) the causes of X. Second, U, is a confounding variable that affects both Z and Y. See Figure 1. Conceptually, Z are the factors of variation where the association with Y can vary across domains. The confounder U is the reason the association can vary. The relationship between Y and Z induced by U needs not be stable across domains. We do not introduce explicit notation for the factors of variation that have invariant relationships with Y. We will see these

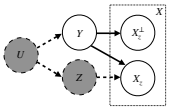


Figure 1: Causal model for the data generating process. We decompose the observed covariate X into latent parts defined by their causal relationships with Z. Solid circles denote observed variables, while shaded circles denote hidden variables.

can be naturally handled implicitly. Slightly abusing notation, we assume that $Z_i, U_i, X_i, Y_i \overset{\text{iid}}{\sim} P^e$ in each domain e. Both Z and U are unobserved. We can now describe the set of domain shifts we consider.

To formalize the notion of invariance, we'll use the concept of *counterfactual invariance* to Z (Veitch et al., 2021). A function f is counterfactually invariant to Z if $f(X(z)) = f(X(z'))$ for all z, z', where $X(z)$ denotes the counterfactual X we would see had Z been z. Learning a predictor that does not depend on the factors of variation Z that induce unstable relationships means learning a predictor that is counterfactually invariant to Z.

To go further in our formalization, we'll need another idea from Veitch et al. (2021): the decomposition of X into (latent) parts defined by their causal relationship with Z. We define $X_z^\perp$ to be the part of X that is not causally affected by Z. More precisely, $X_z^\perp$ is the part of X such that any function of X is counterfactually invariant if and only if it is a function of $X_z^\perp$ alone (that is, $f(X)$ is $X_z^\perp$ measurable). Under weak conditions on Z, $X_z^\perp$ is well defined (e.g., discrete Z suffices) (Veitch et al., 2021). We also introduce X_z to denote the part of X that is not invariant to Z. We make the extra assumption that $X_z^\perp$ does not have a causal effect on X_z (the other direction is ruled out by the definition of $X_z^\perp$).

The optimal counterfactually invariant predictor $g(X)$ is unlikely to be the best predictor in any given domain. The reason is that it excludes Z-dependent information that may in fact be highly predictive in a given domain. Given a new domain e, we would like to be able to quickly learn a new predictor $f^e(X)$ that updates the invariant predictor with domain-specific associations. This update should only depend on X_z because the relation between $X_z^\perp$ and Y is stable. Accordingly, we want to learn a representation $h(X)$ that encapsulates the information in X_z. Moreover, this should be done in a manner such that, given $g(X)$ and $h(X)$, we can learn a good predictor for P^e with only a small number of samples. To formalize this, we'll introduce the following domain-specific predictors: $f^e(X) = g(X) + M^e h(X)$. Here $f^e(X)$ is logits.

3. Observable Signature

The first problem we must confront is how to learn a function $g(X)$ that depends on $X_z^\perp$ alone, and $h(X)$ that depends on X_z alone. Strictly speaking, learning such functions precisely would be impossible, even if we observed Z (Veitch et al., 2021). The reason is that we have access to only observational data, but the two parts of X are defined in terms of the underlying causal structure.

The key observation is that there are two relations that connect $g(X)$ and $h(X)$. The first relation comes from the causal graph. In particular, we want to impose the requirement that $g(X)$ and $h(X)$ satisfy the observable implications of the causal structure. The next theorem gives such an observable implication, which can serve as an observable signature of the causal decomposition.

Theorem 3.1. *If $g(X)$ depends only on $X_z^\perp$ and $h(X)$ depends only on X_z , then, under the causal graph in Figure 1, $g(X) \perp\!\!\!\perp h(X) \mid Y$.*

The usefulness of this theorem is that the conditional independence statement can be measured from data, and enforced in the model training.

The second relation is subtler and it comes from our formulation of domain-specific predictors f^e. Specifically, f^e, as a linear combination of g and h, should minimize the risk in every domain. And by only allowing weights of h to change, we hope g would capture "invariant" information ($X_z^\perp$) and h would learn "unstable" information (X_z).

3.1. Causal Regularization

We enforce $g(X)$ and $h(X)$ to satisfy the conditional independence condition via regularization. Specifically, we want to define a regularizer such that its value goes to zero whenever the conditional independence requirement is met. In general, measuring conditional independence can be hard (Zhang et al., 2012; Fukumizu et al., 2007; Tolstikhin et al., 2016). Instead, we look at a weaker condition that uses the following fact: If $A \perp\!\!\!\perp B \mid D$, then, $\mathbb{E}[A \cdot (B - \mathbb{E}[B|D])] = 0$

This is a necessary but not sufficient condition for conditional independence. However, it is easy to compute and leads to good results in practice (as shown in Section 5). With this identity in hand, we define $C_{\text{cond}}(A, B, D)$, the (infinite data) conditional independent regularization term between random variables A, B given random variable D, and its empirical estimate $\hat{C}_{\text{cond}}$ as follows:

$$C_{\text{cond}}(A, B, D) = \mathbb{E}[A \cdot (B - \mathbb{E}[B|D])]$$

$$\hat{C}_{\text{cond}}(\{(a_i, b_i, d_i)\}_{i=1}^n) = \left\| \frac{1}{n} \sum_i a_i \left(b_i - \frac{1}{|\#j : d_j = d_i|} \sum_{j:d_j=d_i} b_i \right) \right\|$$

where $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$, $\{d_i\}_{i=1}^n$ are samples of A, B and D. Here, the conditional random variable D is assumed to be discrete which is true for the use case in this paper.

4. Learning Algorithm

In principle, we could learn two completely separate functions $g(X)$ and $h(X)$. However, this can be wasteful. For instance, in vision, many low-level features can be reused by different predictors. To address this, we first notice that we can always rewrite the domain-specific classifer as follows

$$f^e(X) = (W^b + W^e)\Phi(X). \tag{1}$$

That is, as a shared representation Φ followed by a fixed linear map W^b defining g and a domain-specific linear map W^e defining $M^e h(X)$.[2] The task is then learning the representation (which we'll parameterize by a neural network), and the invariant and domain-specific linear maps.

In fact, a further simplification is possible: we can fix W^b to be $\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. The reason is that, because Φ is unconstrained, learning W^b doesn't actually add expressive power—any non-zero map suffices.

4.1. Bi-Level Optimization

We have now reduced our task to a bi-level optimization problem.

$$\boxed{\begin{aligned} &\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} \gamma R^e((W^b + W^e)\Phi) + (1-\gamma) R^e(W^b \Phi) \\ &\quad \text{st } W^e \in \arg\min_{W} R^e((W^b + W)\Phi) + \lambda C_{\text{cond}}(W^b\Phi, W\Phi, Y) \end{aligned}} \tag{2}$$

where $\gamma \in [0, 1]$, $\lambda > 0$. The set $\mathcal{E}_{tr}$ consists of all training domains, and $R^e(f)$ is the domain-specific population risk defined as $R^e(f) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}^e}[\ell(f(\mathbf{x}), \mathbf{y})]$ with the cross-entropy loss function ℓ.

Equation (2) is a challenging optimization problem. In general, each constraint calls for an inner optimization routine. So instead of solving Equation (2) directly, we propose to use gradient penalty to make the problem more tractable. Specifically, we translate the condition that the domain-specific risk is optimal (the inner loop) into the condition that the gradient of the domain-specific risk with respect to W^e is 0. Then, we regularize the ℓ_2-norm of this gradient. This is inspired by a similar trick used in Invariant Risk Minimization (Arjovsky et al., 2019).

[2] Consider $\Phi(X) = [g(X)^T, h(X)^T]^T$, $W^b = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and $W^e = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & M^e \end{bmatrix}$.

5. Experiments

The main claims of the paper are:

1. The invariant predictor $g(X)$ will have good performance in new domains, so long as the shifts obey the anti-causal structure.

2. The learned representation Φ enables rapid adaptation to new domains by learning only a linear adjustment term on top of Φ.

3. The learned Φ disentangles the parts of X that are not affected by Z from the parts that are.

To evaluate the above claims, we conduct experiments on synthetic and real-world data. While causal structures of real-world problems like image classification are usually unknown, we find that the anti-causal based method works well on many such problems—suggesting the anti-causal structure is appropriate.

Baselines For each experiment, all methods share a common architecture; they differ only in objective functions or optimization procedures. For invariant learning, we compare with empirical risk minimization (ERM), IRM (Arjovsky et al., 2019) and the MAML (Finn et al., 2017) base learner. To test how well learned representations Φ can enable fast adaptation, we fine-tune linear models on top of the representation. For comparison, we also fine-tune linear layers on top of the representations (penultimate layers) from ERM, IRM, and MAML. In fact, it has been shown recently that fine-tuning the last layer of models trained by ERM has surprisingly good performance on many real-world datasets (Rosenfeld et al., 2022). For MAML, the last layer is trained using the MAML update rule.

5.1. Synthetic Dataset

We generate synthetic data according to the following structural equations (which obey the anti-causal structure):

$$Y \leftarrow \mathrm{Rad}(0.5) \quad X_z^{\perp} \leftarrow Y \cdot \mathrm{Rad}(0.75)$$
$$Z \leftarrow Y \cdot \mathrm{Rad}(\beta_e) \quad X_z \leftarrow Z$$

where input X is $(X_z, X_z^{\perp})$ and $\mathrm{Rad}(\beta)$ means that a random variable is -1 with probability $1-\beta$ and $+1$ with probability β. We create two training domains with $\beta_e \in \{0.95, 0.7\}$, one validation domain with $\beta_e = 0.6$ and one test domain with $\beta_e = 0.1$. Prediction with $X_z^{\perp}$ is stable but has a lower accuracy compared to prediction with X_z during training. If a learning model only chooses the classifier with the best prediction accuracy in training domains and ignores its instability, it will choose X_z as its predictor and end up with only 10% accuracy on the test set. The robust predictor would be $X_z^{\perp}$ with 75% accuracy. On the other hand, in the test domain, $-X_z$ predicts Y with 90% accuracy—so an adaptive predictor is better than the invariant one.

We use a three-layer neural network with hidden size 8 and ReLU activation for Φ and train the neural network with Adam optimizer. The hyperparameters are chosen based on performance on the validation set. For the fine-tuning test, we run 20 steps with a learning rate 10^{-2}. The result is shown in Table 1. Both IRM and the anti-causal model learn good invariant predictors. But the causal model is also equipped with the ability to adapt given a very small amount of data points while the performance of IRM stays the same after fine-tuning. Perhaps unsurprisingly, ERM has a test accuracy of 10%, suggesting that it uses only spurious features X_z.

5.2. Color MNIST

Color MNIST (Arjovsky et al., 2019) modifies the original MNIST dataset. First, we assign label 0 to digits 0-4 and label 1 to digits 5-9. We then flip the label with probability 25% and assign colors to the original images based on the label but with a flip rate $1-\beta^e$. That is, we assign color 0 to images with label 0 with probability β^e. We create two training domains with $\beta^e \in \{0.95, 0.7\}$, a validation domain with $\beta^e = 0.2$ and a test domain with $\beta^e = 0.1$. Ideally, we want to learn invariant predictors based on the shape of the digit—this will achieve 75% accuracy. But the problem is significantly more challenging than the synthetic example because the color is a much easier feature to learn than the shape of the digit, making models more susceptible to spurious correlations. We use a three-layer convolutional neural network for Φ and train the neural network with Adam optimizer. The hyperparameters are chosen based on performance on the validation set. For the fine-tuning test, we run 20 steps with a learning rate 10^{-2}. The result is shown in Table 1. The causal adaptive model learns both the invariant and adaptive structure significantly better than reference baselines.

Causal Disentanglement To understand why the causal model can adapt to test domains in both synthetic and Color MNIST datasets, we plot distributions of activation values of Φ. See Figure 2. We see that the first two coordinates—used as the invariant part of the representation in training—have distributions that do not depend on the value of Z. On the other hand, some other (non-invariant) representation coordinates have activations that change dramatically depending on the value of Z. Thus, the representation effectively disentangles the $X_z^{\perp}$ and X_z features. Importantly, this is achieved with no a priori knowledge of what Z might be, and no observations of it.

References

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020.

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

Bai, H., Sun, R., Hong, L., Zhou, F., Ye, N., Ye, H.-J., Chan, S.-H. G., and Li, Z. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. *arXiv preprint arXiv:2012.09382*, 2020.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Cai, T., Gao, R., Lee, J. D., and Lei, Q. A theory of label propagation for subpopulation shift. *arXiv preprint arXiv:2102.11203*, 2021.

Chen, Y. and BÃ¼hlmann, P. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22(261):1–80, 2021. URL http://jmlr.org/papers/v22/20-1227.html.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, volume 20, pp. 489–496, 2007.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

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.

Jiang, Y. and Verma, N. Meta-learning to cluster. *arXiv preprint arXiv:1910.14134*, 2019.

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.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Lake, B., Salakhutdinov, R., Gross, J., and Tenenbaum, J. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., and Liu, T.-Y. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

Magliacane, S., Van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. Representation learning via invariant causal mechanisms. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=9p2ekP904Rs.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 10–18. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/muandet13.html.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

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)*, pp. 947–1012, 2016.

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

Rojas-Carulla, M., Schölkopf, B., Turner, R. E., and Peters, J. Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, 19:36:1–36:34, 2018. URL http://jmlr.org/papers/v19/16-432.html.

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

Rosenfeld, E., Ravikumar, P., and Risteski, A. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization, 2022. URL https://arxiv.org/abs/2202.06856.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 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. *arXiv preprint arXiv:1911.08731*, 2019.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/625.pdf.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

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

Tolstikhin, I. O., Sriperumbudur, B. K., and Schölkopf, B. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

Wang, J., Lan, C., Liu, C., Ouyang, Y., and Qin, T. Generalizing to unseen domains: A survey on domain generalization. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 4627–4635. ijcai.org, 2021. doi: 10.24963/ijcai.2021/628. URL https://doi.org/10.24963/ijcai.2021/628.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

Wang, Y. and Jordan, M. I. Desiderata for representation learning: A causal perspective, 2021. URL https://arxiv.org/abs/2109.03795.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827. PMLR, 2013.

Zheng, X., Sun, X., Chen, W., and Liu, T.-Y. Causally invariant predictor with shift-robustness. *arXiv preprint arXiv:2107.01876*, 2021.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. *CoRR*, abs/2103.02503, 2021. URL https://arxiv.org/abs/2103.02503.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A. Related Work

Invariant Learning Several papers draw a connection between causality and invariant prediction (e.g., Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2019; Lu et al., 2021; Wang & Jordan, 2021). These papers assume that all domains share a common causal structure, and consider the set of domains induced by arbitrary intervention on any node other than the label Y. In this case, the predictor that has invariant risk across domains is the one that depends only on the causal parents of Y. By contrast, in this paper, we only allow changes of unobserved confounders—resulting in a much smaller set of possible shifts. This weakening allows the invariant predictor to learn to depend on the non-parent parts of the covariates (descendants of Y).

A closely related work is "Invariant Risk Minimization" (Arjovsky et al., 2019), which also seeks to learn a representation Φ of X such that a fixed linear map on top of the representation yields an invariant predictor. The major distinction with the approach here is that we rely on simultaneously learning the non-stable factors of variation in order to identify the representation. How well IRM actually learns invariant predictors in practice is a subject of some debate (Kamath et al., 2021; Rosenfeld et al., 2020; Ahuja et al., 2020). Beyond the anti-causal setup, the practical approach here may circumvent some of these issues.

Other papers also consider settings where the covariates X are not direct causes of Y (Liu et al., 2021; Mitrovic et al., 2021; Ilse et al., 2020). They assume that both X and Y are caused by latent variables that can be divided into stable and non-stable parts. Then, they use generative methods reflecting this assumption. By contrast, the approach in this paper is fully nonparametric—there is no explicit modeling of the generative process. Schölkopf et al. (2012) study the anti-causal direction for two variables that rely heavily on the additive noise model. In this paper, we focus on a different counterfactual implication.

Veitch et al. (2021) introduce the notion of counterfactual invariance to a spurious factor and make some connections with domain shifts. However, they assume Z is known in advance and observed, and rely on this to learn the counterfactually-invariant predictor. In contrast, in this paper we merely assume the existence of some Z—we don't need to know it in advance, and we don't need to measure it directly. However, they use data from only a single domain, whereas we require observations from many distinct domains. We also treat the problem of learning transportable representations, whereas they only handle invariant learning.

Domain Adaptation and Meta Learning There have been numerous fruitful developments in the fields of domain generalization and adaptation (e.g., Zhou et al., 2021; Wang et al., 2021), including ones under various causal assumptions (Zhang et al., 2013; Magliacane et al., 2018; Chen & BÃ¼hlmann, 2021; Subbaswamy et al., 2019; Schölkopf et al., 2012). Our work differs from other domain adaptation methods in that we consider the interplay between both the problem of invariant/robust learning and adaptation. And we are doing so by learning a common representation and training a linear classifier on top to distinguish the two parts.

The adaptive part of the learning model in this paper is also related to meta learning, where the goal is to learn predictors that can quickly adapt to new tasks. Meta learning has been used for supervised learning (Santoro et al., 2016), reinforcement learning (Wang et al., 2016) and even unsupervised learning (Jiang & Verma, 2019). Traditional approaches to meta learning include defining a distribution over the structure of input data to perform inference (Lake et al., 2011) or to use a memory model such as LSTM (Wang et al., 2018). But the dominant models for meta learning are generic gradient-based learning methods such as MAML (Finn et al., 2017) and Reptile (Nichol et al., 2018). Theoretically, Tripuraneni et al. (2021) and Du et al. (2020) also examine the representation power of meta-learning. Although not motivated by causality, they show that if there is a shared common structure, meta learning can be used to reduce sample complexity in unseen domains.

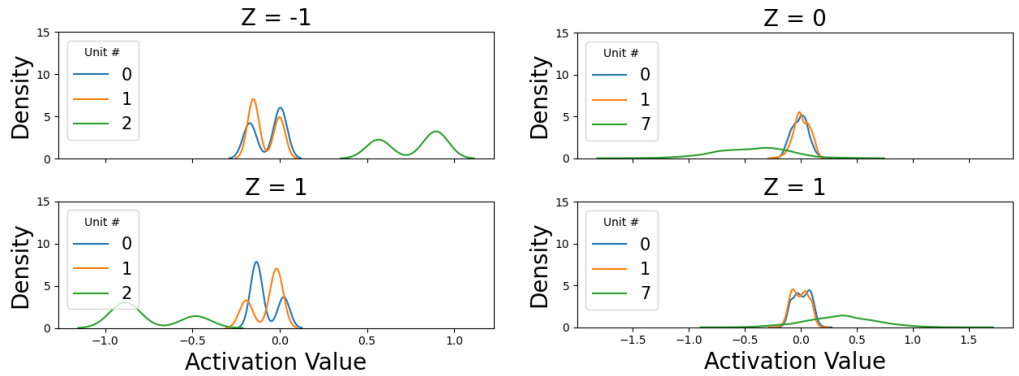
B. Proofs

Theorem 3.1. *If $g(X)$ depends only on $X_z^{\perp}$ and $h(X)$ depends only on X_z , then, under the causal graph in Figure 1, $g(X) \perp\!\!\!\perp h(X) \mid Y$.*

Proof. Reading d-separation from the causal graphs. We have that $X_z \perp\!\!\!\perp X_z^{\perp} \mid Y$. $g(\cdot)$ is a function of $X_z^{\perp}$ and $h(\cdot)$ is a function of X_z. Because function of independent variables are also independent, the theorem is proven. □

Table 1: The causal adaptive model has good invariant and adaptive performance on synthetic datasets and Color MNIST, and it outperforms baseline methods on Camelyon17 for invariant prediction. The table shows accuracy (%). Note that adaptation (n) means that the predictor is tuned on n number of labeled examples. For synthetic datasets, standard errors are over 100 runs. For Color MNIST, it is over 50 runs. For Camelyon17, it is over 5 runs. In each run, the adaptive accuracy is determined by the average of 100 fine-tuning tests for both synthetic datasets and Color MNIST.

	Synthetic Dataset		Color MNIST		Camelyon17
Method	Test Acc.	Adaptation (10)	Test Acc.	Adaptation (10)	Test Acc.
ERM	9.95±0.10	11.57±0.71	28.24±0.51	27.26±0.48	70.77±1.98
IRM	74.91±0.13	74.27±0.47	59.97±0.91	60.16±0.90	71.59±2.76
MAML	17.14±2.22	44.01±3.48	22.18±1.01	75.03±3.30	70.22±2.40
CAUSAL	74.77±0.44	**89.28±0.25**	**70.30±0.71**	**85.25±1.11**	**77.73±1.74**



(a) Synthetic Dataset

(b) Color MNIST

Figure 2: The learned representation Φ disentangles stable and unstable factors of variation. Plots show activation levels of different representation units. Units 0 and 1 are trained as the invariant part of the representation. The invariant units have no dependency on the unstable factor Z, but the other unit has a strong dependency. For synthetic dataset, Z is the random variable defined in the structural equations. For Color MNIST, Z is the color.

C. Additional Experiment Results

C.1. Camelyon17

The goal of the Camelyon17 dataset (Bandi et al., 2018) is to predict the existence of a tumor given a region of tissue. This is a binary classification problem. Data are collected from a small number of hospitals. But there are variations in data collection and processing that could negatively impact model performance on data from a new hospital. We take the individual hospitals to be separate domains. The objective is to generalize to new hospitals not seen in training. The dataset consists of input images with size 96×96 and a binary label that indicates if the central 32×32 region contains any tumor tissue. The dataset can be divided into 5 subsets, each from a different hospital. Following the WILDS benchmark (Koh et al., 2021), we use 3 for training, 1 for validation, and the last one for test.

We use a pre-trained ResNet-18 model for our Φ and train the whole model using Adam optimizer with a learning rate 10^{-4}. For the fine-tuning test, we run 20 iterations with a learning rate 10^{-2}. As shown in Table 1, the causal model has the best invariant accuracy. For adaptive performance, Figure 3 shows that the causal model has large performance improvement when given a small fine-tuning dataset, while other models require more fine-tuning examples to see a significant increase in accuracy.

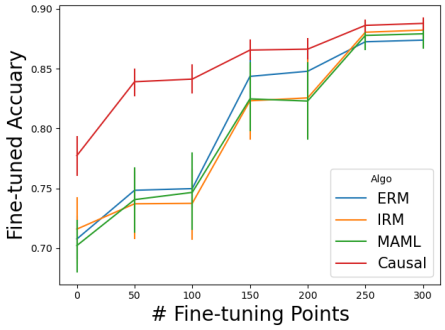


Figure 3: The causal model has good accuracy after fine-tuning with small datasets on Camelyon17. Graph shows accuracy (%) on Camelyon17 for different numbers of fine-tuning examples. Standard errors are over 5 runs. For each run, the accuracy is averaged for 100 fine-tuning tests.