Adapting to Shifts in Latent Confounders using Observed Concepts and Proxies

Matt J. Kusner [1 2] **Ibrahim Alabdulmohsin** [1] **Stephen Pfohl** [1] **Olawale Salaudeen** [1] **Arthur Gretton** [2]
Sanmi Koyejo [1 *] **Jessica Schrouff** [1 *] **Alexander D'Amour** [1 *]

Abstract

We address the problem of unsupervised domain adaptation when the source differs from the target because of a shift in the distribution of a latent confounder. In this case, neither covariate shift nor label shift assumptions apply. When all data is discrete, we show that the optimal target predictor can be non-parametrically identified with the help of concept and proxy variables, available only in the source, and unlabeled data from the target.

1. Introduction

Distribution shift is a fact of many real-world machine learning systems. For example, imagine we have trained a disease prediction model on patients of Hospital P and would like to apply it to patients of Hospital Q. However, these hospitals differ in their patient populations along socioeconomic, demographic, and other axes (Finlayson et al., 2021). How can we find the optimal predictor for Hospital Q, given only labelled data from Hospital P and unlabelled data from Hospital Q? This is the problem of unsupervised domain adaptation (Huang et al., 2006). Without any assumptions on the shift, this question is impossible to answer: the mapping from features X to labels Y could differ across hospitals in arbitrary ways. To address this, one of the most popular assumptions placed on distribution shift is to localize the shift between distributions P and Q in the features (covariates) X, i.e., *covariate shift*: $p(X) \neq q(X)$. There has been a large body of work devoted to estimating predictors for Q under this setting (Shimodaira, 2000; Zadrozny, 2004; Huang et al., 2006; Gretton et al., 2009; Bickel et al., 2009; Chen et al., 2016; Schneider et al., 2020). The key observation is that under this assumption $p(Y \mid X) = q(Y \mid X)$. Therefore, if one makes the source data appear like the target data (e.g., by reweighing the source classifier loss by $q(X)/p(X)$), one can learn an accurate target classifier.

A similar assumption is to localize the shift in the labels Y, i.e., *label shift*: $p(Y) \neq q(Y)$, $p(X \mid Y) = q(X \mid Y)$ (Gart & Buck, 1966; Manski & Lerman, 1977; Rosenbaum & Rubin, 1983; Saerens et al., 2002; Forman, 2008; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Garg et al., 2020; Tachet des Combes et al., 2020). Here one can use a similar approach: learn $q(Y)/p(Y)$ and use it to reweigh a source classifier, adapting it to the target distribution. The assumptions of covariate and label shift can be framed as criteria on the causal structure of the data, shown in Figure 1(a)-(b) (Schölkopf et al., 2012).

However, these assumptions are often overly restrictive for real-world settings, as the shifts encountered are typically more complex (i.e., 'compound' shifts) (Schrouff et al., 2022). Consider the Hospital example where our goal is to learn a mapping from patient electronic health record (EHR) data X to disease risk Y. In order to adapt a predictor in source P to target Q we need to take into account that often $p(Y \mid X) \neq q(Y \mid X)$ and $p(X \mid Y) \neq q(X \mid Y)$. For example, social determinants of health (SDH) (Marmot & Wilkinson, 2005) (e.g., income, education, discrimination, and other societal factors) affects both how one is diagnosed, changing $p(Y \mid X)$, and how often one can visit the hospital for treatment, changing $p(X \mid Y)$.

In this work, we introduce a distribution shift assumption that generalizes both covariate and label shift, allowing for shifts in the marginal distributions of both X and Y. Specifically, the shift from P to Q is located in a latent variable U, which we call *latent shift*: $p(U) \neq q(U)$. This latent U influences, or confounds, all observed variables, and so shifts all observable distributions from P to Q (e.g., in the Hospital example, U could be income level). However, without access to additional observed data, identifying the optimal $q(Y|X)$ is impossible (for more details see Section 3). In order to make progress, we will leverage additional data in the source domain, a strategy inspired by recent work (Arjovsky et al., 2019; Koh et al., 2020).

Our key insight is that we can frame learning the optimal $q(Y|X)$ as a non-parametric identification problem. This suggests a strategy: we can use techniques from causal inference literature that were built to identify the interventional distribution $p(Y|do(X))$, to instead identify

[*] Equal contribution [1] Google Research [2] University College London. Correspondence to: Matt J. Kusner <matt.kusner@gmail.com>.

Presented at ICML Workshop on Principles of Distribution Shift 2022,

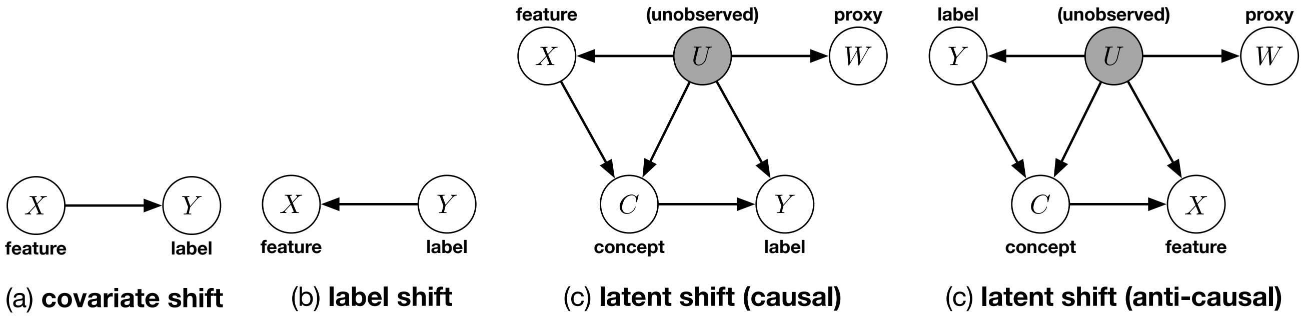

Figure 1. Different domain adaptation assumptions: (a) covariate shift $p(X) \neq q(X)$, (b) label shift $p(Y) \neq q(Y)$, (c)-(d) latent shift $p(U) \neq q(U)$.

$q(Y|X)$. This is non-trivial, as identifying $q(Y|X)$ is harder than $p(Y|do(X))$.[1] Our identification strategy is constructive, and immediately suggests an algorithm for estimating $q(Y|X)$ when data is discrete. We describe why additional data is sufficient for identification, we validate our approach numerically, and we detail useful future directions.

2. Method

Let $X \in [k_X]$ be discrete features (i.e., with k_X possible categories) and $Y \in [k_Y]$ be discrete labels. Let P be the source distribution and Q be the target, with probability mass functions p, q. Let $p(X=i), q(X=i)$ be shorthand for $\mathbb{P}_P(X=i), \mathbb{P}_Q(X=i)$, where $\mathbb{P}$ denotes probability.

Consider the following learning setup: we observe training data $(x_1, y_1), \ldots, (x_n, y_n)$ from the source distribution P. We also observe samples $(x'_1), \ldots, (x'_m)$ drawn from the target distribution Q. In this domain adaptation scenario, our goal is to find the optimal predictor for y' drawn from Q, i.e., $q(Y|X)$. However, without assumptions, this is impossible. We propose the following conditions to generalize covariate and label shift.

A1. *Alongside X, Y we also observe $C \in [k_C]$ and $W \in [k_W]$. All data is generated by the process described in either Figure 1(c) or (d). $U \in [k_U]$ is an unobserved latent variable. Finally all data is* faithful *and* Markov *(Spirtes et al., 2000) (i.e., conditional independences in the data exist iff they exist in the graph).*

Formally, the directed acyclic graphs (DAGs) described in Figure 1 and Assumption A1 are probabilistic graphical models (Pearl, 1988) that describe the data generation process. Specifically, each node V is a random variable, and each edge describes a (unknown) function f_V mapping parent nodes $\text{pa}(V)$ to child nodes V, i.e., $V = f_V(\text{pa}(V))$. These models encode conditional independencies that can be derived via d-separation (Pearl et al., 2000).

These graphs are inspired by two lines of work. The first is on learning with *concepts* (Kumar et al., 2009; Lampert et al., 2009; Koh et al., 2020). Concepts C are high-level, often interpretable, pieces of information that mediate the relationship between X and Y. In many settings, such as healthcare, concepts are readily available (Koh et al., 2020). Continuing our earlier example where X is raw EHR data (e.g., temperature, blood cultures, ...) and Y is disease risk, C could be physician summaries such as the presence and spread of infection. The second is on causal effect estimation with proxy variables (Kuroki & Pearl, 2014; Miao et al., 2018): W is a proxy of U that allows one to identify the causal effect of C on Y (i.e., in the causal setting in Figure 1(c)). In our running example, a useful W is the region where a patient lives, a proxy for income level U. We describe why W and C are needed in Section 3.

A2. *The shift between P and Q is located in U, i.e., there is* latent shift $p(U) \neq q(U)$.

This assumption describes how the difference between P and Q arises: distributions on U or that have U marginalized (i.e., all observed distributions) will shift between P and Q. Whereas, distributions conditional on U do not shift. In particular this setting violates the covariate and label shift invariances: $p(Y|X) = q(Y|X)$ and $p(X|Y) = q(X|Y)$.

A3. *We have that $k_X, k_W \geq k_U$.*

The above is necessary because we need to place some restriction on the ability of U to influence W, X, C, Y. This is more generic than placing a functional restriction on U, all we require is that the support of U is smaller than that of the observed variables.

A4. *For estimation we require that: (i) for every $i \in U$ where $q(i) > 0$ we have that $p(i) > 0$, and (ii) all matrices are assumed to be full rank.*

The above is required for our identification procedure. The first condition ensures that $q(U=i)/p(U=i)$ is well-defined. Overall these assumptions are of two types: (1) **Structural**: Assumptions 1, 2 describe how the data and shifts are structured; (2) **Functional**: Assumptions 3, 4 detail conditions on the functions that generate data.

[1] Identifying $p(Y|do(X))$ only requires identifying a specific target distribution where $q(U|X) = p(U)$ (the one that does not confound X, Y).

Our goal is to estimate the target distribution $q(Y|X)$ given only $(w_1, x_1, c_1, y_1), \dots (w_n, x_n, c_n, y_n)$ from the source distribution P and $x'_1, \dots x'_m$ from the target. Given A1, Figure 1 (c) implies the following decomposition:

$$
\begin{aligned}
q(Y|X) &\overset{(1)}{=} \sum_{i=1}^{k_U} q(Y|X, U=i) q(U=i|X) \qquad (1)\\
&\overset{(2)}{=} \sum_{i=1}^{k_U} p(Y|X, U=i) \frac{q(X|U=i)q(U=i)}{q(X)}\\
&\overset{(3)}{=} \sum_{i=1}^{k_U} p(Y|X, U=i) \frac{p(U=i|X)p(X)q(U=i)}{p(U=i)q(X)}\\
&\overset{(4)}{\propto} \sum_{i=1}^{k_U} p(Y|X, U=i) p(U=i|X) \frac{q(U=i)}{p(U=i)}
\end{aligned}
$$

The first equality (1) is given by the chain rule and marginalization. The second (2) is given by A2: since $q(Y|X, U = i)$ conditions on U we have $q(Y|X, U = i) = p(Y|X, U = i)$. The fractional term is given by Bayes rule. The equality (3) is again given by A2 and Bayes rule: $q(X|U = i) = p(X|U = i) = p(U = i|X)p(X)/p(U = i)$. The proportional (4) is given by the fact that $p(X)/q(X)$ is constant as we are conditioning on these variables on the left-hand side. Therefore, if we can estimate the quantities on the right-hand side, we can estimate $q(Y|X)$ by summing these quantities across U and normalizing. This is non-trivial as we do not observe U. Our main result is the following.

Theorem 1. *Given that the above assumptions hold, all probability mass functions over $W, X, C, Y, \tilde{U}$ in the source P are identifiable, where $\tilde{U}$ is an unknown sorting of U.*

We give a full proof in the Appendix and give a sketch here. The first key observation is that identifying distributions on $\tilde{U}$ is all that is needed, as (a) it satisfies the same independence conditions as U, and (b) our quantity of interest $q(Y|X)$ only requires marginalizing over U, making the order of the categories of U irrelevant to identification. The proof works in two stages: 1. It first demonstrates that $p(W|\tilde{U})$ can be identified, and 2. Shows that once $p(W|\tilde{U})$ is identified, all distributions on $W, X, C, Y, \tilde{U}$ are identified. Stage 1 is done by proving a variation of a result given by Kuroki & Pearl (2014). They demonstrate that when $k_W = k_X = k_U$ and data is generated from the graph of Figure 1(c) then it is possible to identify the causal effect $p(Y|do(C))$ (in Theorem 1 (Kuroki & Pearl, 2014)). Identifying $p(Y|do(C))$ only requires identifying specific distributions involving $\tilde{U}$, in order to remove its contribution to Y, i.e., $p(Y|do(C)) = \sum_{x,u} P(Y|C, X = x, \tilde{U} = u)P(X = x, \tilde{U} = u)$. However, we show by construction, that the result of Kuroki & Pearl (2014) is stronger. First, it recovers $p(W|\tilde{U})$ for Figure 1(c) or Figure 1(d). Second, it allows for identification of all distributions involving $\tilde{U}$.

Algorithm 1 Estimating $q(Y|X)$.

```
Require: source $\mathcal{P} = \{(w_i, x_i, c_i, y_i)\}_{i=1}^n$; target $\mathcal{Q} = \{x_j\}_{j=1}^m$;
    For any variables $G \in [k_G]$, $H \in [k_H]$ let $p(\mathbf{G}|\mathbf{H})$ be a $k_G \times$
    $k_H$ matrix of probabilities s.t. $p(\mathbf{G}|\mathbf{H})_{ij} = p(G=i|H=j)$
Using $\mathcal{P}$, form matrices $\mathbf{A}, \mathbf{B}$ described in eq. (2)
Eigendecompose $\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}^{-1}\Delta\mathbf{S}$ to get $p(W|\tilde{U})$ from $\mathbf{S}^{-1}$
Estimate $p(\tilde{\mathbf{U}}|\mathbf{X}) = p(\mathbf{W}|\tilde{\mathbf{U}})^{-1} p(\mathbf{W}|\mathbf{X})$
Estimate $q(\tilde{\mathbf{U}})/p(\tilde{\mathbf{U}}) = p(\tilde{\mathbf{U}}|\mathbf{X})^{-1}[q(\mathbf{X})/p(\mathbf{X})]$
Estimate $p(\mathbf{Y}|X, \tilde{\mathbf{U}}) = p(\mathbf{Y}|X, \mathbf{W}) \left(\frac{p(\mathbf{W}|\tilde{\mathbf{U}}) \circ p(\tilde{\mathbf{U}}|X)}{p(\mathbf{W}|X)}\right)^{-1}$
Estimate $q(Y|X)$ for any $x_j \in \mathcal{Q}$ via eq. (1)
```

The key observation behind this second result (i.e., Stage 2) is that conditioning on $\tilde{U}$ d-separates W from the rest of the observed variables. This means that factorizing observed distributions using $\tilde{U}, W$ can form linear systems. In these systems, the unknown distributions involving $\tilde{U}$ can be recovered by some function of $p(W|\tilde{U})$ (identified in Stage 1) and observable distributions. Because this result does not rely on the ordering of $\{X, C, Y\}$ in the graph, it applies to both Figures 1(c) and 1(d) (more details in Appendix). Finally, as both stages of the proof are constructive, we can immediately use them to design an approach to estimate $q(Y|X)$. This is shown in Algorithm 1.

3. Uniqueness

Do we really need C, W? And why can't we have additional edges such as $X \rightarrow Y$ in Figure 1(c)? We describe here at a high level why generalizing the graph by removing observed nodes or adding edges prevents non-parametric identification of a simpler quantity $p(Y|do(C))$ (all results will apply to Figure 1(d) by swapping X and Y). While these are not necessary conditions, they are nearly as general as those used in non-parametric identification results in causal inference literature (Miao et al. (2018); Lee & Bareinboim (2021) allow for an additional edge $W \rightarrow Y$).

Can C and/or W be removed? Removing C corresponds to the setting of Pearl (2010), where the goal is to estimate $p(Y|do(X))$. This work assumes one can: (a) observe U without error in a subpopulation (Selén, 1986; Greenland & Lash, 2008), (b) observe $p(W|U)$ (Pearl, 2010), or (c) place a prior distribution on the parameters of $p(W|U)$ to bound $p(Y|do(X))$ (Greenland, 2005). However, these techniques are hard to apply when U is complex, e.g., if U is a collection of social determinants of health (Marmot & Wilkinson, 2005). Here we will not assume it is possible to observe $U, p(W|U)$ or derive a prior for $p(W|U)$. Removing W leads to a generalization of the front-door graph (Pearl et al., 2000) for which causal effects are not non-parametrically identifiable. Removing both C and W, one can only identify $p(Y|do(X))$ if U is observed, an assumption called 'ignorability' (Imbens & Rubin, 2015).

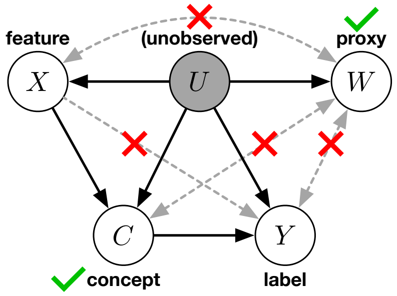

Figure 2. Removing C, W or adding any of the dotted edges prevents non-parametric identification of $p(W|\tilde{U})$ via our approach.

Can we add any additional edges? Identifying $p(W|U)$ (i.e., Stage 1 in the proof of Theorem 1) requires that both $W \perp\!\!\!\perp \{X, C, Y\} \mid U$ and $Y \perp\!\!\!\perp \{W, X\} \mid \{U, C\}$. The first conditional independence means that we cannot have any arrows from X, C, Y to or from W. We do not prove here that this is necessary, but we suspect that it is: currently the only edge that can be added for identifying $p(Y|do(C))$ is $W \rightarrow Y$ Miao et al. (2018); Lee & Bareinboim (2021), and these methods do not identify $p(W|U)$. The only other edge that could be added to the graph and it still be a DAG is $X \rightarrow Y$. However, this would violate the second conditional independence statement as it would make $Y \not\perp\!\!\!\perp X \mid \{U, C\}$. This edge would also render the causal effect unidentifiable under the most generic non-parametric methods (Lee & Bareinboim, 2021).

4. Experimental Verification

We now describe a simple simulation to verify the identification results in the previous section. We simulate data following Figure 1(c)[2] where every variable is categorical and all child variables are log-linear functions of their parents (more details in the Appendix). We sample $n \in \{5e2, 1e3, 2e3, 5e3, 1e4, 2e4, 5e4, 1e5, 2e5, 5e5, 1e6\}$ inputs from each of P and Q, and compare the root mean squared error (RMSE) between a sample estimate of $q(Y|X)$ and (i) an estimate of $p(Y|X)$, (ii) our adaptation approach, i.e., eq. (1), **(iii)** an independent sample estimate of $q(Y|X)$. Here **(iii)** is the error due to sampling, a lower bound for estimation. We average all results over 5 random trials.

The results are shown in Figure 3. Even for small n, the error of $p(Y|X)$ is, at best, nearly 3× higher than approach. The error increase for our method is due to not having enough data to estimate distributions involving $\tilde{U}$ well. As n increases and we estimate the quantities in eq. (1) more accurately, our approach matches the error due to sampling, the estimation lower bound.

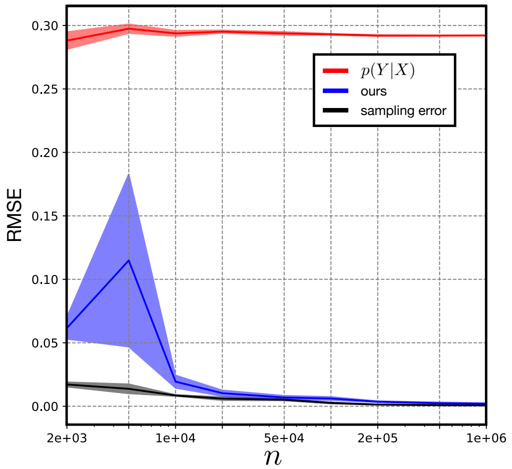

Figure 3. Simulation results for data from the graph of Figure 1(c).

5. Discussion

We presented a result for unsupervised domain adaptation when covariate or label shift assumptions are too restrictive. To do so, we observed that classic results in non-parametric identification of causal effects could be generalized to identify $q(Y|X)$. Because our analysis is constructive, it immediately yields an approach for estimating $q(Y|X)$.

We believe there are multiple interesting directions to further investigate. The biggest shortcoming of this approach is its limitation to discrete data. We believe the identification result extends to the setting where W, X, C, Y are allowed to be continuous. This is because: (i) The only marginalization one needs to perform to set up the matrix equations for identification is over U. So if U is kept discrete, this marginalization is kept finite. (ii) The only conditions on matrices is that they are full rank. This means that all of the proof steps should go through if we replace mass functions with densities. Extending this result further to the case where U is continuous seems possible: it likely requires conditions on U similar to those of Miao et al. (2018).

Even if continuous identification is possible, the estimation approach described in Algorithm 1 is non-trivial to extend to continuous data. Specifically, it requires density estimation of $p(X), q(X)$. This is hard as X is a high-dimensional object in many cases (e.g., EHR, image, text data). A more statistically efficient approach would be to learn a latent variable model, a technique used by Louizos et al. (2017). The difficulty here is to guarantee identification: i.e., Louizos et al. (2017) have no guarantees.

It is also worth deriving estimation guarantees (i.e., consistency guarantees, error bounds) for estimators of $q(Y|X)$. This would help understand if further data in Q could improve estimation. For example, if we also observed C in Q would this more tightly bound the error of $q(Y|X)$?

[2] Obtaining results for Figure 1(d) is trivial as the only change involves swapping X and Y when estimating $p(W|\tilde{U})$.

Acknowledgements We would like to thank Victor Veitch and Alex Brown for valuable discussions and feedback. This work was performed while MJK was at Google.

References

Alexandari, A., Kundaje, A., and Shrikumar, A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pp. 222–232. PMLR, 2020.

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

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pp. 1270–1279. PMLR, 2016.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.*, 385(3):283–286, July 2021.

Forman, G. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

Gart, J. and Buck, A. Comparison of a screening test and a reference test in epidemiologic studies. ii. a probabilistic model for the comparison of diagnostic tests. *American journal of epidemiology*, 83(3):593–602, 1966.

Greenland, S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.

Greenland, S. and Lash, T. Bias analysis in modern epidemiology (pp. 345–380). *Philadelphia, PN: Lippincott Williams & Wilkins.[Google Scholar]*, 2008.

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.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pp. 365–372. IEEE, 2009.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 951–958. IEEE, 2009.

Lee, S. and Bareinboim, E. Causal identification with matrix equations. *Advances in Neural Information Processing Systems*, 34, 2021.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Manski, C. F. and Lerman, S. R. The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, pp. 1977–1988, 1977.

Marmot, M. and Wilkinson, R. *Social determinants of health*. Oup Oxford, 2005.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Pearl, J. On measurement bias in causal inference. In *UAI*, 2010.

Pearl, J. et al. Causality: Models, reasoning and inference. *Cambridge University Press*, 19:2, 2000.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33: 11539–11551, 2020.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *International Conference on Machine Learning*, pp. 459–466, 2012.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

Selén, J. Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, 81(393): 75–81, 1986.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Tachet des Combes, R., Zhao, H., Wang, Y.-X., and Gordon, G. J. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 114, 2004.

Appendix

Recall Theorem 1:

Theorem 1. *Given that the above assumptions hold, all distributions over $W, X, C, Y, \tilde{U}$ are identifiable, where $\tilde{U}$ is an unknown sorting of U.*

Before we prove this we will prove a variant of Theorem 1 of Kuroki & Pearl (2014).

Theorem 2 (variant of Theorem 1 of Kuroki & Pearl (2014))**.** *Given assumptions A1-A4, $p(W|\tilde{U})$ is identifiable.*

Proof. First consider that we can factorize the joint of W, X, Y conditional on C as:

$$p(Y, X, W \mid C) = \sum_{k=1}^{k_U} p(Y \mid C, U = k) p(X \mid C, U = k) p(W \mid U = k) p(U = k \mid C).$$

Next, construct the following matrices based on the decomposition of $p(Y, X, W|C)$ and of its marginal distributions:

$$\mathbf{A} := \begin{bmatrix} 1 & p(W = 1|C) & \cdots & p(W = k_W - 1|C) \\ p(X = 1|C) & p(X = 1, W = 1|C) & \cdots & p(X = 1, W = k_W - 1|C) \\ \vdots & \vdots & \ddots & \vdots \\ p(X = k_X - 1|C) & p(X = k_X - 1, W = 1|C), & \cdots & p(X = k_X - 1, W = k_W - 1|C) \end{bmatrix}$$

$$\mathbf{B} := \begin{bmatrix} p(Y|C) & p(Y, W = 1|C) & \cdots & p(Y, W = k_W - 1|C) \\ p(Y, X = 1|C) & p(Y, X = 1, W = 1|C) & \cdots & p(Y, X = 1, W = k_W - 1|C) \\ \vdots & \vdots & \ddots & \vdots \\ p(Y, X = k_X - 1|C) & p(Y, X = k_X - 1, W = 1|C) & \cdots & p(Y, X = k_X - 1, W = k_W - 1|C) \end{bmatrix}$$

$$\mathbf{R} := \begin{bmatrix} 1 & p(X = 1|C, U = 1) & \cdots & p(X = k_X - 1|C, U = 1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & p(X = 1|C, U = k_U) & \cdots & p(X = k_X - 1|C, U = k_U) \end{bmatrix}$$

$$\mathbf{M} := \begin{bmatrix} p(U = 1|C) & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & \cdots & 0 & p(U = k_U|C) \end{bmatrix} \Delta = \begin{bmatrix} p(Y|C, U = 1) & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & \cdots & 0 & p(Y|C, U = k_U) \end{bmatrix}$$

$$\mathbf{S} := \begin{bmatrix} 1 & p(W = 1|U = 1) & \cdots & p(W = k_W - 1|U = 1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & p(W = 1|U = k_U) & \cdots & p(W = k_W - 1|U = k_U) \end{bmatrix}$$

Then note that

$$\mathbf{A} = \mathbf{R}^\top \mathbf{M} \mathbf{S} \qquad \mathbf{B} = \mathbf{R}^\top \mathbf{M} \Delta \mathbf{S}. \tag{2}$$

We then have that,

$$\begin{aligned} \mathbf{A}^\dagger \mathbf{B} &= \left[\left(\mathbf{A}^\top \mathbf{A} \right)^{-1} \mathbf{A}^\top \right] \mathbf{R}^\top \mathbf{M} \Delta \mathbf{S} \\ &= \left(\mathbf{S}^\top \mathbf{M} \mathbf{R} \mathbf{R}^\top \mathbf{M} \mathbf{S} \right)^{-1} \mathbf{S}^\top \mathbf{M} \mathbf{R} \left(\mathbf{R}^\top \mathbf{M} \Delta \mathbf{S} \right) \\ &\underset{(a)}{=} (\mathbf{S})^{-1} \Delta \mathbf{S}, \end{aligned}$$

where $\mathbf{A}^\dagger$ is the Moore-Penrose psuedoinverse of $\mathbf{A}$ (recall all psuedoinverses are unique and exist). In (a) we assume $\mathbf{S}$ is square and full rank. Further, $\mathbf{R}$ has rank k_U so that $\mathbf{R}\mathbf{R}^\top$ is invertible. In the event that $k_W > k_U$, the invertibility of $\mathbf{S}$ will require coarsening W to ensure equality of the coarsened k'_W and k_U.

Because we have to marginalize U in order to obtain observed distributions, it is only possible to identify U up to an arbitrary permutation. Specifically, let $\tilde{U}$ be a sorting of U such that $p(Y|C, \tilde{U}=1) > p(Y|C, \tilde{U}=2) > \cdots p(Y|C, \tilde{U}=k_U)$.

We may then identify $p(W|\tilde{U})$ from the eigendecomposition of $\mathbf{A}^\dagger \mathbf{B}$, directly employing the argument of Kuroki & Pearl (2014, Appendix) to recover $p(W|\tilde{U})$ from the eigenvectors of $\mathbf{A}^\dagger \mathbf{B}$. □

Now that we have obtained $p(W|\tilde{U})$, we can prove Theorem 1.

Proof. As distributions that only involve $\{W, X, C, Y,\}$ are observable, all we need to prove is that we can identify all distributions involving $\tilde{U}$. Let $\mathcal{V} \subseteq \{W, X, C, Y\}$ and $\mathcal{V}' \subseteq \{W, X, C, Y\} \setminus \mathcal{V}$. All we need to identify are (a) $p(\tilde{U})$, (b) $p(\mathcal{V}|\tilde{U})$, (c) $p(\tilde{U}|\mathcal{V})$, (d) $p(\mathcal{V}|\tilde{U}, \mathcal{V}')$. Note that this is sufficient because (e) $p(\tilde{U}, \mathcal{V}|\mathcal{V}') = p(\mathcal{V}|\tilde{U}, \mathcal{V}')p(\tilde{U}|\mathcal{V}')$ (given by (d) and (c)).

For (a) note that $p(\tilde{\mathbf{U}}) = p(\mathbf{W}|\tilde{\mathbf{U}})^\dagger p(\mathbf{W})$.

For (b) recall we have already identified $p(\mathbf{W}|\tilde{\mathbf{U}})$. So let $\mathcal{V}_{\backslash W} = \mathcal{V} \setminus W$. Then we have that $p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\mathbf{W}) = p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\tilde{\mathbf{U}})p(\tilde{\mathbf{U}}|\mathbf{W}) \Rightarrow p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\tilde{\mathbf{U}}) = p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\mathbf{W})p(\tilde{\mathbf{U}}|\mathbf{W})^\dagger$ (as $\mathcal{V}_{\backslash W} \perp\!\!\!\perp W \mid \tilde{U}$). Note that this is identified because the first term on the right-hand side is observed and the second term can be identified via Bayes rule $p(\tilde{U}|W) = p(W|\tilde{U})p(\tilde{U})/p(W)$. Finally note that $p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}, \mathbf{W}|\tilde{\mathbf{U}}) = p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\tilde{\mathbf{U}})$ (again due to $\mathcal{V}_{\backslash W} \perp\!\!\!\perp W \mid \tilde{U}$), which we have just identified.

For (c) we identified $p(\tilde{\mathbf{U}}|\mathbf{W})$ above by Bayes rule. We then have that $p(\mathbf{W}|\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}) = p(\mathbf{W}|\tilde{\mathbf{U}})p(\tilde{\mathbf{U}}|\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}) \Rightarrow p(\tilde{\mathbf{U}}|\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}) = p(\mathbf{W}|\tilde{\mathbf{U}})^\dagger p(\mathbf{W}|\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}})$, which is identifiable. Finally we have via Bayes rule $p(\tilde{U}|\mathcal{V}_{\backslash W}, W) = p(\mathcal{V}_{\backslash W}, W|\tilde{U})p(\tilde{U})/p(\mathcal{V}_{\backslash W}, W)$ all of which we can identify (via (a) and (b)).

For (d) $p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\mathcal{V}'_{\backslash W}, \mathbf{W}) = p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\tilde{\mathbf{U}}, \mathcal{V}'_{\backslash W})p(\tilde{\mathbf{U}}|\mathcal{V}'_{\backslash W}, \mathbf{W}) \Rightarrow p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\tilde{\mathbf{U}}, \mathcal{V}'_{\backslash W}) = p(\boldsymbol{\mathcal{V}}_{\backslash \boldsymbol{W}}|\mathcal{V}'_{\backslash W}, \mathbf{W})p(\tilde{\mathbf{U}}|\mathcal{V}'_{\backslash W}, \mathbf{W})^\dagger$. The first term on the right-hand side is observed and the second is identified via (c). Finally note that $p(\mathcal{V}_{\backslash W}, W|\tilde{U}, \mathcal{V}'_{\backslash W}) = p(W|\mathcal{V}_{\backslash W}, \tilde{U}, \mathcal{V}'_{\backslash W})p(\mathcal{V}_{\backslash W}|\tilde{U}, \mathcal{V}'_{\backslash W}) = p(W|\tilde{U})p(\mathcal{V}_{\backslash W}|\tilde{U}, \mathcal{V}'_{\backslash W})$ (as $W \perp\!\!\!\perp \mathcal{V}_{\backslash W}|\tilde{U}$), where all right-hand terms are identified. Also that $p(\mathcal{V}_{\backslash W}, |W, \tilde{U}, \mathcal{V}'_{\backslash W}) = p(\mathcal{V}_{\backslash W}, |\tilde{U}, \mathcal{V}'_{\backslash W})$ which is identified. □

Simulation Details

We let $k_U = 3, k_X = 4, k_C = 3, k_Y = 2, k_W = 3$. Let $\sigma(\mathbf{z}) : \mathbb{R}^d \to \Delta^{d-1}$ be the softmax function applied to a vector $\mathbf{z}$, where Δ^{d-1} is the $(d-1)$-dimensional simplex. Let $\mathbf{o}(v)$ be the $|V|$-dimensional one-hot representation of a sample from a categorical variable $v \in V$. The following equations describe how each variable is sampled in the simulation results of Section 4.

$$
\begin{aligned}
p(U) &= \sigma([1, 0.1, 0.1]) \\
q(U) &= \sigma([0.1, 0.1, 1]) \\
p(W \mid U = u) &= \sigma(\mathbf{M}_{W|U}\mathbf{o}(u)) \\
p(X \mid U = u) &= \sigma(\mathbf{M}_{X|U}\mathbf{o}(u)) \\
p(C \mid X = x, U = u) &= \sigma(\mathbf{M}_{C|X}\mathbf{o}(x) + \mathbf{M}_{C|U}\mathbf{o}(u)) \\
p(Y \mid C = c, U = u) &= \sigma(\mathbf{M}_{Y|C}\mathbf{o}(c) + \mathbf{M}_{Y|U}\mathbf{o}(u))
\end{aligned}
$$

where the matrices are defined as:

$$
\mathbf{M}_{W|U} := \begin{bmatrix} 5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 5 \end{bmatrix} \mathbf{M}_{X|U} := \begin{bmatrix} 5 & 5 & 0.5 \\ 5 & 0.5 & 5 \\ 0.5 & 5 & 5 \\ 5 & 0.5 & 5 \end{bmatrix} \mathbf{M}_{C|X} := \begin{bmatrix} 5 & 5 & 0.5 & 0.5 \\ 5 & 0.5 & 5 & 5 \\ 0.5 & 5 & 5 & 0.5 \end{bmatrix} \mathbf{M}_{C|U} := \begin{bmatrix} 5 & 5 & 0.5 \\ 5 & 0.5 & 5 \\ 0.5 & 5 & 5 \end{bmatrix}
$$

$$
\mathbf{M}_{Y|C} = \mathbf{M}_{Y|U} := \begin{bmatrix} 5 & 5 & 0.5 \\ 0.5 & 5 & 5 \end{bmatrix}.
$$

These were chosen to ensure that the shift from $p(U)$ to $q(U)$ caused changes in the distributions of observed variables.