Learning Invariant Representations under General Interventions on the Response

Anonymous Author(s)
Affiliation
Address
email

Abstract

It has become increasingly common nowadays to collect observations of feature and response pairs from different environments. As a consequence, one has to apply learned predictors to data with a different distribution due to distribution shifts. One principled approach is to adopt the structural causal models to describe training and test models, following the invariance principle which says that the conditional distribution of the response given its predictors remains the same across environments. However, this principle might be violated in practical settings when the response is intervened. A natural question is whether it is still possible to identify other forms of invariance to facilitate prediction in unseen environments. To shed light on this challenging scenario, we introduce invariant matching property (IMP) which is an explicit relation to capture interventions through an additional feature. This leads to an alternative form of invariance that enables a unified treatment of general interventions on the response. We analyze the asymptotic generalization errors of our method under both the discrete and continuous environment settings, where the continuous case is handled by relating it to the semiparametric varying coefficient models. We present algorithms that show competitive performance compared to existing methods over various experimental settings.

1 Introduction

How to make reliable prediction in unseen environments that are different from training environments is a challenging problem, which is fundamentally different from the classical machine learning settings [38, 48, 11]. Modeling these distribution shifts in a principled way is of great importance in many fields including robotics, medical imaging, and environmental science. Apparently, this problem is ill-posed without any constraints on the relationship between training and test distributions, as the test distribution may be arbitrary. Consider the problem of predicting the response Y given its predictors $X = (X_1, ..., X_d)^\top$ in unseen environments. To model distribution changes across different environments (or training and test distributions), we follow the approach of using *structural causal models* (SCMs) [32, 35] to model different data-generating mechanisms. The common assumption is that the assignment for Y does not change across environments (or Y is not intervened), which allow for natural formulations of the invariant conditional distribution of Y given a subset of X [43, 49, 34, 35, 40, 20, 8]. The underlying principle is known as invariance, autonomy or modularity [19, 1, 32, 24]. We provide a brief review of the related works that follow this principle in Appendix A.

In practical settings, however, the structural assignment of Y might change across environments, namely, Y might be intervened. How to relax this assumption in a principled way is one of the main motivations in our work. We propose to explore alternative forms of invariance, and make

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

an attempt in this direction by focusing on linear SCMs. Concretely, the assignment for Y allows general interventions

$$Y^e = (\beta^e)^\top X^e + \varepsilon_Y^e,$$

where Y can be intervened through coefficient β^e and/or the noise ε_Y^e, to capture the dependence of structural assignment across different environments. We consider a multi-environment regression setting for domain adaption: There are multiple training data (X^e, Y^e) for $e \in \mathcal{E}^{\text{train}}$ that are generated from a training model and one test data (indexed by e^{test}) from a test model; we assume the training model and test model follow SCMs with the same graph structure, but we allow β^e and the mean and variance of ε_Y^e to be arbitrarily different under the two models. To avoid the setting to be ill-posed, a key necessary condition is that Y^e needs to have at least one child in the SCMs, as prediction is not possible otherwise given that Y^e may change arbitrarily over environments. The main challenge lies in whether it is still possible to identify other forms of invariance to facilitate prediction in the test environment. We propose an alternative form of invariance $\mathcal{P}_e(Y|\phi_e(X)) = \mathcal{P}_h(Y|\phi_h(X))$ that is enabled by a family of conditional invariant transforms $\Phi \ni \phi_e, \phi_h$. Under general interventions on Y, we provide explicit constructions of such transforms by developing *invariant matching property (IMP)*, a deterministic relation between an estimator of Y and X along with an additional predictor constructed from X.

2 Background and Problem Formulation

Consider a linear acyclic SCM $\mathcal{M}$ over (X, Y) (see Appendix B.1), the coefficients and noise distributions may change when (X, Y) is observed in different environments (e.g., different experiment settings for data collection). In the following, we use interventions on the SCM $\mathcal{M}$ to model such changes. Let $\mathcal{E}^{\text{all}}$ denote the set of all possible environments [1], which consists of *multiple* training environments $\mathcal{E}^{\text{train}}$ and one test environment $\{e^{\text{test}}\}$ such that $\mathcal{E}^{\text{all}} = \mathcal{E}^{\text{train}} \cup \{e^{\text{test}}\}$. For each $e \in \mathcal{E}^{\text{all}}$, an acylcic linear SCM over (X^e, Y^e) is given by

$$\mathcal{M}^e : \begin{cases} X^e = \gamma^e Y^e + B^e X^e + \varepsilon_X^e & (1) \\ Y^e = (\beta^e)^\top X^e + \varepsilon_Y^e. & (2) \end{cases}$$

A variable from $\{X_1, \ldots, X_d, Y\}$ is intervened if the parameters or noise distribution in its assignment changes over different $e \in \mathcal{E}^{\text{all}}$. This formulation of $\mathcal{M}^e$ is fairly general, and we discuss several special cases in Appendix C.

We observe n^e i.i.d. samples $\{(x_1, y_1), ..., (x_{n^e}, y_{n^e})\}$ from each training environment distribution $\mathcal{P}_e$ for $e \in \mathcal{E}^{\text{train}}$, but in the test environment e^{test} we only observe m i.i.d. samples $\{x_1, ..., x_m\}$ from $\mathcal{P}_{\text{test}}$. The goal is to learn a function $f : \mathcal{X} \to \hat{\mathcal{Y}}$ that works well on e^{test} in the sense that it minimizes the *test population loss*

$$\mathcal{L}_{\text{test}}(f) := \mathsf{E}_{(X,Y)\sim\mathcal{P}_{\text{test}}}[l(Y, f(X))]. \tag{3}$$

where l is the square loss function $l(y, \hat{y}) = (y - \hat{y})^2$. To make this problem tractable, we assume that (X, Y) under $\mathcal{P}_{\text{test}}$ and $\mathcal{P}_e$ are generated according to the SCM $\mathcal{M}^e$ but we *allow for general types of interventions*.

It is well-known that if Y is not intervened, a general form of invariance principle applies, assuming the existence of some subset $S \subseteq \{1, ..., d\}$ such that $\mathcal{P}_e(Y|X_S) = \mathcal{P}_h(Y|X_S)$ holds for any $e, h \in \mathcal{E}^{\text{all}}$. The main challenge in our setting comes from the general interventions on Y, making the traditional invariance principle not applicable. In this work, we propose to exploit an alternative form of invariance to tackle this problem.

Definition 1. *A function $\phi \in (\mathcal{E}^{all}, \mathbb{R}^d) \to \mathbb{R}^q$ is called a* conditional invariant transform *if the following* invariance property *holds for any $e, h \in \mathcal{E}^{all}$*

$$\mathcal{P}_e(Y|\phi_e(X)) = \mathcal{P}_h(Y|\phi_h(X)). \tag{4}$$

Under general intervention settings, we denote this class of conditional invariant transforms as Φ, and we provide *explicit* characterizations of it via the invariant matching property (IMP) (see Definition 2). For each $\phi \in \Phi$, the invariance property (4) enables us to compute

$$f_{\phi_e}(x) = g \circ \phi_e(x) = \mathsf{E}_{\mathcal{P}_e}[Y|\phi_e(x)], \tag{5}$$

[1] We use training (or test) environments and observable (or unseen) environments interchangeably.

for any $e \in \mathcal{E}^{\text{all}}$, where the function $g : \mathbb{R}^q \to \hat{\mathcal{Y}}$ is invariant across environments and is nonlinear in general. Equivalently, this solves a relaxed version of (3) by minimizing $\mathcal{L}_{\text{test}}(f_\phi)$ over $\{\phi \in \Phi\}$.

3 A Motivating Example

Example 1. *Consider* $(Y^e, X^e), e \in \mathcal{E}^{toy} = \{1, 2\}$, *with* $X^e := (X_1^e, X_2^e, X_3^e)^\top$ *satisfying the following linear acyclic SCM (illustrated in Figure. 1),*

$$\mathcal{M}_{toy}^e : \begin{cases} Y^e = a^e X_1^e + X_2^e + N_Y^e \\ X_3^e = Y^e + X_1^e + N_3^e, \end{cases}$$

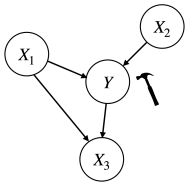

Figure 1: Directed acyclic graph $\mathcal{G}(\mathcal{M}_{\text{toy}}^e)$.

where X_1^e, X_2^e, N_3^e, *and* N_Y^e *are independent and* $\mathcal{N}(0, 1)$*-distributed for every* $e \in \mathcal{E}^{toy}$. *Since* (Y^e, X^e) *is multivariate Gaussian, the MMSE estimator of* Y^e *given* X^e *is*

$$\begin{aligned} \mathsf{E}_{\mathcal{P}_e}[Y|X] =& X^\top \left(\mathsf{E}_{\mathcal{P}_e}[XX^\top]\right)^{-1} \mathsf{E}_{\mathcal{P}_e}[XY] \\ =& \frac{1}{2}(a^e - 1)X_1^e + \frac{1}{2}X_2^e + \frac{1}{2}X_3^e, \end{aligned}$$

Similarly, one can compute $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2] = (1 + a^e)X_1^e + X_2^e$. *As a result, there exists a deterministic linear relation, which we refer to as* matching,

$$\mathsf{E}_{\mathcal{P}_e}[Y|X] = \lambda \mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2] + \eta^\top X^e, \tag{6}$$

with coefficients $\lambda = 1/2$ *and* $\eta = (-1, 0, 1/2)^\top$ *that are* invariant *with respect to the environment. Moreover, one can verify that* $\mathcal{P}_e(Y|X_1, X_3, \mathsf{E}[X_3|X_1, X_2])$ *is invariant since the corresponding conditional mean and variance are invariant. A prediction model in* (6) *with invariant coefficients is often not unique when it exists. One can show that* $\mathsf{E}_{\mathcal{P}_e}[Y|X] = -X_1^e + \frac{1}{2}X_2^e + X_3^e - \frac{3}{2}\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, X_3]$. *However, such invariant relations do not hold for* $\mathsf{E}_{\mathcal{P}_e}[X_1|X_2, X_3]$. *To further illustrate the invariant relations, we provide simulations in Appendix D. Moreover, in Appendix E, we extend Example 1 to allow for interventions on* X_1, X_2, *and* Y *through the means and/or variances of the noise variables.*

4 Invariant Matching Property

In this section, we generalize the invariant relations observed in Example 1 to a class of such relations for $\mathcal{M}^e$, $e \in \mathcal{E}^{\text{all}}$. To handle non-Gaussian cases (beyonnd Example 1), we choose to adopt the *linear MMSE* estimators (see Appendix B.2) for constructing *linear* invariant relations. To simplify presentation, we focus on (X^e, Y^e) with zero means for each $e \in \mathcal{E}^{\text{all}}$, while the non-zero mean settings can be handled by introducing the constant one as an additional predictor.

Definition 2. *For* $k \in \{1, \ldots, d\}$, $R \subseteq \{1, \ldots, d\} \setminus k$, *and* $S \subseteq \{1, \ldots, d\}$, *we say that the tuple* (k, R, S) *satisfies the* invariant matching property *(IMP) if, for every* $e \in \mathcal{E}^{all}$,

$$\mathsf{E}_{l,\mathcal{P}_e}[Y \mid X_S] = \lambda \mathsf{E}_{l,\mathcal{P}_e}[X_k \mid X_R] + \eta^\top X^e, \tag{7}$$

for some $\lambda \in \mathbb{R}$ *and* $\eta \in \mathbb{R}^d$ *that do not depend on* e. *We denote* $\mathcal{I}_{\mathcal{M}} := \{(k, R, S) : \text{ (7) holds}\}$ *for model* $\mathcal{M}$, *and we call* $(\eta^\top, \lambda)^\top$ *the* matching parameters.

Observe that $\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S]$ is not directly applicable to the test environment due to its components depending on e, but those components are fully captured by $\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R]$. We formally define this class of additional features as follows.

Definition 3. *For any* $k \in \{1, \ldots, d\}$ *and* $R \subseteq \{1, \ldots, d\} \setminus k$, *we call* $\mathsf{E}_{l,\mathcal{P}_e}[X_k \mid X_R]$ *a* prediction module. *If a prediction module satisfies an IMP for some* $S \subseteq \{1, \ldots, d\}$, *we call it a* matched prediction module *for* S.

Now we discuss the relationship between the IMP and the invariance property $\mathcal{P}_e(Y|\phi_e(X)) = \mathcal{P}_h(Y|\phi_h(X))$ in (4). Define $\phi_e^{(k,R,S)}(X^e) := (X^e_{S'}, \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R])^\top$, where $X^e_{S'}$ is a row vector for some $S' \subseteq \{1,\dots,d\}$. In general, the invariance of the matching parameters $\{\lambda, \eta\}$ does not imply that the invariance property (4) some $\phi_e^{(k,R,S)}(X^e)$. In Section 5, we will characterize a class of IMPs that each satisfies (4). *It is crucial to note that we only use the IMP to identify the transform to satisfy the invariance property.* When the invariance property is in place, one can apply the general conditional expectation $f_{\phi_e}(x) = \mathsf{E}_{\mathcal{P}_e}[Y|\phi_e(x)]$ as in (5), since the linear estimator from the IMP is in general sub-optimal for the non-Gaussian cases.

It is noteworthy that since $\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R]$ is a linear function of X^e_R, the matching parameters are not unique given a single environment $e \in \mathcal{E}^{\text{train}}$. We show that *two* training environments are sufficient for the identification of the matching parameters in Appendix F.1. Furthermore, we provide a natural decomposition of the IMP in Appendix (F.2) to illustrate the connections among the estimator $\mathsf{E}_{l,\mathcal{P}_e}[Y \mid X_S]$, prediction module $\mathsf{E}_{l,\mathcal{P}_e}[X_k \mid X_R]$, and causal function $\mathsf{E}_{\mathcal{P}_e}[Y \mid X_{PA(Y)}]$.

5 Characterization of Invariant Matching Properties

First, we consider model $\mathcal{M}^e$ with interventions only on Y through the coefficients β^e [2]. To distinguish the parents of Y with varying and invariant coefficients, we decompose β^e in $\mathcal{M}^e$ into two parts α^e and β. Without loss of generality, we assume that $\alpha^e_j \neq 0$ if and only if α^e_j is a non-constant function of e, and we define $PE = \{j \in \{1,\dots,d\} : \alpha^e_j \neq 0\}$. Recall that prediction modules do not rely on the response Y but the relations between the predictors for each environment. When Y is unobserved (or equivalently, substituting Y in (2) into (1)), the relations between the predictors are as follows,

$$X^e = \left(\gamma\left(\alpha^e + \beta\right)^\top + B\right) X^e + \gamma\varepsilon_Y + \varepsilon_X, \tag{8}$$

where $\gamma\varepsilon_Y + \varepsilon_X$ a vector of dependent random variables when γ is not a zero vector. If α^e vanishes from (8), the distribution of X^e becomes invariant with respect to environments, but the distribution of Y^e changes arbitrarily due to the change of α^e, which makes the prediction problem ill-posed. Observe that α^e is non-vanishing in (8) only if γ is not a zero vector, which brings up the following key assumption.

Assumption 1. *When Y is intervened, we assume that Y has at least one child.*

The first and second matching properties provided in Appendix F.2 enable us to characterize the tuples (k, R, S)'s that satisfy IMPs through the characterizations of S (for the first matching property) and (k, R) (for the second matching property) separately. In the following theorem, we show that a class of IMPs implied by the first and second invariant matching properties satisfy the invariance property (4).

Theorem 1. *For model $\mathcal{M}^{e,1}$, the first and second matching properties hold in the following cases.*

1. On the first MP: *For each $S \subseteq \{1,\dots,d\}$ such that $PE \subseteq S$, the first matching property holds.*
2. On the second MP: *For each $k \in \{1,\dots,d\} \setminus PE$ and $R \subseteq \{1,\dots,d\} \setminus k$ such that $PE \subseteq R$, the second matching property holds.*

For any tuple (k, R, S) above such that $R \subseteq S$, if $\lambda_X \neq 0$ in the second matching property, then $\phi_e(X^e) = (X^e_S, \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R])^\top$ satisfies (4). *Furthermore, $\mathcal{L}_{test}(f_\phi)$ is minimized by any ϕ with $S = \{1, ..., d\}$.*

It is noteworthy that Assumption 1 is a necessary condition for $\lambda_X \neq 0$, and we provide a sufficient condition for $\lambda_X \neq 0$ in a concrete setting with $S = \{1,\dots,d\}$ in Proposition 2 in Appendix G. More details regarding the characterization of IMPs are provided in Appendix G, along with the characterization of IMP with interventions on both the predictors and response.

Due to space limit, we provide our algorithms developed for both discrete and continuous environment setting in Appendix I. To handle the continuous environment setting, we bridge our framework with

[2] Note that, in the non-zero mean settings, this model covers the shift intervention on Y through the varying coefficient of the predictor that is a constant one.

the profile likelihood estimators developed in the semiparametric literature. A note on the profile likelihood estimation for semiparameteric varying coefficient models can be found in Appendix K. Our algorithms are examined over various synthetic data sets under different intervention settings in Appendix J. Finally, we analyze the asymptotic generalization errors of the proposed estimators for both discrete and continuous environment settings in the Appendix L.

References

[1] John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2018.

[4] J Andrew Bagnell. Robust supervised learning. In *AAAI*, pages 714–719, 2005.

[5] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[6] Maurice S Bartlett. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111, 1951.

[7] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19, 2006.

[8] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[9] Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22:1–80, 2021.

[10] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[12] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[13] Jianqing Fan and Tao Huang. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057, 2005.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[15] Rui Gao, Xi Chen, and Anton J Kleywegt. Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[16] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

[17] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016.

[18] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[19] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

[20] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

[21] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

[22] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[23] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724, 2013.

[24] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[25] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

[26] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David Carlson. On target shift in adversarial domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 616–625. PMLR, 2019.

[28] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

[29] Yue-pok Mack and Bernard W Silverman. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3): 405–415, 1982.

[30] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

[31] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[32] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[33] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 451–482. 2022.

[34] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[35] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[36] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[37] Niklas Pfister, Evan G Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

[38] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[39] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *Advances in Neural Information Processing Systems*, 31, 2018.

[40] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[41] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

[42] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[43] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[44] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[45] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

[46] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. 2012.

[47] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.

[48] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):1–40, 2016.

[49] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

Contents

A Related Works **10**

B Definitions **10**

B.1 Linear SCM 10

B.2 Linear MMSE estimator 11

C Types of Interventions on Linear SCMs **11**

D Simulations on Example 1 **11**

E Another Toy Example **12**

F Properties of the IMP **12**

F.1 Identification of the Matching Parameters 12

F.2 A Decomposition of the IMP 13

G More on the Characterization of IMPs **14**

G.1 Interventions on the Response 14

G.2 Interventions on both Predictors and Response 14

H Proofs for the Theoretical Results in Section 5 and Section G **15**

H.1 Proof of Theorem 1 15

H.2 Proof of Proposition 2 17

H.3 Proof of Proposition 2 18

H.4 Proof of Corollary 2 19

H.5 Proof of Theorem 2 20

I Algorithms **20**

I.1 Discrete Environments 20

I.2 Continuous Environments 22

J Experiments **23**

J.1 Discrete environments 23

J.2 Interventions on both X and Y (continuous) 24

J.3 Robustness 25

K Note on Semi-parametric Varying Coefficient Models and Profile Least-Squares Estimation **25**

L Asymptotic Generalization Error **27**

M Proofs for the Theoretical Results in Section L **27**

M.1 Technical Lemmas for the Proof of Theorem 3 27

M.2 Proof of Theorem 3 31
M.3 Proof of Corollary 3 32
M.4 Proof of Corollary 4 33

A Related Works

The invariance-based causal prediction initiated in [34] (also see [31] and [8] and references therein) assumes that the conditional distribution of Y given a set of predictors $X_S \subseteq \{X_1, ..., X_d\}$ is invariant in all environments, i.e., $\mathcal{P}_e(Y|X_S) = \mathcal{P}_h(Y|X_S)$ for environments e and h, where (X, Y) is generated according to the joint distribution $\mathcal{P}_e := \mathcal{P}_e^{X,Y}$. Focusing on linear SCMs, it assumes the existence of a linear model that is invariant across environments, with an unknown noise distribution and arbitrary dependence among predictors (see extensions to nonlinear [22] and time series [36] settings). Following this framework, theoretical guarantees for domain adaption have been developed in [40, 30]. More recently, a multi-environment regression method for domain adaption called the *stabilized regression* [37] explicitly enforces stability (based on a weaker version of invariance $\mathsf{E}_{\mathcal{P}_e}[Y|X_S = x_s] = \mathsf{E}_{\mathcal{P}_h}[Y|X_S = x_s]$) by introducing the *stable blanket*, which is a refined version of the *Markov blanket* to promote generalization. The tradeoff between predictive performance on training and test data has been studied via regularization under shift interventions [42]. Motivated by [34], the invariant risk minimization (IRM) [2] imposes $\mathcal{P}_e(Y|\phi(X)) = \mathcal{P}_h(Y|\phi(X))$, where ϕ is invariant across environments, leading to a bi-leveled optimization problem that is not practical. Several relaxed versions of IRM have been proposed in [2], but they behave very differently from the original IRM (see, e.g., [41, 25]). For a framework of the out-of-distribution setting from a causal perspective with a focus on minimizing the worst-case risk, see [10] and references therein. *In this line of invariance-based work, the fundamental assumption is that interventions on the target variable Y is not allowed*.

In [9], the authors have provided a systematic treatment of domain adaption using the SCMs to enable analysis and comparisons of domain adaption methods, which leads to the conditional invariant residual matching (CIRM) method. The CIRM and its variants combine the domain invariant projection (DIP)-type methods (see [5, 14] and the generalized label shift to handle target label perturbation [27, 47]) with the idea of conditional invariance penalty (appeared in [17, 21] under slightly different settings) that assumes the existence of conditional invariant components (CICs) in the anticausal setting where Y causes X. Theoretical guarantees have been provided for the prediction performance under shift interventions on Y, while numerical studies are provided for interventions on the noise variance of Y [9]. *It has also been pointed out that the general mixed-causal-anticausal domain adpation problem remain open*. We aim to shed light on this challenging setting by constructing explicit conditional invariant transforms.

The role of causality in facilitating domain adaptation problem is first articulated in [43], focusing on causal and anticausal predictions. Reweighting methods have been extensively studied for covariate shift [38, 45, 46], which assumes that only the feature distribution changes over environments while the conditionals remain the same. The label shift, which aligned with the anticausal setting, has attracted much attention recently [28, 3, 16]. Many other interesting domain adaption methods have been developed but they are less related to this work. The performance bounds using Vapnik-Chervonenkis (VC) theory has been initiated in [7]. There are fundamental work from the robust statistics perspective including distributional robust learning [4, 23, 44, 15, 26, 12] and adversarial machine learning [18, 39].

B Definitions

B.1 Linear SCM

Consider a response $Y \in \mathbb{R}$ and a vector of predictors $X = (X_1, \ldots, X_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d$, a linear SCM over (X, Y) is defined by

$$\mathcal{M} : \begin{cases} X = \gamma Y + BX + \varepsilon_X \\ Y = \beta^\top X + \varepsilon_Y, \end{cases}$$

where $\beta, \gamma \in \mathbb{R}^d$, $B \in \mathbb{R}^{d \times d}$, $\varepsilon_X = (\varepsilon_{X_1}, \ldots, \varepsilon_{X_d})^\top$, and the noise variables $\{\varepsilon_{X_1}, \ldots, \varepsilon_{X_d}\}$ and ε_Y are jointly independent. We use $\mathcal{G}(\mathcal{M})$ to denote the directed acyclic graph induced by $\mathcal{M}$, with edges determined by the non-zero coefficients in $\mathcal{M}$.

B.2 Linear MMSE estimator

For a target variable $Y \in \mathbb{R}$ given a vector of predictors $X \in \mathbb{R}^p$, the LMMSE estimator is define as

$$\mathsf{E}_l[Y|X] := (\theta^{\text{ols}})^\top (X - \mathsf{E}[X]) + \mathsf{E}[Y],$$

where $\theta^{\text{ols}} := \text{Cov}(X, X)^{-1} \text{Cov}(X, Y)$ is called the population ordinary least squares (OLS) estimator. With a slight abuse of notation, we write $\mathsf{E}_{l,\mathcal{P}_e}[Y \mid X]$ to denote the LMMSE of Y given X with respect to $(X, Y) \sim \mathcal{P}_e$.

C Types of Interventions on Linear SCMs

This formulation of the interventions on $\mathcal{M}$ as in $\mathcal{M}^e$ is fairly general. From the structural perspective, it consists of causal, anticausal, and mixed-causal-anticausal settings [43]. It should be noted that we only adopt the linear SCM rather than the fully specified SCMs as in [33], since learning the functional forms can be more complicated than the prediction problem we aim to solve. Regarding the interventions types, we discuss several special cases to put it into perspective.

1. *Shift interventions on X or Y:* A variable X_j is intervened through a shift if the mean of the noise variable $\varepsilon^e_{X,j}$ changes with $e \in \mathcal{E}^{\text{all}}$. For the shift intervention on Y, the mean of ε^e_Y changes .
2. *Interventions on the coefficients of X or Y:* A variable X_j is intervened through coefficients if the coefficients $\{\gamma^e_j, B^e_{j\cdot}\}$ change with $e \in \mathcal{E}^{\text{all}}$. For Y, the change is on the coefficient vector β^e.
3. *Interventions on the noise variance of X or Y:* Similar to shift interventions, a variable X_j or Y is intervened if its noise variance changes.

D Simulations on Example 1

Let $X_4 := \mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2]$, $X_5 := \mathsf{E}_{\mathcal{P}_e}[X_2|X_1, X_3]$, and $X_6 := \mathsf{E}_{\mathcal{P}_e}[X_1|X_2, X_3]$. We illustrate the linear invariant relations that corresponding to X_4 and X_5 as provided in Example 1. Recall that such invariant relations do not hold for X_6.

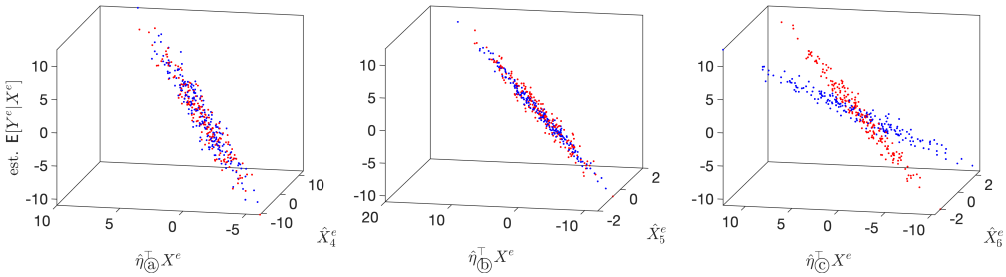

Figure 2: Illustrations of invariant relations in Example 1. In (a), we illustrate the linear invariant relation (6) by visualizing estimates of the tuple $(X^e_4, \eta^\top X^e, \mathsf{E}_{\mathcal{P}_e}[Y|X])$ (corresponding to the $\{x, y, z\}$ axes). Similarly, for (b) and (c), we verify such a relation for X^e_5 and X^e_6. The overlaps between the red dots ($e = 1$) and the blue dots ($e = 2$) in (a) and (b) indicate the invariant linear relations. However, the red and blue dots in (c) are not aligned, implying that no invariant relations as in (6) hold for $(\mathsf{E}_{\mathcal{P}_e}[Y|X], X^e_6)$.

Regarding the estimation of the MMSE estimators and the coefficients in the linear relations, we take the estimation procedure for Figure 2.(a) as an example in the following. For $e \in \{1, 2\}$, X^e_4 and $\mathsf{E}_{\mathcal{P}_e}[Y|X]$ are estimated using OLS. Then, η (denoted as $\eta_{ⓐ}$ in Figure 2.(a)) is estimated using OLS by regressing $\mathsf{E}_{\mathcal{P}_e}[Y|X]$ on (X^e_4, X^e) using the pooled data.

E Another Toy Example

We extend Example 1 to allow for interventions on X_1, X_2, and Y through the means and/or variances of the noise variables (see Fig. 3).

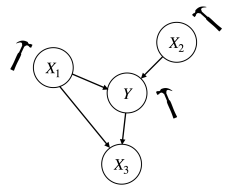

Figure 3: $\mathcal{G}(\mathcal{M}^e_{\text{toy}})$ with interventions on (X_1, X_2).

Example 2. *Consider training model $\mathcal{M}^e_{toy}$ in Example 1 with additional shift interventions and/or interventions on the noise variances. The results are summarized in the following.*

1. *Under shift interventions on Y and X_1 and an intervention on the variance of X_1, the two invariant relations in Example 1 hold with additional intercept terms.*
2. *When Y is intervened through the variance of N_Y, the relations in Example 1 will not hold. In this case, new relations can be established if $\mathsf{E}_{\mathcal{P}_e}[Y|X]$ in Example 1 is replaced by $\mathsf{E}_{\mathcal{P}_e}[Y|X_{PA(Y)}] = \mathsf{E}_{\mathcal{P}_e}[Y|X_1, X_2]$.*
3. *When X_2 is intervened through either the mean or variance, a relation as in Example 1 that is based on $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, X_3]$ will not hold.*
4. *Combining the interventions above, there will be one invariant relation left,*

$$\mathsf{E}_{\mathcal{P}_e}[Y|X_1, X_2] = \mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2] - X_1^e + b, \tag{9}$$

for some intercept $b \in \mathbb{R}$. It is noteworthy that (9) will fail to hold if X_3 is intervened. However, due to the intervention on the noise variance of Y, $\mathcal{P}_e(Y|X_S, \mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2])$ is not invariant for any $S \subseteq \{1, 2, 3\}$, since $\mathrm{Var}_{\mathcal{P}_e}(Y|X_S, \mathsf{E}_{\mathcal{P}_e}[X_3|X_1, X_2])$ changes with e.

F Properties of the IMP

F.1 Identification of the Matching Parameters

We rewrite (7) in a compact form as

$$\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S] = \theta^\top \widetilde{X}^e, \tag{10}$$

where

$$\widetilde{X}^e := (X_1^e, \ldots, X_d^e, \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R])^\top, \tag{11}$$

and $\theta = (\eta^\top, \lambda)^\top$ denotes the matching parameter.

Proposition 1. *For a tuple (k, R, S) that satisfies an IMP, the matching parameter θ can be uniquely identified in $\mathcal{E}^{train}$ if $|\mathcal{E}^{train}| \geq 2$ and*

$$\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R = x] \neq \mathsf{E}_{l,\mathcal{P}_h}[X_k|X_R = x] \tag{12}$$

for some $e, h \in \mathcal{E}^{train}$ and $x \in \mathcal{X}_R$.

Proof. Let $\mathcal{E}^{\text{train}} = \{e_1, \ldots, e_n\}$. For a tuple (k, S, R) that satisfies the IMP, let $\hat{\boldsymbol{Y}} = (\mathsf{E}[Y^{e_1}|X_S^{e_1}], \ldots, \mathsf{E}[Y^{e_n}|X_S^{e_n}])^\top$ and

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} (X^{e_1})^\top & \mathsf{E}_l[X_k^{e_1}|X_R^{e_1}] \\ \vdots & \vdots \\ (X^{e_n})^\top & \mathsf{E}_l[X_k^{e_n}|X_R^{e_n}] \end{bmatrix} := [\boldsymbol{X} \quad v],$$

where the rows of $\tilde{\boldsymbol{X}}$ are independent. According to (10), we have $\hat{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\theta$. Then, if $\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}]$ is invertible, we have

$$(\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}])^{-1}\mathsf{E}[\tilde{\boldsymbol{X}}^\top \hat{\boldsymbol{Y}}] = (\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}])^{-1}\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}]\theta = \theta. \tag{13}$$

Now, we prove the invertibility. Observe that

$$\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}] = \begin{bmatrix} \mathsf{E}[\boldsymbol{X}^\top \boldsymbol{X}] & \mathsf{E}[\boldsymbol{X}^\top v] \\ \mathsf{E}[v^\top \boldsymbol{X}] & \mathsf{E}[v^\top v] \end{bmatrix},$$

where $\mathsf{E}[\boldsymbol{X}^\top \boldsymbol{X}] = \sum_{i=1}^n \mathsf{E}[X^{e_i}(X^{e_i})^\top]$ is inveritble since it is a sum of positive-definite matrices. Then, $\mathsf{E}[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}]$ is invertible if and only if

$$\mathsf{E}[v^\top v] - \mathsf{E}[v^\top \boldsymbol{X}]\mathsf{E}^{-1}[\boldsymbol{X}^\top \boldsymbol{X}]\mathsf{E}[\boldsymbol{X}^\top v] \neq 0.$$

This is equivalent to

$$\mathsf{E}[(v - \boldsymbol{X}\beta)^\top (v - \boldsymbol{X}\beta)] \neq 0,$$

where $\beta := \mathsf{E}^{-1}[\boldsymbol{X}^\top \boldsymbol{X}]\mathsf{E}[\boldsymbol{X}^\top v]$. This is true since there is no $b \in \mathbb{R}^d$ such that $v = \boldsymbol{X}b$ almost surely by our assumption in (12). Therefore, θ is uniquely determined by (13).

□

F.2 A Decomposition of the IMP

In our toy examples, recall that the IMPs are derived by first computing $\mathsf{E}_{\mathcal{P}_e}[Y|X_S]$ and $\mathsf{E}_{\mathcal{P}_e}[X_k|X_R]$ separately and then fitting a linear relation from $(\mathsf{E}_{\mathcal{P}_e}[X_k|X_R], X_S)$ to $\mathsf{E}_{\mathcal{P}_e}[Y|X_S]$. These two steps reveal a natural decomposition of the IMP, which we term as the first and second matching properties below.

Definition 4. *We say that $S \subseteq \{1, \ldots, d\}$ satisfies the* first matching property *if, for every $e \in \mathcal{E}^{all}$,*

$$\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S] = \lambda_Y \mathsf{E}_{\mathcal{P}_e}[Y|X_{PA(Y)}] + \eta_Y^\top X^e, \tag{14}$$

for some $\lambda_Y \in \mathbb{R}$ and $\eta_Y \in \mathbb{R}^d$ that do not depend on e.

First, observe that the first matching property holds for $S = PA(Y)$ since $\mathsf{E}_{l,\mathcal{P}_e}[Y|X_{PA(Y)}] = \mathsf{E}_{\mathcal{P}_e}[Y|X_{PA(Y)}] = (\beta^e)^\top X^e$. The first matching property concerns the set S such that the components in $\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S]$ that depends on e are fully captured by the causal function $\mathsf{E}_{\mathcal{P}_e}[Y|X_{PA(Y)}]$. However, this invariant relation is not directly useful for the prediction of $Y^{e^{\text{test}}}$, since the causal function can change arbitrarily with e. To this end, we identify another invariant relation from $\mathcal{M}^e$ which is called the second matching property.

Definition 5. *For $k \in \{1, \ldots, d\}$ and $R \subseteq \{1, \ldots, d\} \setminus k$, we say that a tuple (k, R) satisfies the* second matching property *if, for every $e \in \mathcal{E}^{all}$,*

$$\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R] = \lambda_X \mathsf{E}_{\mathcal{P}_e}[Y|X_{PA(Y)}] + \eta_X^\top X^e, \tag{15}$$

for some $\lambda_X \in \mathbb{R}$ and $\eta_X \in \mathbb{R}^d$ that do not depend on e.

It is straightforward to see that, if $\lambda_X \neq 0$ in the second matching property, the first and second matching properties imply the IMP as follows,

$$\begin{aligned}\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S] &= \frac{\lambda_Y}{\lambda_X}\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R] + \left(\eta_Y - \frac{\lambda_Y}{\lambda_X}\eta_X\right)^\top X^e \\ &:= \lambda \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R] + \eta^\top X^e.\end{aligned}$$

For prediction tasks under SCMs, the causal function often plays a central role. Our first and second matching properties show how the LMMSE estimator $\mathsf{E}_{l,\mathcal{P}_e}[Y|X_S]$ and the matched prediction module $\mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R]$ are connected with the causal function, respectively. Together, the two individual connections make up the IMP (illustrated in Fig. 4).

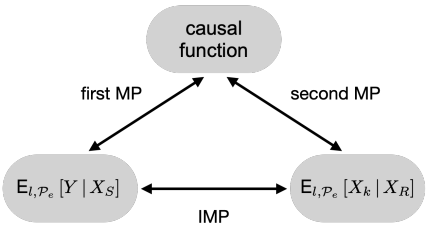

Figure 4: A triangular relation consists of the first, second, and invariant matching properties.

G More on the Characterization of IMPs

G.1 Interventions on the Response

Corollary 1. *For model $\mathcal{M}^{e,1}$, the first and second matching properties hold in the following cases.*

1. On the first MP: *The first matching property holds for $S = \{1, \ldots, d\}$.*
2. On the second MP: *For each $k \in \{j \in MB(Y) : \alpha_j^e = 0\}$ and $R = -k := \{1, \ldots, d\} \setminus k$, the second matching property holds.*

Proposition 2. *Under Assumption 1, for each (k, R) in Corollary 1, we have $\lambda_X \neq 0$ in the second matching property if $B_{-k,k}$ is not in the following hyperplane,*

$$w^\top x + b = 0, \tag{16}$$

where $w \in \mathbb{R}^{(d-1)}$ and $b \in \mathbb{R}$ are determined by the parameters in $\mathcal{M}^{e,1}$ other than $B_{-k,k}$.

The explicit expressions of w and b using the parameters in $\mathcal{M}^{e,1}$ are provided in the proof of Proposition 2. For generic choices of the parameters, the second matching property holds with $\lambda_X \neq 0$ since $B_{k,-k}$ is not necessarily on the hyperplane described in (16).

When Y is additionally intervened through the noise variance, the proof of Theorem 1 will break down in general (see Remark 2 in Appendix H.1). However, recall that the first matching property holds for $S = PA(Y)$ by definition. In this case, we provide an example for the second matching property in the following corollary.

Corollary 2. *Under Assumption 1, if Y is intervened through the noise variance in model $\mathcal{M}^{e,1}$, the second matching property holds for $k \in CH(Y)$ such that $k \notin DE(i)$ for any $i \in CH(Y) \setminus k$, and $R = \{1, \ldots, d\} \setminus DE(Y)$.*

The resulting IMPs no longer satisfy the invariance property (4), but we can use the IMP directly for the prediction of $Y^{e^{\text{test}}}$.

Remark 1. *To sum up, the class of IMPs constructed under interventions on Y only through the coefficients and shifts will in general imply the invariance property* (4)*, but it is not the case under interventions on Y through the noise variance.*

G.2 Interventions on both Predictors and Response

To generalize the setting when only Y is intervened to the general setting when X and Y are both intervened, an idea is to the merge the setting when only Y is intervened with the one when only X is intervened. The later setting has been studied in the stabilized regression framework [37]. The following set of predictors is identified (see Definition 3.4 therein),

$$X^{\text{int}}(Y) = CH^I(Y) \cup \big\{ j \in \{1, \ldots, d\} \mid \exists i \in CH^I(Y) \text{ such that } j \in DE(X_i) \big\},$$

which contains the intervened children of Y (denoted by $CH^I(Y)$) and the descendants of such children. This useful notion can be defined for each $X_j \in \{X_1, \ldots, X_d\}$, denoted by $X^{\text{int}}(X_j)$ for each X_j. When only X is intervened, the invariance principle $\mathcal{P}_e(Y|X_S) = \mathcal{P}_h(Y|X_S)$ holds

for $S^* = \{1, \dots, d\} \setminus X^{\text{int}}(Y)$; the Markov blanket of Y defined with respect to X_{S^*} is called the *stable blanket* of Y in [37]. In other words, by excluding the predictors in $X^{\text{int}}(Y)$, the target Y is blocked from the interventions on X when conditioning on X_{S^*}. This holds when Y is additionally intervened, as if only Y is intervened given X_{S^*}. In order for S^* to include as least one child of Y as in Assumption 1, we need the following assumption.

Assumption 2. *When Y is intervened, we assume that Y has at least one child that is not intervened and that child is not a descendent of some intervened child of Y.*

Based on the observation above, we identify an important class of IMPs for the general setting in the following Theorem.

Theorem 2. *For the training model $\mathcal{M}^e$ without the intervention on the noise variance of Y, the first and second matching properties hold in the following cases.*

1. On the first MP: *For $S = \{1, \dots, d\} \setminus X^{int}(Y)$, the first matching property holds.*
2. On the second MP: *For each $k \in \{1, \dots, d\} \setminus \{PE \cup X^{int}(Y)\}$, and $R = \{1, \dots, d\} \setminus \{k, X^{int}(X_k) \cup X^{int}(Y)\}$, the second matching property holds.*

Furthermore, if $\lambda_X \neq 0$ in the second matching property, then $\phi_e(X^e) = (X^e_S, \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R])^\top$ satisfies (4).

Similar to the argument in the proof of Theorem 1, the class of ϕ's from Theorem 2 will lead to the same test population loss, as they depend on the same S that is fixed in this setting. Assumption 2 is necessary for $\lambda_X \neq 0$, while sufficient conditions for $\lambda_X \neq 0$ can be found similarly as in Proposition 2. When Y is additionally intervened through the noise variance, the second matching property in Theorem 2 still holds (see Remark 3 in Appendix H.4).

H Proofs for the Theoretical Results in Section 5 and Section G

By introducing an environmental random variable $E \in \mathcal{E}^{\text{all}}$, we define a mixture of $\mathcal{M}^e$'s, $e \in \mathcal{E}^{\text{all}}$, as follows,

$$\mathcal{M} : \begin{cases} X = \gamma(E)Y + B(E)X + \varepsilon_X(E) \\ Y = \beta^\top(E)X + \varepsilon_Y(E), \end{cases}$$

where E in a root node in $\mathcal{G}(\mathcal{M})$ and the noise variables are jointly independent given E. We do not specify the distribution of E but assume that E does not have a degenerate distribution. Under this formulation, the invariance property (4) is equivalent to

$$Y \perp\!\!\!\perp E \,|\, \phi(E, X), \tag{17}$$

and the invariant, first, and second matching properties can be equivalently written as

$$\mathsf{E}_l[Y|X_S, E = e] = \lambda \mathsf{E}_l[X_k|X_R, E = e] + \eta^\top X, \tag{18}$$

$$\mathsf{E}_l[Y|X_S, E = e] = \lambda_Y \mathsf{E}[Y|X_{PA(Y)}, E = e] + \eta_Y^\top X, \tag{19}$$

$$\mathsf{E}_l[X_k|X_R, E = e] = \lambda_X \mathsf{E}[Y|X_{PA(Y)}, E = e] + \eta_X^\top X, \tag{20}$$

for $e \in \mathcal{E}^{\text{all}}$. As a special case of $\mathcal{M}$, we define a mixture of $\mathcal{M}^{e,1}$'s as

$$\mathcal{M}^1 : \begin{cases} X = \gamma Y + BX + \varepsilon_X & (21) \\ Y = (\alpha(E) + \beta)^\top X + \varepsilon_Y, & (22) \end{cases}$$

where only Y is intervened through the coefficients.

The proofs of Theorems 1, 2 and Corollaries 1, 2 will be presented under this formulation.

H.1 Proof of Theorem 1

For the first part, let $Z(E) = \alpha^\top(E)X$ denote an additional node in the acyclic graph $\mathcal{G}(\mathcal{M}^1)$, then the assignment of Y in (22) becomes

$$Y = Z(E) + \beta^\top X + \varepsilon_Y, \tag{23}$$

where E is no longer a parent of Y. Note that $Z(E) = \mathsf{E}[Y|X_{PA(Y)}, E] - \beta^\top X$ as we assume both X and Y have zero means. Since E is a root node, observe that E and Y can be d-connected through only two types of paths as follows,

1. $E \to Z(E) \to Y$,
2. $E \to Z(E) \leftarrow X_i \to \cdots \to X_l \leftarrow \cdots \leftarrow Y$,

where $i \in PE$, and $X_i \to \cdots \to X_l \leftarrow \cdots \leftarrow Y$ is a V-structure for some $l \in DE(i) \cap DE(Y)$. Note that the second type of path does not exist if Assumption 1 is not satisfied or $CH(i) \setminus Y$ is empty.

We start by showing that the d-separation $Y \perp\!\!\!\perp_{\mathcal{G}} E \,|\, \{Z(E), X_S\}$ holds given $PE \subseteq S$. First, the first path is immediately blocked by $Z(E)$. Second, for any $s \in CH(i) \setminus Y$, the path $E \to Z(E) \leftarrow X_i \to X_s$ is blocked by $\{Z(E), X_i\}$. Thus, the second path is blocked given $Z(E)$ and X_S. According to the Markov property of SCMs [32], the d-separation $Y \perp\!\!\!\perp_{\mathcal{G}} E \,|\, \{Z(E), X_S\}$ implies

$$Y \perp\!\!\!\perp E \,|\, \{Z(E), X_S\}. \tag{24}$$

Now, we prove that the above conditional independence implies the first matching property (19).

By definition, the LMMSE estimators only rely on the (finite) first two moments of the variables. Thus we start with the case when $(X, Y)|_{E=e}$ is jointly Gaussian, for each $e \in \mathcal{E}^{\text{all}}$. First we have

$$\begin{aligned}\mathsf{E}_l[Y|X_S, E=e] &\overset{(a)}{=} \mathsf{E}[Y|X_S, E=e] \\ &\overset{(b)}{=} \mathsf{E}[Y|Z(e), X_S, E=e],\end{aligned}$$

where (a) follows from the Gaussian assumption on $(X, Y)|_{E=e}$ and (b) from the fact that $Z(e)$ is a function of $\{X_S, E=e\}$ given our assumption that $PE \subseteq S$. This implies that

$$\mathsf{E}_l[Y|X_S, E] = \mathsf{E}[Y|Z(E), X_S, E] \overset{(a)}{=} \mathsf{E}[Y|Z(E), X_S],$$

where (a) follows from the conditional independence relation (24). Using the Gaussian assumption again, we have $\mathsf{E}[Y|Z(e), X_S] = \mathsf{E}_l[Y|Z(e), X_S]$. Thus putting all the pieces together, we obtain

$$\mathsf{E}_l[Y|X_S, E=e] = \mathsf{E}_l[Y|Z(e), X_S] \tag{25}$$

$$= aZ(e) + b^\top X_S, \tag{26}$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}^{|S|}$ that are not functions of E. When $(X, Y)|_{E=e}$ is non-Gaussian, one can replace it with Gaussian random variables with the matching first and second moments. Then the same argument leading to (26) still holds. Then, the first matching property follows from the fact that $Z(E) = \mathsf{E}[Y|X_{PA(Y)}, E] - \beta^\top X$.

Similarly, for the second part, we first show the following d-separation $X_k \perp\!\!\!\perp_{\mathcal{G}} E \,|\, \{Z(E), X_R\}$. Observe that X_k and E can be d-connected through two types of paths as follows,

1. $E \to Z(E) \to Y \cdots X_k$,
2. $E \to Z(E) \leftarrow X_i \cdots X_k$,

with $i \in \{j \in \{1, \ldots, d\} : \alpha_j(E) \neq 0\}$, where $Y \cdots X_k$ denotes any directed path between Y and X_k, and similarly for $X_i \cdots X_k$. The two types of paths are immediately blocked by $\{Z(E), X_R\}$ under our assumption that $PE \subseteq R$. We thus have the following when $X|_{E=e}$ is jointly Gaussian,

$$\begin{aligned}\mathsf{E}_l[X_k|X_R, E=e] &= \mathsf{E}_l[X_k|Z(e), X_R, E=e] \\ &= \mathsf{E}_l[X_k|Z(e), X_R] \\ &= cZ(e) + d^\top X_R\end{aligned} \tag{27}$$

for some $c \in \mathbb{R}$ and $d \in \mathbb{R}^{|R|}$ do not depend on E. The non-Gaussian cases can be handled in the same way as before, and thus the second property follows again from $Z(E) = \mathsf{E}[Y|X_{PA(Y)}, E] - \beta^\top X$. Observe that we have $\lambda_X = c$ in the second matching property and Assumption 1 is a necessary condition for $\lambda_X \neq 0$.

Finally, given $R \subseteq S$, we have that (27) with $c \neq 0$ (i.e., $\lambda_X \neq 0$) provides a one-to-one mapping between $\{Z(E), X_S\}$ and $\{\mathsf{E}_l[X_k|X_R, E], X_S\}$. Therefore, the conditional independence (24) is equivalent to

$$Y \perp\!\!\!\perp E \,\big|\, \{\mathsf{E}_l[X_k|X_R, E],\, X_S\}. \tag{28}$$

This implies that, using our previous notation, $\phi_e(X^e) = (X^e_S,\, \mathsf{E}_{l,\mathcal{P}_e}[X_k|X_R])^\top$ satisfies the invariance property (4), which implies[3] $\mathsf{E}_{\mathcal{P}_e}[Y|X_S] = \mathsf{E}_{\mathcal{P}_e}[Y|\phi_e(X)]$. Effectively, $\mathsf{E}_{\mathcal{P}_e}[Y|\phi_e(X)]$ serves as a representation of $\mathsf{E}_{\mathcal{P}_e}[Y|X_S]$ that is invariant. Note that Φ is not empty when $\lambda_X \neq 0$ holds with $R \subseteq S$. This implies that any $\phi \in \Phi$ with $S = \{1, \ldots, d\}$ minimizes $\mathcal{L}_{\text{test}}(f_\phi)$, since the optimality of $\phi \in \Phi$ only relies on the corresponding S.

Remark 2. *When ε_Y is replace by $\varepsilon_Y(E)$, we consider the following two cases.*

1. *The mean of ε^e_Y is a function of e, and its variance is a constant.*
2. *The variance of ε^e_Y is a function of E, and its mean can be either a function of e or a constant.*

For the first case, we can introduce $X_{d+1} := 1$ that is a parent of Y with $\alpha_{d+1} := \mathsf{E}[\varepsilon_Y(E)|E]$. Additionally, X_{d+1} is a parent of every X_j such that $\varepsilon_{X,j}$ has a non-zero mean. Specifically, the coefficient of X_{d+1} in the assignment of X_j will be $\mathsf{E}[\varepsilon_{X,j}]$. Thus, the problem reduces to the setting when ε^e_X and ε^e_Y have zero means, which has been proved in Theorem 1. For the second case, however, the varying variance of $\varepsilon_Y(E)$ cannot be separated as the mean, thus E is always a parent of Y and the path $E \to Y$ cannot be blocked by Z or any $X_S \subseteq \{X_1, \ldots, X_d\}$, i.e., the proof of Theorem 1 breaks down.

H.2 Proof of Proposition 2

The following lemma is a slight extension of Lemma 3.6 from [37], where we consider linear models with dependent noise variables rather than linear SCMs considered in [37].

Lemma 1. *Consider $V \in \mathbb{R}$ and $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ satisfying a linear model,*

$$\begin{bmatrix} V \\ X \end{bmatrix} = \begin{bmatrix} 0 & h^\top \\ g & A \end{bmatrix} \begin{bmatrix} V \\ X \end{bmatrix} + \begin{bmatrix} e_V \\ e_X \end{bmatrix} := \tilde{A} \begin{bmatrix} V \\ X \end{bmatrix} + e, \tag{29}$$

where $e_V \in \mathbb{R}$, $g, h, e_X \in \mathbb{R}^p$, and $A \in \mathbb{R}^{p \times p}$. Assume that $I - \tilde{A}$ is invertible, the population OLS estimator when regressing V on X is given by

$$\begin{aligned} \theta^{ols} = &\left[c - g^\top \tilde{\Sigma}^{-1} \left(I - c\sigma^2 g g^\top \tilde{\Sigma}^{-1}\right) v\right] h \\ &+ \left(I - A^\top\right) \left[c\sigma^2 \tilde{\Sigma}^{-1} g + \tilde{\Sigma}^{-1} \left(I - c\sigma^2 g g^\top \tilde{\Sigma}^{-1}\right) v\right], \end{aligned}$$

where $c := \left(1 + \sigma^2 g^\top \tilde{\Sigma}^{-1} g\right)^{-1}$, $\tilde{\Sigma} := \Sigma + v g^\top + g v^\top$, and

$$\mathrm{Cov}(e, e) := \begin{bmatrix} \sigma^2 & v^\top \\ v & \Sigma \end{bmatrix},$$

with $\sigma^2 := \mathrm{Var}(e_V)$, $\Sigma := \mathrm{Cov}(e_X, e_X)$, and $v := \mathrm{Cov}(e_X, e_V)$.

Proof. Let $P = [0_{1\times d}, I_{d\times d}]$, we continue from Equation A.7 in [37] as follows,

$$\begin{aligned} \theta^{\text{ols}} &= \mathrm{Cov}^{-1}(X, X)\, \mathrm{Cov}(X, V) \\ &= \left\{ P(I - \tilde{A})^{-1} \begin{bmatrix} \sigma^2 & v^\top \\ v & \Sigma \end{bmatrix} \left[(I - \tilde{A})^{-1}\right]^\top P^\top \right\}^{-1} \\ &\qquad\qquad P(I - \tilde{A})^{-1} \begin{bmatrix} \sigma^2 \\ v \end{bmatrix} + h, \end{aligned}$$

and it has been shown that

$$P(I - \tilde{A})^{-1} = [w, M],$$

[3] We use the shorthand $\phi_e(X)$ for $\phi_e(X^e)$ since the expectation is with respect to $\mathcal{P}_e$.

for $M = (I - A - gh^\top)^{-1}$ and $w = Mg$. Then,

$$\begin{aligned}\theta^{\text{ols}} &= \left[\sigma^2 ww^\top + wv^\top M^\top + Mvw^\top + M\Sigma M^\top\right]^{-1} \\ &\qquad\qquad \left(\sigma^2 w + Mv\right) + h \\ &:= \left[\sigma^2 ww^\top + M\tilde{\Sigma}M^\top\right]^{-1} \left(\sigma^2 w + Mv\right) + h,\end{aligned}$$

where $\tilde{\Sigma} := \Sigma + vg^\top + gv^\top$ and it was computed in [37] using the Sherman-Morrison formula [6] that

$$\begin{aligned}&\left[\sigma^2 ww^\top + M\tilde{\Sigma}M^\top\right]^{-1} \\ =&(I - A - gh^\top)^\top \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) M^{-1},\end{aligned}$$

with $c := \left(1 + \sigma^2 g^\top \tilde{\Sigma}^{-1} g\right)^{-1}$. Then, some simple algebra leads to

$$\begin{aligned}\theta^{\text{ols}} &= h + c\sigma^2 (I - A - gh^\top)^\top \tilde{\Sigma}^{-1} g \\ &\qquad + (I - A - gh^\top)^\top \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) v \\ &= \left[c - g^\top \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) v\right] h \\ &\quad + \left(I - A^\top\right) \left[c\sigma^2 \tilde{\Sigma}^{-1} g + \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) v\right].\end{aligned}$$

□

H.3 Proof of Proposition 2

For each tuple $(k, R, S), k \in \{j \in MB(Y) : \alpha_j^e = 0\}$. $R = -k := \{1, \ldots, d\} \setminus k$, $S = \{1, \ldots, d\}$, we prove that $\lambda_X \neq 0$. Equivalently, we prove that $\mathsf{E}_l[X_k | X_{-k}, E = e]$ is a non-constant function of e.

First, recall that, when the target variables Y is unobserved, the relations between the predictors in $\mathcal{M}^{e,1}$, are described by

$$X = \left(\gamma(\beta + \alpha^e)^\top + B\right) X + \gamma\varepsilon_Y + \varepsilon_X.$$

Now, we rewrite the above equation in the same form as the linear model (29) as follows,

$$\begin{aligned}&\begin{bmatrix} X_k \\ X_{-k} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \gamma_k(\beta + \alpha^e)_{-k}^\top + B_{k,-k} \\ \gamma_{-k}\beta_k + B_{-k,k} & \gamma_{-k}(\beta + \alpha^e)_{-k}^\top + B_{-k,-k} \end{bmatrix} \begin{bmatrix} X_k \\ X_{-k} \end{bmatrix} \\ &\qquad\qquad + \varepsilon_X + \gamma\varepsilon_Y,\end{aligned}$$

where the top-left element of the coefficient matrix is zero, i.e., $\gamma_k(\beta_k + \alpha_k^e) + B_{k,k} = 0$ since $\alpha_k^e = 0$ (by assumption), $B_{k,k} = 0$ (due to acyclicity), and $\gamma_k\beta_k = 0$ (since X_k cannot be both a child and a parent of Y). Now, by Lemma 1, the population OLS estimator when regressing X_k on X_{-k} given $E = e$ is

$$\begin{aligned}\theta^{\text{ols},k}(e) &= \left[c - g^\top \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) v\right] h \\ &\quad + \left(I - A^\top\right) \left[c\sigma^2 \tilde{\Sigma}^{-1} g + \tilde{\Sigma}^{-1} \left(I - c\sigma^2 gg^\top \tilde{\Sigma}^{-1}\right) v\right] \\ &:= ah + (I - A^\top)b, \qquad (30)\end{aligned}$$

where $h = \gamma_k(\beta + \alpha^e)_{-k} + B_{k,-k}^\top$, $v = \gamma_k \sigma_Y^2 \gamma_{-k}$, $g = \beta_k \gamma_{-k} + B_{-k,k}$, $A = \gamma_{-k}(\beta + \alpha^e)_{-k}^\top + B_{-k,-k}$, $\Sigma = \text{Cov}(N_{X_{-k}}, N_{X_{-k}})$, $\tilde{\Sigma} = \Sigma + vg^\top + gv^\top$, and $c = \left(1 + \sigma^2 g^\top \tilde{\Sigma}^{-1} g\right)^{-1}$. Note that $a \in \mathbb{R}$ and $b \in \mathbb{R}^{(d-1)}$ are not functions of e and

$$a = g^\top b + c(1 + \sigma^2 g^\top \tilde{\Sigma}^{-1} g) = g^\top b + 1. \qquad (31)$$

1. $k \in CH(Y)$: First, observe that $\beta_k = 0$ implies $g = B_{-k,k}$. By plugging h and A into (30), we obtain

$$\theta^{\text{ols},k}(e) = (a - \gamma_{-k}^\top b)\alpha_{-k}^e + a\gamma_k\beta_{-k} + (I - \beta_{-k}\gamma_{-k}^\top - B_{-k,-k}^\top)b,$$

where α_{-k}^e in the first term is non-vanishing only if

$$a - \gamma_{-k}^\top b = 1 + (B_{-k,k} - \gamma_{-k})^\top b \neq 0,$$

where we use (31).

2. $k \in PA(Y)$: Observe that $v = 0_{(d-1)}$ and $\gamma_k = 0$, then

$$\begin{aligned}\theta^{\text{ols},k}(e) &= c\alpha^e + c\beta + c\sigma^2\left(I - A^\top\right)\Sigma^{-1}g \\ &= c(1 + \sigma^2\beta_k\gamma_{-k}^\top\Sigma^{-1}\gamma_{-k} \\ &\quad + \sigma^2\gamma_{-k}^\top\Sigma^{-1}B_{-k,k})\alpha_{-k}^e + c\beta \\ &\quad + c\sigma^2(I - \beta_{-k}\gamma_{-k}^\top - B_{-k,-k}^\top)\Sigma^{-1}g,\end{aligned}$$

where the first term is not vanishing only if

$$1 + \sigma^2\beta_k\gamma_{-k}^\top\Sigma^{-1}\gamma_{-k} + \sigma^2\gamma_{-k}^\top\Sigma^{-1}B_{-k,k} \neq 0.$$

3. $k \in \{j : \exists i \in CH(X_j) \text{ such that } i \in CH(Y)\}$: Again, we have $v = 0_{(d-1)\times 1}$ and $\gamma_k = 0$. Additionally, we have $\beta_k = 0$, then α_{-k}^e in $\theta^{\text{ols},k}(e)$ is non-vanishing only if

$$1 + \sigma^2\gamma_{-k}^\top\Sigma^{-1}B_{-k,k} \neq 0.$$

H.4 Proof of Corollary 2

First, we extract the assignments of (X_k, X_R)'s from (8) with $E = e$, where $k \in CH(Y)$ and $k \notin DE(X_i)$ for any $i \in CH(Y) \setminus k$ and $R = \{1, \ldots, d\} \setminus DE(Y)$ as follows,

$$\begin{bmatrix} X_k \\ X_R \end{bmatrix} = \begin{bmatrix} 0 & \gamma_k(\beta + \alpha^e)_R^\top + B_{k,R} \\ 0_{|R|\times 1} & B_{R,R} \end{bmatrix} \begin{bmatrix} X_k \\ X_R \end{bmatrix} + \begin{bmatrix} \varepsilon_{X,k} + \gamma_k\varepsilon_Y^e \\ \varepsilon_{X,R} \end{bmatrix},$$

where the zeros in the coefficient matrix are due to the fact that all the descendants of X_k are excluded from X_R since $DE(X_k) \subseteq DE(Y)$ and $R = \{1, \ldots, d\} \setminus DE(Y)$. Note that X_k is a child of Y that is not a descendent of any other child of Y, thus any removed node $j \in \{1, \ldots, d\} \setminus \{k, R\}$ can not be the parent of any remaining node $i \in \{k, R\}$.

Now, using Lemma 1, the population OLS estimator when regressing X_k on X_R given $E = e$ is

$$\theta^{\text{ols}}(e) = \gamma_k(\beta + \alpha^e)_R + B_{k,R}^\top.$$

This implies

$$\mathsf{E}_l[X_k|X_R, E = e] = \gamma_k(\beta + \alpha^e)^\top X + B_{k,\cdot}X,$$

where we use the fact that $\beta_j = \alpha_j^e = B_{k,j} = 0$ for $j \notin R$ (recall that $\{1, \ldots, d\} \setminus R$ does not contain either the parents of X_k or parents of Y). Therefore,

$$\mathsf{E}_l[X_k|X_R, E = e] = \gamma_k\mathsf{E}[Y|X_{PA(Y)}, E = e] + B_{k,\cdot}X,$$

which is the second matching property.

Remark 3. *Observe that $\theta^{ols}(e)$ does not depend on $B_{R,R}$ or the covariance of $\varepsilon_{X,R}$, thus the second matching property holds even if every $X_j \in X_R$ is intervened.*

H.5 Proof of Theorem 2

For the first part, we only need to prove that Y and E are d-connected only through the arrow $E \to Y$ when conditioning on X_S, $S = \{1, \dots, d\} \setminus X^{\text{int}}(Y)$. The rest is simply the proof of the first part of Theorem 1. Since E is a root node, Y and E can only be d-connected through the two types of paths,

1. $E \to \cdots \to Y$,
2. $Y \to \cdots \to X_j \leftarrow \cdots \leftarrow E$,

where $j \in DE(Y) \cap DE(E)$, and $E \to \cdots \to Y$ denotes any directed path from E to Y. For the first type of paths, since $PA(Y) \subseteq S$, the only unblocked path when conditioning on X_S is $E \to Y$. For the second type of paths that are V-structures, we have $j \notin S$ since $X^{\text{int}}(Y)$ contains all the intervened children of Y and the descendants of such children, implying that the V-structures are always blocked when conditioning on X_S. Thus, the only unblocked path between E and Y is $E \to Y$ when conditioning on X_S.

For the second part, similarly, for $k \in \{1, \dots, d\} \setminus \{PE \cup X^{\text{int}}(Y)\}$ and $R = \{1, \dots, d\} \setminus \{k, X^{\text{int}}(X_k) \cup X^{\text{int}}(Y)\}$, we prove that X_k is d-connected with E only through paths that contain the subpath $E \to Y$ when conditioning on X_R (i.e., the first type of path below). Following the same idea as in the first part, Y and E are d-connected only through the arrow $E \to Y$ when conditioning on X_R since $X^{\text{int}}(Y) \cap R = \varnothing$ and $PA(Y) \subseteq R$. Then, since X_k is not intervened (i.e., $E \notin PA(X_k)$), the variables E and X_k can only be d-connected through two types of paths as follows,

1. $E \to Y \to \cdots \to X_k$
2. $E \to \cdots \to X_i \leftarrow \cdots \leftarrow X_k$,

where $i \in R$ and we use the fact that $k \notin X^{\text{int}}(Y)$ (i.e., the node k is not an intervened child of Y or a descendant of an intervened child of Y).

To handle the second type of paths, we will need two technical results. (I) If $i \in R$, then $j \in R$ for any $j \in PA(i)$ such that $j \neq k$. This can be proved by contradiction. If $j \notin R$ (i.e., $j \in X^{\text{int}}(Y) \cup X^{\text{int}}(X_k)$), then we have $i \in X^{\text{int}}(Y) \cup X^{\text{int}}(X_k)$ by the definition of $X^{\text{int}}(Y)$ and $X^{\text{int}}(Y)$, i.e., $i \notin R$. (II) For $i \in R$, observe that X_i can only be a child of X_k, otherwise the path $X_i \leftarrow X_j \leftarrow \cdots \leftarrow X_k$ is blocked by $X_j \in PA(X_i)$ when conditioning on X_R, since $i \in R$ implies $j \in R$.

Now, we proved that the second type of paths are always blocked when conditioning on X_R, by focusing on $E \to \cdots \to X_i$. For any $X_i \in CH(X_k)$, the subpath $E \to \cdots \to X_i$ cannot be $E \to X_i$, since $i \in R$ implies that X_i is not an intervened child of X_k. Finally, the path $E \to \cdots \to X_l \to X_i$ is blocked by $X_l \in PA(X_i)$ when conditioning on X_R, since $i \in R$ implies that $l \in R$.

I Algorithms

For each $e \in \mathcal{E}^{\text{train}}$, we are given the i.i.d. training data $\boldsymbol{X}_e \in \mathbb{R}^{n_e \times d}, \boldsymbol{Y}_e \in \mathbb{R}^{n_e}$, and we observe the i.i.d. test data $\boldsymbol{X}^\tau \in \mathbb{R}^{m \times d}$ and aim to predict $\boldsymbol{Y}^\tau \in \mathbb{R}^m$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with $n := \sum_{i=1}^{|\mathcal{E}^{\text{train}}|} n_e$ denote the pooled data of $\boldsymbol{X}_e$'s. In this section, we present the implementation of our method starting with the case when e is sampled from a discrete distribution with a finite support. In this setting, we expect to have $n_e \gg 1$ for every $e \in \mathcal{E}^{\text{train}}$ in general, thus it is possible to do estimation based on the data from each single environment. The challenging setting of continuous environments will be handled afterwards.

I.1 Discrete Environments

To implement our method, the main task is to identify the set of IMPs $\mathcal{I}_{\mathcal{M}}$ from the training data. For each tuple (k, R, S) in Algorithm 1, we test the following null hypothesis

$$\mathcal{H}_0 : \text{There exists } \theta \in \mathbb{R}^{d+1} \text{ such that (10) holds.}$$

We propose two test procedures.

1. *Test of the Deterministic Relation:* Since the IMPs are linear and deterministic (i.e., noiseless), we test whether the residual vector $\boldsymbol{R} \in \mathbb{R}^n$ of fitting an IMP on (k, R, S) is a zero vector or not using the test statistics,

$$T = \frac{1}{n}\boldsymbol{R}^\top \boldsymbol{R},$$

where $\boldsymbol{R}$ is a pooled data vector of $\boldsymbol{R}_e$'s defined below. To fit an IMP, we first estimate the two LMMSE estimators in (7) using OLS for each environment,

$$\hat{\boldsymbol{L}}_{e,1} := (\boldsymbol{X}_{e,S}^\top \boldsymbol{X}_{e,S})^{-1}\boldsymbol{X}_{e,S}^\top \boldsymbol{Y}_e\,,$$
$$\hat{\boldsymbol{L}}_{e,2} := (\boldsymbol{X}_{e,R}^\top \boldsymbol{X}_{e,R})^{-1}\boldsymbol{X}_{e,R}^\top \boldsymbol{X}_{e,k}\,.$$

Let $\hat{\boldsymbol{L}}_1 \in \mathbb{R}^n$ and $\hat{\boldsymbol{L}}_2 \in \mathbb{R}^n$ denote the pooled data of $\hat{\boldsymbol{L}}_{e,1}$'s and $\hat{\boldsymbol{L}}_{e,2}$'s, respectively. *It is noteworthy that $\hat{\boldsymbol{L}}_2$ only depends on $\boldsymbol{X}$, thus $\hat{\boldsymbol{L}}_2^\tau$ for the test data can be computed similarly using $\boldsymbol{X}^\tau$.* Next we estimate the matching parameter using OLS on the pooled data (recall that the matching parameter cannot be identified using the data from a single environment). The OLS estimator of the matching parameter is

$$\hat{\theta} := (\hat{\eta}^\top, \hat{\lambda})^\top = ([\boldsymbol{X}_S, \hat{\boldsymbol{L}}_2]^\top [\boldsymbol{X}_S, \hat{\boldsymbol{L}}_2])^{-1}[\boldsymbol{X}_S, \hat{\boldsymbol{L}}_2]^\top \hat{\boldsymbol{L}}_1\,.$$

For each $e \in \mathcal{E}^{\text{train}}$, we obtain the residual vector of fitting an IMP

$$\boldsymbol{R}_e = \hat{\boldsymbol{L}}_{e,1} - \hat{\lambda}\hat{\boldsymbol{L}}_{e,2} - \boldsymbol{X}_{e,S}\hat{\eta}\,.$$

2. *Approximate Test of Invariant Residual Distributions*: According to the invariance property (5), we test whether the residual when regressing $\boldsymbol{Y}$ on $[\boldsymbol{X}_S, \boldsymbol{L}_2]$ has constant mean and variance. Specifically, we use the t-test and F-test with corrections for multiple hypothesis testing from [34] (see Section 2.1 Method II). The test yields a p-value.

The test statistic from the first procedure and the p-value from the second procedure quantify how likely an IMP holds (i.e., the smaller the more likely), and thus we will refer to either one of them as an IMP score denoted by s_{IMP}. Let $\hat{\mathcal{I}} = \{(k, R, S) : s_{\text{IMP}}(k, R, S) < c_{\text{IMP}}\}$ denote the set of IMPs identified from the training data, where c_{IMP} is some cutoff parameter. Then, since IMPs are not equally predictive in general, we focus on the most predictive ones by introducing the mean squared prediction error as a prediction score s_{pred}, and we select the set of IMPs that are more predictive $\widehat{\mathcal{I}_{\text{pred}}} = \{(k, R, S) \in \hat{\mathcal{I}} : s_{\text{pred}}(k, R, S) < c_{\text{pred}}\}$ with some cutoff parameter c_{pred}. For the second IMP score that is a p-value, the cutoff parameter c_{IMP} is simply a significance level that is fixed to 0.05 in this work. For choosing the rest cutoff parameters, we follow a bootstrap procedure from [36] with one subtle difference: We sample the same amount of bootstrap samples from each environment rather than sampling over the pool data as in [36], since our procedure involves estimations using the data from each environment.

Algorithm 1 Invariant Prediction using the IMP (discrete)

procedure IDENTIFY IMPS FROM THE TRAINING DATA
  for $k \in \{1, \ldots, d\}$, $S \subseteq \{1, \ldots, d\}$, $R \subseteq S \setminus k$ **do**
    Compute the IMP score s_{imp} and the prediction score s_{pred} for $i = (k, R, S)$
    Regress $\boldsymbol{Y}$ on $[\boldsymbol{X}_S, \hat{\boldsymbol{L}}_2]$ to obtain f_i
  Identify $\hat{\mathcal{I}}$ and $\widehat{\mathcal{I}_{\text{pred}}}$
procedure PREDICTION ON THE TESTING DATA
  $\hat{\boldsymbol{Y}}^\tau = \frac{1}{|\widehat{\mathcal{I}_{\text{pred}}}|}\sum_{i \in \widehat{\mathcal{I}_{\text{pred}}}} f_i(\boldsymbol{X}_S^\tau, \hat{\boldsymbol{L}}_2^\tau)$

In practice, there can be spurious IMPs that have extremely small IMP scores but have large prediction scores, e.g., when Y is independent of X_S and X_k is independent of X_R. To this end, we will pre-select (k, R, S)'s with prediction scores smaller than the median of the all the computed prediction scores before identifying $\hat{\mathcal{I}}$. If the regression function f_i in Algorithm 1 is chosen to be linear, one can use the IMP directly, i.e.,

$$\hat{\boldsymbol{Y}}_{\text{IMP}}(k, R, S) = [\boldsymbol{X}_S, \hat{\boldsymbol{L}}_2]\hat{\theta}, \tag{32}$$

which we call the *discrete IMP estimator* denoted by IMP_{d}. To make use of all the IMPs selected in $\widehat{\mathcal{I}_{\text{pred}}}$, we use an averaging step for the prediction of $\boldsymbol{Y}^\tau$ in Algorithm 1.

I.2 Continuous Environments

To model continuous environments, we introduce an environmental variable U that is a *continuous* random variable with support $\mathcal{U}$. Apparently, this is a much more challenging setting compared with the discrete environment case, as we only have one training data sample for each $u \in \mathcal{U}$, making the OLS a poor estimate of $\mathsf{E}_{l,\mathcal{P}_u}[Y|X_S]$. Fortunately, it turns out that we can leverage the *semi-parametric varying coefficient* (SVC) models [13] (see Appendix **??**) to remedy this issue. In particular, we estimate $\mathsf{E}_{l,\mathcal{P}_u}[Y|X_S]$ by fitting,

$$Y = M + \beta^\top Z + N \;\text{ with }\; M = \alpha^\top(U)W, \tag{33}$$

where N is independent of U and the two vectors of predictors $W \in \mathbb{R}^p$ (for the varying coefficient) and $Z \in \mathbb{R}^q$ (for the invariant coefficients) with $p + q = |S|$. Since we assume $N \perp\!\!\!\perp U$, we focus on the settings when Y is not intervened through the noise variance.

Remark 4. *Our estimation procedure for the discrete environments can also be formulated under the SVC model with a discrete random variable U, where we treat all the coefficients as varying coefficients (i.e., $\beta = \mathbf{0}$), and IMP$_d$ becomes an estimate of M.*

An SVC model over $(Y^\tau, W^\tau, Z^\tau, N^\tau)$ for the test data can be defined similarly, where $\sigma^2 = \mathsf{E}[(N^\tau)^2]$ is the population generalization error of the IMP estimator. Observe that the linear SCM $\mathcal{M}^u$ in (**??**) can be viewed as a collection of SVC models parameterized by $U = u$. Thus the estimation tasks for the linear SCMs from continuous environments can greatly benefit from the existing theories developed for SVC models. More precisely, we employ the following estimate

$$\widehat{\mathsf{E}_{l,\mathcal{P}_u}[Y|X_S]} = \hat{M}|_{U=u} + \hat{\beta}^\top Z,$$

where the profile least-squares estimation of β and M proposed in [13] can be found in Appendix **??**. Similarly, $\mathsf{E}_{l,\mathcal{P}_u}[X_k|X_R]$ can be estimated by fitting another semi-parametric varying coefficient model

$$V = M_V + \beta_V^\top Z_V + N_V \;\text{ with }\; M_V = \alpha_V^\top(U)W, \tag{34}$$

where V denotes any X_k, and X_R is divided into $Z_V \in \mathbb{R}^r$ and W.

It is noteworthy that the two SVC models share the same set of predictors with varying coefficients, which we explain below. A challenge for fitting such models is that the vector of predictors with varying coefficients, namely W, needs to be known. For continuous environments, we focus on discovering IMPs that can be decomposed into the first and second matching properties. Thus, since the causal function captures the predictors with varying coefficients, the first and second matching properties imply that the vector W is simply X_{PE}, i.e., the parents of Y with varying coefficients in $\mathcal{M}^u$, for both models.

Based on this observation and Theorem 1, we replace the exhaustive search over (k, R, S) in Algorithm 1 by a search over (P, k, R, S) according to the conditions in Theorem 1 with $PE = P$. That is, we choose (P, k, R, S) from

$$P \subseteq \{1, \ldots, d\}, \quad k \in \{1, \ldots, d\} \setminus P,$$
$$P \subseteq S \subseteq \{1, \ldots d\}, \quad P \subseteq R \subseteq S \setminus k,$$

such that $W = X_P$, $Z = X_{S\setminus P}$, $V = X_k$, and $Z_V = X_{R\setminus P}$.

Unlike IMP$_\text{d}$ in (32), we make use of the fact that β is invariant and propose the *continuous IMP estimator* denoted by IMP$_\text{c}$ as follows

$$\hat{\boldsymbol{Y}}_{\text{IMP}}(P, k, R, S) = [\boldsymbol{W}, \hat{\boldsymbol{M}}_V]\hat{w} + \boldsymbol{Z}\hat{\beta}, \tag{35}$$

with the matching parameter $w \in \mathbb{R}^{p+1}$ estimated by

$$\hat{w} = ([\boldsymbol{W}, \hat{\boldsymbol{M}}_V]^\top[\boldsymbol{W}, \hat{\boldsymbol{M}}_V])^{-1}[\boldsymbol{W}, \hat{\boldsymbol{M}}_V]^\top\hat{\boldsymbol{M}}. \tag{36}$$

The data matrices for the two models (e.g., $\boldsymbol{W} \in \mathbb{R}^{n\times p}$) can be defined accordingly and we provide the details in Appendix M.1. In this case, the residual vector for the first IMP score is given by

$$\boldsymbol{R} = \boldsymbol{M} - [\boldsymbol{W}, \hat{\boldsymbol{M}}_V]\hat{w}.$$

Note the the second IMP score is not applicable for continuous environments due to the small sample size in each environment, thus we focus the first IMP score.

Remark 5. *The IMP$_d$ and IMP$_c$ are similar in spirit, the IMP$_c$ relies on the first and second matching properties for identifying W (so we can reuse Z in* (35)*), whereas IMP$_d$ directly tests the IMP since the estimation process treats all the coefficients as varying coefficients (also see the proof of Corollary 4).*

J Experiments

The prediction performance is measured by the mean residual sum of squares (RSS) on the test environments. We compare our method with several baseline methods: Ordinary Least Squares (OLS), stabilized regression (SR) [37], anchor regression (AR) [42]. We have also compared with domain invariant projection (DIP) [5], conditional invariance penalty (CIP) [21], conditional invariant residual matching (CIRM) [9], and invariant risk minimization (IRM) [2]; it turns out that the empirical performance of these methods are not as competitive as the other baselines in our experimental settings, thus we do not report them below.

The two IMP scores lead to two versions of our algorithm, and we refer to the first one IMP and the second one IMP_{inv} (as it tests the invariance of the noise mean and variance). We focus on linear functions f_i's in Algorithm 1, namely, we use the IMP estimators. For the profile likelihood estimation, we adopt the Epanechnikov kernel $k(u) = 0.75 \max(1 - u^2, 0)$ with the bandwidth fixed to be 0.1. We test DIP, CIP, CIRM, and their variants provided in [9] with the default parameters. For the anchor regression, we use a 5-fold cross-validation procedure to select the hyper-parameter γ from $\{0, 0.05, 0.1, \ldots, 0.5\}$. The significance levels are fixed to be 0.05 for all methods. We randomly simulate 500 data sets for each experiment, if not mentioned otherwise.

J.1 Discrete environments

First, we generate linear SCMs $\mathcal{M}^e$'s without interventions. For each $e \in \mathcal{E}^{\text{train}} = \{1, \ldots, 5\}$ or $e \in \mathcal{E}^{\text{test}} = \{6, \ldots, 10\}$, we randomly generate a linear SCM with 11 variables as follows. The graph $\mathcal{G}(\mathcal{M}^e)$ is specified by a lower triangular matrix of i.i.d. $\text{Bernoulli}(1/2)$ random variables. The response Y is randomly selected from the 11 variables and we require that Y has a least one parent and one child in $\mathcal{G}(\mathcal{M}^e)$. For each linear SCM, the non-zero coefficients are sampled from $\text{Unif}[-1.5, -0.5] \cup [0.5, 1.5]$ and the noise variables are standard normal. For each training or test environment, we simulate i.i.d. data of sample size 300.

J.1.1 Interventions on X

Since the baseline methods have been examined extensively under shift interventions, we focus on shift interventions on X for the comparison. The general interventions on X will be considered in Section J.3. Specifically, for each training environment, we randomly selected 4 predictors to be intervened through shifts sampled from $\text{Unif}[-2, 2]$. For each test environment, the shifts are sampled from $\text{Unif}[-10, 10]$. In Fig. 5, IMP_{inv} performs similarly to SR since they share a similar idea when only X is intervened. But our IMP method can potentially improve upon these two methods, since there are IMPs beyond the ones that imply invariance (see an example in (**??**)), which is left for future work. Due to the averaging steps of IMP, IMP_{inv}, and SR, they have smaller variances compared with OLS and AR (a similar result has been reported in [36]).

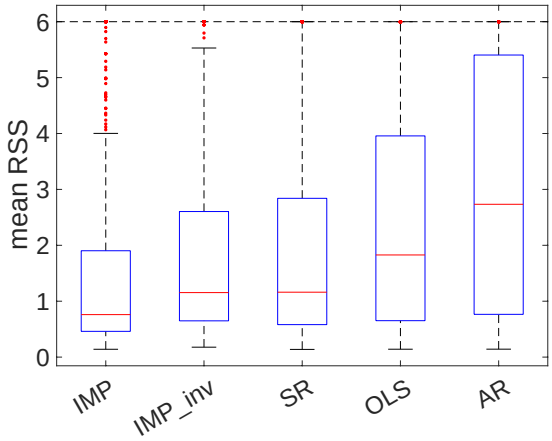

Figure 5: Experiment 7.1.1

J.1.2 Interventions on Y

We consider the response Y to be intervened through both the coefficients and shifts. We randomly select $n_p \sim \text{Unif}\{1, \ldots, |PA(Y)|\}$ of parents of Y to have varying coefficients. For each training environment, we add perturbation terms sampled from $\text{Unif}[-2, 2]$ to the original coefficients. For each test environment, the perturbations are sampled from $\text{Unif}[-10, 10]$. The shift intervention on Y is the same as the shift interventions on X in Section J.1.1. In this setting, since none of the baseline methods allow interventions on Y through the coefficients, they cannot even improve upon OLS. In Fig. 6, IMP performs slightly better than IMP_{inv}, which may due to the fact that the IMP method aims to find all possible IMPs, but the IMP_{inv} only looks for IMPs that imply invariance.

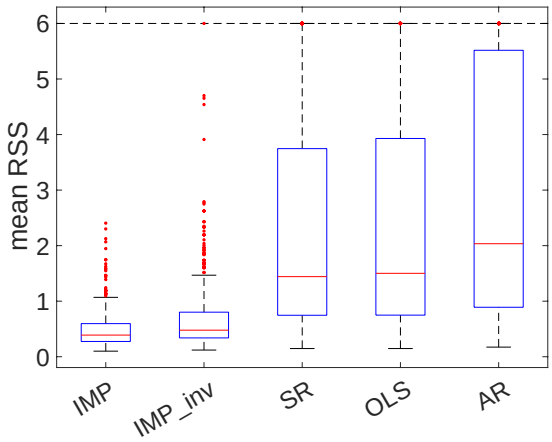

Figure 6: Experiment 7.1.2

J.1.3 Interventions on both X and Y

The setting of interventions on both X and Y is simply a combination of the two setting above. We require that Y has at least one child that is not intervened when selecting the predictors to be intervened. In this challenging setting, our method outperforms the baselines by a large margin.

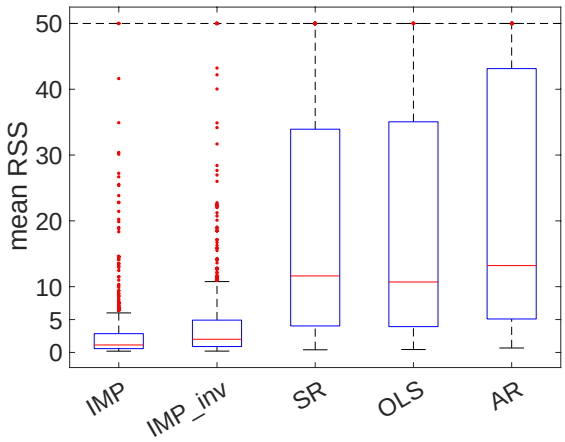

Figure 7: Experiment 7.1.3

J.2 Interventions on both X and Y (continuous)

In this more challenging setting, we only compare with OLS, since AR only considers shift interventions while the other baselines are not developed for continuous environments, and also recall

that IMP_{inv} is proposed for discrete environments. First, we define $\{U_1, \ldots, U_{800}\}$ sampled from $\text{Unif}[0, 1]$ and $\{U_1^\tau, \ldots, U_{800}^\tau\}$ sampled from $\text{Unif}[1, 2]$ as the training and test environments, respectively. Similar to the setting of discrete environments, we randomly generate the linear SCMs without interventions first and then add interventions to the model. Due to the high computational complexity, we focus on graphs with 5 nodes where 2 predictors are intervened. The interventions on the coefficients and shift interventions are defined by adding a perturbation term $a \sin(2\pi w U_i)$, where w is sampled from $\text{Unif}[0.5, 2]$. The parameter a is fixed to be 2 for the training environments (i.e., the same range as for the discrete case) and 5 for the test environments. Since the parameter space is much smaller than the previous experiments, we only generated 100 data sets. Overall, the performance of our IMP algorithm is similar to that in the discrete setting.

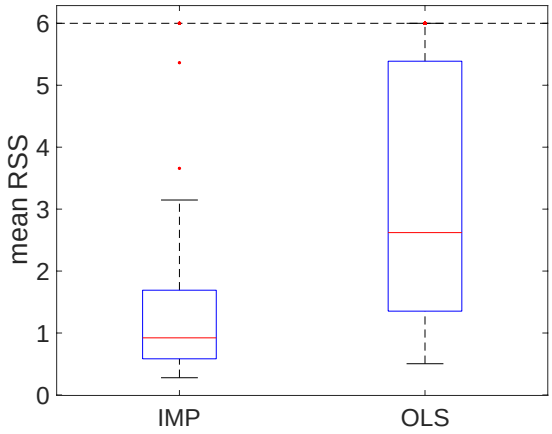

Figure 8: Experiment 7.2

J.3 Robustness

In the previous experiments, we only consider shift interventions on X and interventions on Y other than the noise variance. In this experiment, we consider an extreme case when only a child of Y with the highest causal ordering is not intervened (i.e., Assumption 2 is satisfied), *all other variables* are intervened through *every parameter*. Specifically, a shift intervention or an intervention on the coefficient is defined by adding a perturbation term to the original parameter. The perturbation term is sampled from $\text{Unif}[-2, 2]$ for the training data, and from $\text{Unif}[-5, 5]$ for the test data. The intervened noise variances are sampled from $\text{Unif}[0.75, 1.25]$ for each training environment, and from $\text{Unif}[0.5, 1.5]$ for each test environment. To test how sensitive our method is with respect to Assumption 2 in this challenging setting, we gradually add interventions to the child of Y that is not intervened, where the shifts and coefficient interventions are sampled from $\text{Unif}[-2\lambda, 2\lambda]$ and $\text{Unif}[-5\lambda, 5\lambda]$ for the training and test environments, respectively. The intervened noise variances are sample from $\text{Unif}[1 - 0.25\lambda, 1 + 0.25\lambda]$ for training, and from $\text{Unif}[1 - 0.5\lambda, 1 + 0.5\lambda]$ for testing. The parameter $\lambda \in [0, 1]$ controls the intervention strength. Due to the results in Section J.1.3, it would not be informative to compare with the baseline methods, so we focus on our IMP method. Note that our IMP_{inv} is also not included since IMPs will not imply invariance in this case (see Remark 1). Overall, as shown in Fig. 9, the median of the mean RSS is not too sensitive with respect to mild interventions on the child that is not intervened, but the variance increases rapidly.

K Note on Semi-parametric Varying Coefficient Models and Profile Least-Squares Estimation

First, we introduce the semi-parametric varying coefficient model following most of the notation in [13]. Consider $Y \in \mathbb{R}$, $U \in \mathcal{U}$, and two vectors of predictors $W = (W_1, \ldots, W_p)^\top$ and $Z = (Z_1, \ldots, Z_q)^\top$ such that Z_j's have invariant coefficients, a semi-parametric varying coefficient

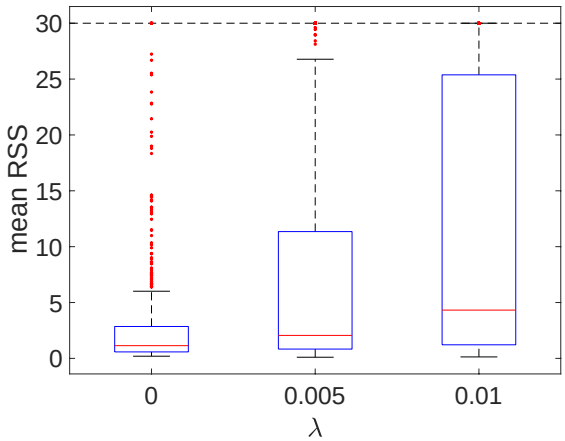

Figure 9: Experiment 7.3

model over (U, Y, W, Z) is defined by

$$Y = M + \beta^\top Z + N, \quad M = \alpha^\top(U)W, \tag{37}$$

where N is independent of (U, W, Z).

We briefly introduce the profile least-squares estimator of β proposed in [13]. Denote n i.i.d. samples of (U, Y, Z, N) as $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{W} = (\boldsymbol{W_1}, \ldots, \boldsymbol{W_n})^\top$, $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ip})^\top$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\top$, $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iq})^\top$, and $\boldsymbol{N} = (N_1, \ldots, N_n)^\top$. We thus have $\boldsymbol{M} = (M_1, \ldots, M_n)^\top$ with $M_i = \alpha^\top(U_i)\boldsymbol{W}_i$. Let $\boldsymbol{K}_u = \mathrm{diag}(K_h(U_1 - u), \ldots, K_h(U_n - u))$ for some kernel function $K_h(\cdot) = K(\cdot/h)/h$ with bandwidth h, and

$$\tilde{\boldsymbol{W}}_u = \begin{bmatrix} \boldsymbol{W_1}^\top & \frac{U_1 - u}{h}\boldsymbol{W_1}^\top \\ \vdots & \vdots \\ \boldsymbol{W_n}^\top & \frac{U_n - u}{h}\boldsymbol{W_n}^\top \end{bmatrix} \in \mathbb{R}^{n \times 2p}.$$

For u in a neighborhood of $u_0 \in \mathcal{U}$, assume that each function $\alpha_i(u)$ in (37) can be approximated locally by the first-order Taylor expansion $\alpha_i(u) \approx \alpha_i + b_i(u - u_0)$. Then the varying coefficients $\alpha(U_1), \ldots, \alpha(U_n)$ can be estimated by solving the weighted local least squares problem

$$\min_{\{a_i, b_i\}} \sum_{k=1}^{n} \left[\boldsymbol{Y}_k - \sum_{j=1}^{q} \beta_j Z_{kj} - \sum_{i=1}^{p} (a_i + b_i(U_k - u))W_{ki} \right]^2 \cdot K_h(U_k - u), \tag{38}$$

which has the solution

$$[\hat{a}_1(u), \ldots, \hat{a}_p(u), h\hat{b}_1(u), \ldots, h\hat{b}_p(u)] = (\tilde{\boldsymbol{X}}_u^\top \boldsymbol{K}_u \tilde{\boldsymbol{X}}_u)^{-1} \tilde{\boldsymbol{X}}_u^\top \boldsymbol{K}_u (\boldsymbol{Y} - \boldsymbol{Z}\beta),$$

for $u \in \{U_1, \ldots, U_n\}$. Then, each variable $\boldsymbol{M}_i$ can be estimated by $\hat{\boldsymbol{M}}_i = \sum_{j=1}^{p} \hat{a}_j(u)\boldsymbol{X}_{ij}$, and thus the vector $\boldsymbol{M}$ can be estimated by

$$\hat{\boldsymbol{M}}(\beta) = \begin{bmatrix} [\boldsymbol{W}_1^\top \quad 0_{1\times p}]\{\tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1} \tilde{\boldsymbol{W}}_{u_1}\}^{-1} \tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1} \\ \vdots \\ [\boldsymbol{W}_n^\top \quad 0_{1\times p}]\{\tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n} \tilde{\boldsymbol{W}}_{u_n}\}^{-1} \tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n} \end{bmatrix} (\boldsymbol{Y} - \boldsymbol{Z}\beta) := A(\boldsymbol{Y} - \boldsymbol{Z}\beta), \tag{39}$$

which depends on the unknown parameter β that will be estimated below. Substituting $\hat{\boldsymbol{M}}(\beta)$ into the vector form of (37), we obtain $(I-A)\boldsymbol{Y} = (I-A)\boldsymbol{Z}\beta + \boldsymbol{N}$. The profile least-squares estimator of β is given by

$$\hat{\beta} = \{\boldsymbol{Z}^\top (I-A)^\top (I-A)\boldsymbol{Z}\}^{-1} \cdot \boldsymbol{Z}^\top (I-A)^\top (I-A)\boldsymbol{Y}. \tag{40}$$

Finally, by replacing β in (39) with $\hat{\beta}$, the final form of the estimator for $\boldsymbol{M}$ is given by

$$\hat{\boldsymbol{M}} = A(\boldsymbol{Y} - \boldsymbol{Z}\hat{\beta}). \tag{41}$$

L Asymptotic Generalization Error

In this section, we provide the asymptotic generalization errors (as $n, m \to \infty$) of the IMP_c and IMP_d estimators for (k, R, S)'s that satisfy IMPs, i.e., $(k, R, S) \in \mathcal{I}_{\mathcal{M}}$. Recall that $\sigma^2 = \mathsf{E}[(N^\tau)^2]$ is the population generalization error of the IMP estimators. Due to the estimations on both training and test data, the asymptotic generalization error can be decompose to the error terms depend on the training data size n and the test data size m as follows. Let $c_n = \left\{\frac{\log(1/h)}{nh}\right\}^{1/2} + h^2$, where h is the kernel bandwidth (see Appendix K for details).

Theorem 3. *For any $(k, R, S) \in \mathcal{I}_{\mathcal{M}}$, under the technical assumptions in Appendix M.1, the asymptotic generalization error of the IMP_c estimator is given by*

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i^\tau - Y_i^\tau)^2 = \sigma^2 + O_p(c_n \vee n^{-1/2}) + O_p(c_m \vee m^{-1/2}).$$

The following corollary considers the setting when the amount of unlabeled training and test data grows in a higher order than the amount of labels in the training data. The generalization error due to the estimation on the test data disappears.

Corollary 3. *Given i.i.d. training data of size n with $0 < l_n < n$ labels and test data of size m, if $\max(\frac{l_n}{n}, \frac{l_n}{m}) \to 0$ as $\min(m, n) \to \infty$, under the technical assumptions in Appendix M.1, the asymptotic generalization error of the IMP_c estimator is given by*

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i^\tau - Y_i^\tau)^2 = \sigma^2 + O_p(c_{l_n} \vee l_n^{-1/2}).$$

The setting of discrete environments can be viewed as a special case of continuous environments, where the error term $c_n \vee n^{-1/2}$ due to the kernel estimation procedure is replaced by an error term from multiple OLS estimations.

Corollary 4. *For any $(k, R, S) \in \mathcal{I}_{\mathcal{M}}$, under the technical assumptions in Appendix M.1, the asymptotic generalization error of the IMP_d estimator is given by*

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i^\tau - Y_i^\tau)^2 = \sigma^2 + O_p(a_n) + O_p(a_m),$$

where $a_n = (\min_{e\in\mathcal{E}^{train}} n_e)^{-1/2}$ and $a_m = m^{-1/2}$.

This asymptotic generalization error heavily depends on the environment with the smallest sample size, which also supports the fact that IMP_d should not be employed for continuous environment settings.

M Proofs for the Theoretical Results in Section L

M.1 Technical Lemmas for the Proof of Theorem 3

First, we present some technical assumptions and two technical lemmas from [13]. Let $c_n' = \left\{\frac{\log(1/h)}{nh}\right\}^{1/2}$ and $c_n = c_n' + h^2$.

1. U has a bounded support $\mathcal{U}$ and has density function $f(\cdot)$ that is Lipschitz continuous and bounded away from 0.

2. For each U_i, the matrix $\mathsf{E}[W^\top W|U_i]$ is non-singular, and the matrices $\mathsf{E}[W^\top W|U_i]$, $\big(\mathsf{E}[W^\top W|U_i]\big)^{-1}$, and $\mathsf{E}[WZ^\top|U_i]$ are Lipschitz continuous.

3. $\alpha_1(u), \ldots, \alpha_p(u)$ have continuous second derivatives.

4. $K(\cdot)$ is a symmetric density function.

Lemma 2 ([29]). *Let $(U_1, Y_1), \ldots, (U_n, Y_n)$ be i.i.d. random vectors in $\mathbb{R}^2$. Assume that $\mathsf{E}[|Y|^s] < \infty$ and $\sup_x \int |y|^s f(u, y)dy < \infty$, where $f(u, y)$ is the density of (U, Y). Let K be a bounded positive function with a bounded support that satisfies a Lipschitz condition. Given that $n^{2\varepsilon-1}h \to \infty$ for some $\varepsilon < 1 - s^{-1}$, then*

$$\sup_u \left| \frac{1}{n} \sum_{i=1}^n K_h(U_i - u)Y_i - \mathsf{E}[K_h(U_i - u)Y_i] \right| = O_p(c_n').$$

Lemma 3 ([13]). *Under suitable assumptions, $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ as $n \to \infty$, where $\Sigma = \mathrm{Var}(N) \cdot C^{-1}$ with*

$$C = \mathsf{E}[ZZ^\top] - \mathsf{E}\big[\mathsf{E}[ZW^\top|U]\mathsf{E}[WW^\top|U]\mathsf{E}[WZ^\top|U]\big].$$

In the following lemma, we provide the rates of several quantities needed for the proof of Theorem 3.

Lemma 4. *Under the same assumptions as in Lemma 3,*

$$\begin{aligned}
R_1 &= \frac{1}{n}(\hat{\boldsymbol{M}} - \boldsymbol{M})^\top(\hat{\boldsymbol{M}} - \boldsymbol{M}) = O_p(c_n^2 \vee n^{-1}),\\
R_2 &= \frac{1}{n}(\hat{\boldsymbol{M}} - \boldsymbol{M})^\top[\boldsymbol{W}, \boldsymbol{M}] = O_p(c_n \vee n^{-1/2}),\\
R_3 &= \frac{1}{n}(\hat{\boldsymbol{M}} - \boldsymbol{M})^\top \boldsymbol{Z} = O_p(c_n \vee n^{-1/2}),\\
R_4 &= \frac{1}{n}(\hat{\boldsymbol{M}} - \boldsymbol{M})^\top \boldsymbol{N} = O_p(c_n).
\end{aligned}$$

Proof. First, (41) and the vector form of (37) give

$$\begin{aligned}
\hat{\boldsymbol{M}} - \boldsymbol{M} &= A(\boldsymbol{Y} - \boldsymbol{Z}\hat{\beta}) - \boldsymbol{M}\\
&= (A - I)\boldsymbol{M} + A\boldsymbol{Z}(\beta - \hat{\beta}) + A\boldsymbol{N}.
\end{aligned}$$

Observe that $R_2 \sim R_4$ can be defined through

$$\begin{aligned}
I_1 &= \frac{1}{n}\boldsymbol{M}^\top(A^\top - I)\boldsymbol{P},\\
I_2 &= \frac{1}{n}(\beta - \hat{\beta})^\top \boldsymbol{Z}^\top A^\top \boldsymbol{P},\\
I_3 &= \frac{1}{n}\boldsymbol{N}^\top A^\top \boldsymbol{P},
\end{aligned}$$

where $\boldsymbol{P}$ can be replaced by $\boldsymbol{W}$, $\boldsymbol{Z}$, $\boldsymbol{M}$, or $\boldsymbol{N}$ (note that the rows of $\boldsymbol{P}$ are i.i.d.). We also have $R_1 = \sum_{i=4}^{9} I_i$, where

$$
\begin{aligned}
I_4 &= \frac{1}{n}\boldsymbol{M}^\top (A^\top - I)(A - I)\boldsymbol{M}, \\
I_5 &= \frac{1}{n}(\beta - \hat{\beta})^\top \boldsymbol{Z}^\top A^\top A \boldsymbol{Z}(\beta - \hat{\beta}), \\
I_6 &= \frac{1}{n}\boldsymbol{N}^\top A^\top A \boldsymbol{N}, \\
I_7 &= \frac{1}{n}\boldsymbol{M}^\top (A^\top - I)\Big\{ A\boldsymbol{Z}(\beta - \hat{\beta}) + A\boldsymbol{N} \Big\}, \\
I_8 &= \frac{1}{n}(\beta - \hat{\beta})^\top \boldsymbol{Z}^\top A^\top \Big\{ (A - I)\boldsymbol{M} + A\boldsymbol{N} \Big\}, \\
I_9 &= \frac{1}{n}\boldsymbol{N}^\top A^\top \Big\{ (A - I)\boldsymbol{M} + A\boldsymbol{Z}(\beta - \hat{\beta}) \Big\}.
\end{aligned}
$$

It can be shown that $I_1, I_3 = O_p(c_n)$ (we use this as a shorthand for $I_1 = O_p(c_n), I_3 = O_p(c_n)$, as I_1 and I_3 are not equal), $I_2 = O_p(n^{-1/2})$, $I_5 = O_p(n^{-1})$, $I_4, I_6 = O_p(c_n^2)$, $I_8 = O_p(c_n n^{-1/2})$, $I_7, I_9 = O_p(c_n^2 \vee c_n n^{-1/2})$, which implies $R_1 = O_p(c_n^2 \vee n^{-1})$, $R_2, R_3 = O_p(c_n \vee n^{-1/2})$ and $R_4 = O_p(c_n)$. The techniques for proving the rates of $I_1 \sim I_9$ are similar; observe that all the components in $I_4 \sim I_9$ are already computed to obtain $I_1 \sim I_3$, thus we only provide the proof for $I_1 \sim I_3$ for simplicity of presentation.

Using Lemma 2, it has been shown in [13] that $\tilde{\boldsymbol{W}}_{u_i}^\top \boldsymbol{K}_{u_i} \tilde{\boldsymbol{W}}_{u_i}$ can be equivalently expressed as

$$
\begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} = nf(U_i)\mathsf{E}[WW^\top | U_i] \otimes \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \{1 + O_p(c_n)\}, \tag{42}
$$

where $\mu_2 = \int_{\mathcal{U}} u^2 K(u) du$, and the four block matrices are $B_1 = S_0(U_i)$, $B_2 = B_3 = S_1(U_i)$, and $B_4 = S_2(U_i)$, with respect to

$$
S_k(U_i) = \sum_{j=1}^{n} \left(\frac{U_j - U_i}{h} \right)^k \boldsymbol{W}_j \boldsymbol{W}_j^\top K_h(U_j - U_i).
$$

Since the techniques for proving (42), omitted in [13], will be used repeatedly in the rest of this paper, we provide the proof for completeness. Applying Lemma 2, it holds uniformly in $u \in \mathcal{U}$ that $S_k(u)$ can be expressed as

$$
\begin{aligned}
&\mathsf{E}\left[\left(\frac{U - u}{h} \right)^k \mathsf{E}[WW^\top | U] K_h(U - u) \right] \{1 + O_p(c_n')\} \\
&= \int_{\mathcal{V}} v^k \mathsf{E}[WW^\top | U = hv + u] K(v) f(hv + u) dv \\
&\qquad\qquad \cdot \{1 + O_p(c_n')\} \qquad (43) \\
&= \int_{\mathcal{V}} v^k K(v) \left[\mathsf{E}[WW^\top | U = u] + vO(h) \right] \\
&\qquad\qquad \cdot [f(u) + vO(h)] \, dv \{1 + O_p(c_n')\} \qquad (44) \\
&= \begin{cases} \mathsf{E}[WW^\top | U = u] f(u) \mu_k \{1 + O_p(c_n)\}, & k \text{ even} \\ O(h) + O_p(c_n'), & k \text{ odd} \end{cases} \qquad (45)
\end{aligned}
$$

where (43) is due to the change of variable $V = (U - u)/h$, (44) uses the Lipschitz continuity assumptions on $\mathsf{E}[XX^T | U]$ and $f(\cdot)$, and (45) is by the symmetry of the kernel function $K(\cdot)$. Similarly, we obtain

$$
\begin{aligned}
&\tilde{\boldsymbol{W}}_{u_i}^\top \boldsymbol{K}_{u_i} \boldsymbol{M} \\
&= \begin{bmatrix} \sum_{j=1}^{n} \boldsymbol{W}_j \boldsymbol{W}_j^\top \alpha(U_j) K_h(U_j - U_i) \\ \sum_{j=1}^{n} \frac{U_j - U_i}{h} \boldsymbol{W}_j \boldsymbol{W}_j^\top \alpha(U_j) K_h(U_j - U_i) \end{bmatrix} \\
&= nf(U_i)\mathsf{E}[WW^\top | U_i] \alpha(U_i) \otimes (1 \quad 0)^\top \{1 + O_p(c_n)\}. \qquad (46)
\end{aligned}
$$

Recall the expression of A in (39), we have

$$
\begin{aligned}
&(A-I)\boldsymbol{M} \\
&= \begin{bmatrix} [\boldsymbol{W}_1^\top \quad \mathbf{0}]\{\tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1}\tilde{\boldsymbol{W}}_{u_1}\}^{-1}\tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1}\boldsymbol{M} \\ \vdots \\ [\boldsymbol{W}_n^\top \quad \mathbf{0}]\{\tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n}\tilde{\boldsymbol{W}}_{u_n}\}^{-1}\tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n}\boldsymbol{M} \end{bmatrix} - \boldsymbol{M}
\end{aligned}
$$

Using (42) and (46), we obtain

$$
\begin{aligned}
I_1 &= \frac{1}{n}\boldsymbol{M}^\top (A^\top - I)\boldsymbol{P} \\
&= \frac{1}{n}\sum_{i=1}^{n}(M_i\{1+O_p(c_n)\} - M_i)\boldsymbol{P}_i = O_p(c_n),
\end{aligned}
$$

where $\boldsymbol{P}_i$ denotes the ith row of $\boldsymbol{P}$ and the last equality is due to the law of large numbers. For I_2, similarly as above, we compute

$$
A\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{W_1}^\top \left\{\mathsf{E}[WW^T|U_1]\right\}^{-1}\mathsf{E}[WZ^\top|U_1] \\ \vdots \\ \boldsymbol{W_n}^\top \left\{\mathsf{E}[WW^T|U_n]\right\}^{-1}\mathsf{E}[WZ^\top|U_n] \end{bmatrix}\{1+O_p(c_n)\}.
$$

Since $\beta - \hat{\beta} = O_p(n^{-1/2})$ by Lemma 3, we obtain

$$
\begin{aligned}
I_2 &= \frac{1}{n}(\beta-\hat{\beta})^\top \boldsymbol{Z}^\top A^\top \boldsymbol{P} \\
&= (\beta-\hat{\beta})^\top \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W_i}^\top \left\{\mathsf{E}[WW^T|U_i]\right\}^{-1} \\
&\qquad\qquad \cdot \mathsf{E}[WZ^\top|U_i]\{1+O_p(c_n)\}\boldsymbol{P}_i \\
&= O_p(n^{-1/2}),
\end{aligned}
$$

where, again, the last equality is due to the law of large numbers. Finally, for

$$
A\boldsymbol{N} = \begin{bmatrix} [\boldsymbol{W}_1^\top \quad \mathbf{0}]\{\tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1}\tilde{\boldsymbol{W}}_{u_1}\}^{-1}\tilde{\boldsymbol{W}}_{u_1}^\top \boldsymbol{K}_{u_1}\boldsymbol{N} \\ \vdots \\ [\boldsymbol{W}_n^\top \quad \mathbf{0}]\{\tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n}\tilde{\boldsymbol{W}}_{u_n}\}^{-1}\tilde{\boldsymbol{W}}_{u_n}^\top \boldsymbol{K}_{u_n}\boldsymbol{N} \end{bmatrix},
$$

the same argument for (45) leads to

$$
\tilde{\boldsymbol{W}}_{u_i}^\top \boldsymbol{K}_{u_i}\boldsymbol{N} = nf(U_i)\mathsf{E}[WN^\top|U_i]\{1+O_p(c_n)\},
$$

where $\mathsf{E}[WN^\top|U_i] = 0$ since N is independent of W and U_i, and N has a zero mean. Thus, by the law of large numbers,

$$
I_3 = \frac{1}{n}\boldsymbol{N}^\top A^\top \boldsymbol{P} = \frac{1}{n}\sum_{i=1}^{n} O_p(c_n)\boldsymbol{P}_i = O_p(c_n).
$$

□

Proofs of Theorem 3 and Corollaries 3 and 4

To reuse model (37) for the prediction of Y^τ, the main challenge comes from M that changes with U (while β remains invariant). First, we introduce some notation for the proofs. We define $(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{W}_i)$ and $(\boldsymbol{Z}, \boldsymbol{Z}_i, \boldsymbol{M}, \boldsymbol{N})$ in the same way as in Appendix K. Similarly, we define $\boldsymbol{Z}_V, \boldsymbol{M}_V, \boldsymbol{N}_V$, and all the corresponding data matrices for the test data (e.g., $\boldsymbol{Y}^\tau$), and let $\sigma^2 = \mathsf{E}[(N^\tau)^2]$. With this notation, the IMP (7) implies that

$$
M = \lambda M_V + \zeta^\top W := [W, M_V]w, \tag{47}
$$

for some $\zeta \in \mathbb{R}^p$. Let $\hat{\boldsymbol{W}}_V := [\boldsymbol{W}, \hat{\boldsymbol{M}}_V]$ and $\hat{\boldsymbol{W}}_V^\tau := [\boldsymbol{W}^\tau, \hat{\boldsymbol{M}}_V^\tau]$. Then, the OLS estimator of w according to the above equation is given by

$$\hat{w} = (\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{W}}_V)^{-1} \hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{M}}. \tag{48}$$

We predict $\boldsymbol{Y}^\tau$ using the continuous IMP estimator

$$\hat{\boldsymbol{Y}}^\tau = \tilde{\boldsymbol{W}}_V^\tau \hat{w} + \boldsymbol{Z}^\tau \hat{\beta}, \tag{49}$$

where $\hat{\beta}$, $\hat{\boldsymbol{M}}$, and $\hat{\boldsymbol{M}}_V$ are provided in Appendix K.

Lemma 5. *Under assumptions 1)$\sim$ 4), it holds that*

$$\hat{w} - w = O_p(c_n \vee n^{-1/2}).$$

Proof. Using the fact that

$$\hat{\boldsymbol{W}}_V = [\boldsymbol{W}, \boldsymbol{M}_V] + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V],$$

we obtain

$$\begin{aligned}
&\frac{1}{n}\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{W}}_V \\
&= \frac{1}{n}[\boldsymbol{W}, \boldsymbol{M}_V]^\top [\boldsymbol{W}, \boldsymbol{M}_V] + \frac{1}{n}[\boldsymbol{W}, \boldsymbol{M}_V]^\top [\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V] \\
&\quad + \frac{1}{n}[\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V]^\top [\boldsymbol{W}, \boldsymbol{M}_V] + \frac{1}{n}[\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V]^\top \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot [\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V] \\
&= \mathsf{E}\Big[(W^\top, M_V)^\top (W^\top, M_V)\Big]\{1 + O_p(c_n \vee n^{-1/2})\},
\end{aligned}$$

where we use R_1 and R_2 from Lemma 4 and the law of large numbers. Similarly,

$$\begin{aligned}
\frac{1}{n}\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{M}} &= \frac{1}{n}[\boldsymbol{W}, \boldsymbol{M}_V]^\top \boldsymbol{M} + \frac{1}{n}[\boldsymbol{W}, \boldsymbol{M}_V]^\top (\hat{\boldsymbol{M}} - \boldsymbol{M}) \\
&\quad + \frac{1}{n}[\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V]^\top (\hat{\boldsymbol{M}} - \boldsymbol{M}) \\
&\quad + \frac{1}{n}[\boldsymbol{0}, \hat{\boldsymbol{M}}_V - \boldsymbol{M}_V]^\top \boldsymbol{M} \\
&= \mathsf{E}\Big[(W^\top, M_V)^\top M\Big]\{1 + O_p(c_n \vee n^{-1/2})\}.
\end{aligned}$$

Note that the rate is not directly implied by Lemma 4, but it can be proved using the same techniques demonstrated in the proof Lemma 4. Therefore, using (47),

$$\hat{w} = (\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{W}}_V)^{-1} \hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{M}} = w\{1 + O_p(c_n \vee n^{-1/2})\}.$$

□

M.2 Proof of Theorem 3

Using (49) and $\boldsymbol{Y}^\tau = \boldsymbol{M}^\tau + \boldsymbol{Z}^\tau \beta + \boldsymbol{N}^\tau$, we derive

$$\begin{aligned}
&\hat{\boldsymbol{Y}}^{\boldsymbol{\tau}} - \boldsymbol{Y}^\tau \\
=&[\boldsymbol{W}^\tau, \hat{\boldsymbol{M}}_V^\tau]\hat{w} + \boldsymbol{Z}^\tau \hat{\beta} - [\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau] w - \boldsymbol{Z}^\tau \beta - \boldsymbol{N}^\tau \\
=&[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau](\hat{w} - w) + [\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau](\hat{w} - w) \\
&\qquad + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau] w + \boldsymbol{Z}^\tau (\hat{\beta} - \beta) - \boldsymbol{N}^\tau.
\end{aligned}$$

Then, the generalization error is given by

$$\frac{1}{m}(\hat{\boldsymbol{Y}}^{\boldsymbol{\tau}} - \boldsymbol{Y}^{\tau})^\top(\hat{\boldsymbol{Y}}^{\boldsymbol{\tau}} - \boldsymbol{Y}^{\tau}) = \sum_{i=1}^{10} J_i$$

where

$$\begin{aligned}
J_1 &= \frac{1}{m}(\hat{w} - w)^\top[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]^\top \\
&\qquad\qquad [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau](\hat{w} - w), \\
J_2 &= \frac{1}{m}(\hat{w} - w)^\top[\boldsymbol{X}^\tau, \boldsymbol{M}_V^\tau]^\top[\boldsymbol{X}^\tau, \boldsymbol{M}_V^\tau](\hat{w} - w), \\
J_3 &= \frac{1}{m}w^\top[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]^\top[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]w, \\
J_4 &= \frac{1}{m}(\hat{\beta} - \beta)^\top(\boldsymbol{Z}^\tau)^\top\boldsymbol{Z}^\tau(\hat{\beta} - \beta), \\
J_5 &= \frac{1}{m}(\boldsymbol{N}^\tau)^\top\boldsymbol{N}^\tau, \\
J_6 &= \frac{1}{m}(\hat{w} - w)^\top[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]^\top\Big\{[\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau] \\
&\quad \cdot (\hat{w} - w) + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]w + \boldsymbol{Z}^\tau(\hat{\beta} - \beta) - \boldsymbol{N}^\tau\Big\}, \\
J_7 &= \frac{1}{m}(\hat{w} - w)^\top[\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau]^\top\Big\{[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau] \\
&\quad \cdot (\hat{w} - w) + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]w + \boldsymbol{Z}^\tau(\hat{\beta} - \beta) - \boldsymbol{N}^\tau\Big\}, \\
J_8 &= \frac{1}{m}w^\top[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]^\top\Big\{[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau] \\
&\quad \cdot (\hat{w} - w) + [\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau](\hat{w} - w) + \boldsymbol{Z}^\tau(\hat{\beta} - \beta) - \boldsymbol{N}^\tau\Big\}, \\
J_9 &= \frac{1}{m}(\hat{\beta} - \beta)^\top(\boldsymbol{Z}^\tau)^\top\Big\{[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau](\hat{w} - w) \\
&\quad + [\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau](\hat{w} - w) + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]w + \boldsymbol{N}^\tau\Big\}, \\
J_{10} &= -\frac{1}{m}(\boldsymbol{N}^\tau)^\top\Big\{[\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau](\hat{w} - w) \\
&\quad + [\boldsymbol{W}^\tau, \boldsymbol{M}_V^\tau](\hat{w} - w) + [\boldsymbol{0}, \hat{\boldsymbol{M}}_V^\tau - \boldsymbol{M}_V^\tau]w + \boldsymbol{Z}^\tau(\hat{\beta} - \beta)\Big\}.
\end{aligned}$$

We can show the following rates of these terms through simple applications of Lemmas 3, 4, 5, along with the law of large number and the central limit theorem. $J_2, J_3 = O_p(c_n^2 \vee n^{-1})$, $J_4 = O_p(n^{-1})$, $J_5 = \sigma^2 + O_p(m^{-1/2})$,

$$\begin{aligned}
J_{10} &= O_p(c_m \vee m^{-1/2}) \cdot O_p(c_n \vee n^{-1/2}) \\
&\quad + O_p(c_n \vee n^{-1/2}) + O_p(c_m \vee m^{-1/2}) + O_p(n^{-1/2}),
\end{aligned}$$

and $J_1, J_6 \sim J_9$ have higher orders compared with either $O_p(c_n \vee n^{-1/2})$ or $O_p(c_m \vee m^{-1/2})$ in J_{10}. This completes the proof of Theorem 3.

M.3 Proof of Corollary 3

We use l as the shorthand notation for l_n in the proof. Since the estimation of β, $\boldsymbol{M}$, and w are based on the labeled data, we have $\hat{\beta} - \beta = O_p(l^{-1/2})$ in Lemma 3 and $R_1, R_2, R_3 = O_p(c_l \vee l^{-1/2})$

and $R_4 = O_p(c_l)$ in Lemma 4. In the proof of Lemma 5, the rate of $\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{W}}_V$ remains the same since the estimation of $\boldsymbol{M}_V$ only uses the unlabeled data but the rate of $\hat{\boldsymbol{W}}_V^\top \hat{\boldsymbol{M}}$ now depends on l. This observation implies $\hat{w} - w = O_p(c_l \vee l^{-1/2})$. Now, observe that $J_2 = O_p(c_l^2 \vee l^{-1})$, $J_3 = O_p(c_m \vee m^{-1/2})$, $J_4 = O_p(l^{-1})$, $J_5 = \sigma^2 + O_p(m^{-1/2})$, and $J_{10} = O_p(c_l \vee l^{-1/2}) + O_p(m^{-1/2})$. Since $\max(\frac{l}{n}, \frac{l}{m}) \to 0$ as $\min(n, m) \to \infty$, we get $\sum_{i=1}^{10} J_i = O_p(c_l \vee l^{-1/2})$.

M.4 Proof of Corollary 4

According to the definition of the discrete IMP estimator, all the coefficients are treated as varying coefficients (i.e., $\beta = \mathbf{0}$ and $\beta_V = \mathbf{0}$), and $\boldsymbol{M}$ is estimated by performing the OLS for each environment and then putting the estimates into one vector, thus $R_1 = O_p(a_n^2)$ and $R_2, R_4 = O_p(a_n)$ in Lemma 4, with $a_n = (\min_{e \in \mathcal{E}^{\text{train}}} n_e)^{-1/2}$ by the asymptotic normality of the OLS estimators (note that X and Z are assumed to have finite fourth moments in Lemma 3). Accordingly, we have $\hat{w} - \hat{w} = O_p(a_n)$ in Lemma 5 using the law of large numbers. Setting $Z^\tau = \mathbf{0}$ in $J_1 \sim J_{10}$, we obtain $J_2, J_3 = O_p(a_n^2)$, $J_4 = 0$, $J_5 = \sigma^2 + O_p(m^{-1/2})$, $J_{10} = O_p(a_n) + O_p(a_m)$, and the other terms have higher orders compared with $O_p(a_n)$ or $O_p(a_m)$. Similar to Theorem 3, the asymptotic generalization error is dominated by J_{10}.