Learning invariant representations with missing data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Spurious correlations allow deep models to predict well during training but poorly on related test populations. Recent work has shown that models that satisfy particular independencies involving the correlation-inducing *nuisance* variable have guarantees on their test performance. However, enforcing independencies requires nuisances to be observed during training. But nuisances such as demographics or image background labels are often missing. Enforcing independence on just the observed data does not imply independence on the entire population. In this work, we derive the missing-MMD estimator used for invariance objectives under missing nuisances. On simulations, missing-MMDs enable improvements in test performance similar to those achieved by using fully-observed data.

1 Introduction

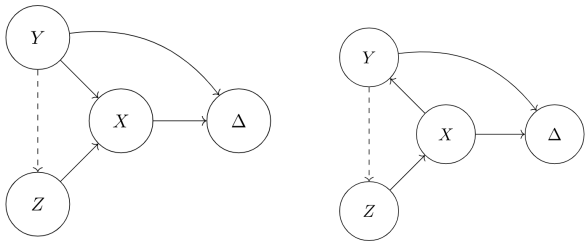
Spurious correlations, or *shortcuts*, allow models to predict well on populations during development, but have worse than chance performance on related populations at test time [4, 16, 24, 12, 7, 21]. For example, diabetes is associated with high body mass index (BMI) in the United States. But in India and Tawain, diabetes also co-occurs with low and average BMI [26]. Nuisance variables, such as BMI, due to their shifting relationship with the label, in this case diabetes, can lead a model to exploit correlations in training data but generalize poorly in test sets of interest.

Invariant prediction methods are designed to guarantee performance on test distributions of interest when training data exhibits nuisance correlations [15, 1, 12, 24, 16]. These methods enforce (conditional) independence between models and nuisance (given label). But they require the nuisance to be specified explicitly and observed. However, in large health datasets, nuisances are often missing. For example, not all people who report diabetes status may report other correlated conditions (e.g., hypertension, BMI, depression) or demographics (e.g., gender, ethnicity). In image classification, when a prediction should be based on an object (*is this a water bird or land bird?*) and not its background (e.g., water or trees) [21] we may only have background labels for a subset of images.

It is necessary to handle missingness appropriately to improve generalization at deployment. But extending invariant prediction methods to missingness is hard. The difficulty stems from adapting the objectives, usually measures of dependence such as Maximum Mean Discrepancys (MMDs) or Mutual Information (MI) that require sampling from the full data.

We propose missing-MMDs, a new objective that measures nuisance-model dependence under missingness. We first show the consequences of naively enforcing independence on only the nuisance-observed subset of data: independence on this subset does not imply independence on the full data, and solutions on the full data objective may be discarded by optimization on the observed subset. Next, we derive the missing-MMD and show that it is consistent when one can predict either the missingness variable or the nuisance, i.e. the estimators are *doubly-robust* [2]. Finally, in simulations we show that missing-MMDs improve test accuracy over using observed-only data and perform close to ground-truth estimation with no missingness.

Submitted to NeurIPS DistShift Workshop 2021. Do not distribute.



(a) Anti-Causal (b) Causal

Figure 1: Generative processes. The $Y \to Z$ is dashed to emphasize that the $Z|Y$ may change at test time. Δ determines missingness of Z and in general depends on both X and Y.

2 Notation and Background

Notation. Let X denote features such as demographics or physiological signals. Let Y be a label such as disease status. Let Z denote the nuisance, which could be another disease correlated with Y, demographics, or part of the data collection process such as artifacts in imaging data. Denote the nuisance missingness indicator as Δ. Instead of (X, Y, Z), we observe $(X, Y, \Delta, \tilde{Z} = \Delta Z)$, where $\tilde{Z} = Z$ when $\Delta = 1$ and Z is unobserved otherwise. Let $h_X = h(X)$ denote a model to predict Y.

Assumptions and scope. We require ignorability $Z \perp\!\!\!\perp \Delta | X, Y$, (similar to "no unobserved confounders" [8]). Some distributions that satisfy this are shown in Figure 1. We require positivity $0 < \epsilon \leq P(\Delta = 1|X)$ to ensure that we observe Z for any X with positive probability. We focus on the graph in Figure 1(a) with binary nuisances. The presented method can extend to continuous nuisances and to other graphs (e.g. Figure 1(b)). We pick one measure of dependence, the MMD.

OOD generalization. Nuisance-based prediction arises in training data when Z is predictive of Y and associated with X, causing models to use information about Z in X to predict Y. This is a problem when the test set is expected to have a different (Y, Z) relationship from the training set.

Formally, define a family of distributions where the only differences are in $P(Z \mid Y)$: indexing by D, let the family be $\mathcal{F} = \{P_D(X, Y, Z) = P(Y)P_D(Z|Y)P(X|Y, Z)\}$. We assume that we see train on $P_{train} \in \mathcal{F}$ and may see any other member at test time. When $P_{test}(Z|Y) \neq P_{train}(Z|Y)$, in general $P_{train}(Y \mid X) \neq P_{test}(Y|X)$ and a model built on P_{train} can generalize poorly. Let $\overline{Y} = 1 - Y$ and define P_{train} by: $Y \sim \mathcal{B}(0.5)$, $P_{train}(Z = 1 \mid Y = 1) = P_{train}(Z = 0 \mid Y = 0) = 0.9Y + 0.1\overline{Y}$ and $X \sim [\mathcal{N}(Y, 10), \mathcal{N}(Z, 1)]$. The MLE solution to predicting Y from X on P_{train} is a linear classifier with a higher weight on X_2 because of the smaller variance. However, if $\forall a \in \{0, 1\}$, $P_{test}(Z = a \mid Y = a) = 0.9\overline{Y} + 0.1Y$, then the second coordinate relationship with the label is flipped: higher values correspond to label $Y = 1$, unlike in P_{train}.

When it is possible to anticipate and observe nuisances during training, enforcing certain independence constraints [12, 24, 16] helps guarantee performance regardless of the nuisance-label relationship. For example, [12, 16] construct a distribution with an independent nuisance and label and, under this distribution, enforce different independence constraints on the representation: while [12] enforce $h_X \perp\!\!\!\perp Z$, [16] enforce $(Y, h_X) \perp\!\!\!\perp Z$. Both give bounds on the performance of the resulting models on test distributions with arbitrary $p_D(Z \mid Y)$. [24] enforce a different independence $h_X \perp\!\!\!\perp Z \mid Y$ directly on the training data and show that this produces invariant risk regardless of the underlying nuisance-label correlation. We work with [24] because their method works on a generalize of the model in [12] and enforces independence using MMD, which can be computed in closed form, unlike the MI in [16]. We summarize the argument for using $h_X \perp\!\!\!\perp Z \mid Y$ in the anti-causal graph in Appendix D. In this work we study making the independence objectives work when the nuisance is subject to missingness so that they can be used for such guarantees.

Measures of dependence. To enforce independencies it is necessary to measure dependence, i.w. measuring distance between a joint and its product of marginals. Integral Probability Metrics (IPMs) define a metric on distributions and a special case, the kernel-based MMD [6], has a closed form. Let $X_1 \sim P, X_2 \sim Q$. Let X_j' be an independent copy of X_j. For kernel k,

$$\text{MMD}(X_1, X_2) = \mathbb{E}[k(X_1, X_1')] + \mathbb{E}[k(X_2, X_2')] - 2\,\mathbb{E}[k(X_1, X_2)]\,. \tag{1}$$

MMDs can measure dependence between X, Z by computing $\text{MMD}(p(X, Z), p(X)p(Z))$.

Missingness. Two parts of the data distribution help estimate $\mathbb{E}[Z]$: $G_X \triangleq \mathbb{E}[\Delta \mid X]$ and $m_X \triangleq \mathbb{E}[Z \mid X]$ (Appendix F). The *doubly-robust* (DR) estimator combines both and is consistent when *either* G_X or m_X are [17, 2, 10]:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{\Delta \tilde{Z}}{G_X} - \frac{\Delta - G_X}{G_X} m_X\right]. \tag{2}$$

More on DR estimation in Appendix G. This provides extra assurance when neither G_X nor m_X is known. In this work we adapt this estimator from $\mathbb{E}[Z]$ to means $\mathbb{E}[f(X, Z)]$ and $\mathbb{E}_{X|Z=z}[f(X)]$.

3 Invariant representations with missing data

For the anti-causal model, [24] enforce $h_X \perp\!\!\!\perp Z | Y = y$ for $y \in \{0, 1\}$ by minimizing the MMD,

$$\max_h \quad \log p(y|h_X) - \lambda \cdot \text{MMD}\left(p(h_X|Z=1, Y)\,,\, p(h_X|Z=0, Y)\right)\,. \tag{3}$$

We first demonstrate what goes wrong when enforcing this only on samples with Z observed and then derive an estimator of MMD that enforces independence on the full population under missingness.

3.1 Failures on observed-only data

Restricting to $\Delta = 1$ amounts to sampling from $X, Z|Y, \Delta = 1$ instead of $X, Z|Y$. For each $y \in \{0, 1\}$, this enforces $h_X \perp\!\!\!\perp Z | Y = y, \Delta = 1$ instead of $h_X \perp\!\!\!\perp Z | Y = y$.

Proposition 1. There exist distributions on (X, Y, Δ, Z) such that there exist $h_X^\star$ that satisfy $h_X^\star \perp\!\!\!\perp Z | Y = y$ but not $h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1$.

The proof is in Appendix E. Proposition 1 implies that restricting the data would wrongly discard a solution. This is particularly bad when $h_X^\star$ is the optimal solution to Equation (3).

3.2 The missing-MMD

We present estimators for $\text{MMD}(p(h_X|Z=1), p(h_X|Z=0))$ One can compute the conditional MMD to enforce $h_X \perp\!\!\!\perp Z | Y = y$ simply by restricting samples to those with $Y = y$. For a kernel k,

$$\text{MMD}\left[p(h_X|Z=1), p(h_X|Z=0)\right] = \mathop{\mathbb{E}}_{\substack{X|Z=1 \\ X'|Z'=1}}\left[k_{XX'}\right] + \mathop{\mathbb{E}}_{\substack{X|Z=0 \\ X'|Z'=0}}\left[k_{XX'}\right] - 2 \mathop{\mathbb{E}}_{\substack{X|Z=1 \\ X'|Z'=0}}\left[k_{XX'}\right], \tag{4}$$

where $k_{XX'} = k(h_X, h_{X'})$. The challenge comes from missingness in the conditioning set. We derive G_X-based and m_X-based estimators in in Appendix H. In Proposition 2, we combine them.

Proposition 2. Assume positivity, ignorability, and, for each X, either $G_X = \mathbb{E}[\Delta|X]$ or $m_X = \mathbb{E}[Z|X]$. Let $\tilde{Z}_1 = \tilde{Z}$, $\tilde{Z}_0 = 1 - \tilde{Z}$, $m_{X,1} = m_X$ and $m_{X,0} = 1 - m_X$, for $j \in \{0, 1\}$. Let $k_{X,X'} = k(h_X, h_{X'})$, $G_{XX'} = G_X G_{X'}$, and $N(z, z') = P(Z = z)P(Z' = z')$. Then:

$$\mathop{\mathbb{E}}_{\substack{X|Z=j \\ X'|Z'=j'}}\left[k_{XX'}\right] = \frac{1}{N(j, j')} \mathop{\mathbb{E}}_{\substack{X,\Delta,Z \\ X',\Delta',Z'}}\left[\left(\frac{\Delta\Delta' \tilde{Z}_j \tilde{Z}'_{j'}}{G_{XX'}} - \frac{\Delta\Delta' - G_{XX'}}{G_{XX'}} m_{X,j} m_{X',j'}\right) k_{XX'}\right]. \tag{5}$$

Proof in Appendix H. We can use estimate the terms in Equation (4) and plug them into the objective in Equation (3). This depends on constant $\mathbb{E}[Z]$ in $N(j, j')$, which can be estimated with Equation (2). Altogether, Equation (5) is a ratio of two expectations. Each can be consistently estimated with Monte-Carlo by WLLN and the ratio is consistent by Slutsky's Theorem. We discuss recent work on noisy groups labels in Appendix C.

4 Experiments

We investigate the correspondence between accuracy, dependence, and missingness. We use:

- standard prediction of $Y|X$ with no invariance objective (NONE) (should be worst)

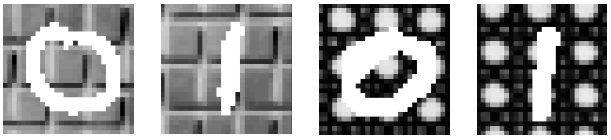


Figure 2: Textured MNIST with digits 0,1 on two textures from the Brodatz dataset.

- invariant prediction using the full data as if Z were fully observed (FULL) (should be best)
- invariant prediction using only the data with Z observed (OBS)
- the DR estimator with learned auxiliary distributions (DR) and true G_X (DR+)

Each table shows mean $\pm$ standard deviation over 3 seeds. Each seed draws a new dataset.

Experiment 1: Simulation. The training and validation sets are drawn as follows:

$$Y \sim \mathcal{B}(0.5),\ Z \sim \mathcal{B}(.9Y + .1\overline{Y}),\ X \sim [\mathcal{N}(Y - Z, \sigma_X^2), \mathcal{N}(Y + Z, \sigma_X^2)]\,. \tag{6}$$

The test set is changed to have the opposite nuisance-label relationship: $Z \sim \mathcal{B}(.1Y + .9\overline{Y})$. We draw Δ such that the optimal $h_X^\star = (X_1 + X_2)/2$ does not satisfy $h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1$ (Appendix E):

$$A = (X_1 + X_2)/2, \quad B = -(X_1 - X_2)/2, \quad Q = \mathbb{1}\left[A > 0.6\right]\mathbb{1}\left[B < 0.6\right]\,.$$

Then $\Delta \sim \mathcal{B}(Q + 0.2\overline{Q})$. The results are in Table 1. The DR estimators including the one with the learned G_X, m_X achieves indistinguishable performance to the full-data MMD.

Table 1: Simulation. $\lambda = 1$. Simulation results show the basic trend described. DR estimators indistinguishable from full data estimator.

	NONE	OBS	FULL	DR	DR+
TR MMD	0.21 ± 0.04	0.05 ± 0.04	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01
TR ACC	0.89 ± 0.00	0.87 ± 0.00	0.86 ± 0.01	0.85 ± 0.01	0.84 ± 0.02
TE ACC	0.67 ± 0.02	0.77 ± 0.02	0.80 ± 0.01	0.81 ± 0.02	0.81 ± 0.01

Experiment 2: Textured MNIST. Following [5], we correlate MNIST digits with textures from the Brodatz dataset (Figure 2). This captures the idea that MMDs can be used to ignore image backgrounds, even when background labels are not all available. The missingness distribution is chosen to be a proxy for Z by measuring the average pixel intensity of X. More details in Appendix B and results in Table 2. The DR estimators including the one with a learned G_X, m_X achieve close to the full performance.

Table 2: MNIST $\lambda = 1$. DR estimator achieves close to full performance as measured by 0 full MMD and high test accuracy. OBS is notably high variance.

	NONE	OBS	FULL	DR	DR+
TR MMD	2.05 ± 0.18	0.02 ± 0.04	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01
TR ACC	0.90 ± 0.01	0.74 ± 0.03	0.76 ± 0.01	0.77 ± 0.0	0.76 ± 0.01
TE ACC	0.13 ± 0.01	0.63 ± 0.17	0.74 ± 0.01	0.72 ± 0.04	0.73 ± 0.01

5 Conclusion

The estimator we present can be used to optimize invariance objectives under nuisance missingness and, unlike using observed data only, the presented estimator equals the full data objective when either auxiliary model can be learned. Estimation of the full data objective is necessary to preserve the theoretical properties of invariance objectives, especially to make sure that the model does not suffer from nuisance-induced prediction just on sub-populations whose nuisances tend to be missing.

References

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

[2] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[3] D. A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 279–292, 1983.

[4] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[5] I. J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013. URL http://arxiv.org/abs/1308.4214.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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

[8] M. A. Hernan and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 1st edition, 2021.

[9] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[10] J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

[11] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[12] M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally-motivated shortcut removal using auxiliary labels. *arXiv preprint arXiv:2105.06422*, 2021.

[13] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, volume 29, pages 2208–2216, 2016.

[14] J. Pearl. *Causality*. Cambridge university press, 2009.

[15] J. Peters, P. Bühlmann, and N. 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.

[16] A. Puli, L. H. Zhang, E. K. Oermann, and R. Ranganath. Predictive modeling in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520*, 2021.

[17] J. Robins and A. G. Rotnitzky. Comment on the bickel and kwon article, 'inference for semiparametric models: Some questions and an answer'. *Statistica Sinica*, 11:920–936, 01 2001.

[18] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

[19] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[20] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[21] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[22] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[23] M. J. Van der Laan, M. Laan, and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

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

[25] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.

[26] E. C. WHO. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet (London, England)*, 363(9403):157–163, 2004.

A Full Experiments

Table 3: Simulation. $\lambda = 1$.

	NONE	OBS	FULL	DR	DR+	REG	IPCW	IPCW+
TR MMD	0.21 ± 0.04	0.05 ± 0.04	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.01
TR ACC	0.89 ± 0.00	0.87 ± 0.00	0.86 ± 0.01	0.85 ± 0.01	0.84 ± 0.02	0.86 ± 0.00	0.84 ± 0.01	0.84 ± 0.01
TE ACC	0.67 ± 0.02	0.77 ± 0.02	0.80 ± 0.01	0.81 ± 0.02	0.81 ± 0.01	0.79 ± 0.02	0.82 ± 0.02	0.81 ± 0.00

Table 4: Simulation. $\lambda = 5$.

	NONE	OBS	FULL	DR	DR+	REG	IPCW	IPCW+
TR MMD	0.21 ± 0.04	0.03 ± 0.02	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.0	0.00 ± 0.01	0.00 ± 0.0	0.00 ± 0.0
TR ACC	0.89 ± 0.0	0.85 ± 0.02	0.84 ± 0.01	0.82 ± 0.01	0.78 ± 0.06	0.84 ± 0.00	0.81 ± 0.02	0.81 ± 0.03
TE ACC	0.67 ± 0.02	0.78 ± 0.02	0.83 ± 0.01	0.82 ± 0.02	0.77 ± 0.04	0.82 ± 0.01	0.81 ± 0.02	0.80 ± 0.01

Table 5: MNIST $\lambda = 1$.

	NONE	OBS	FULL	DR	DR+	REG	IPCW	IPCW+
TR MMD	2.05 ± 0.18	0.02 ± 0.04	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	$0.07 \pm 0.12.$	0.03 ± 0.06
TR ACC	0.90 ± 0.01	0.74 ± 0.03	0.76 ± 0.01	0.77 ± 0.00	0.76 ± 0.01	0.76 ± 0.01	0.67 ± 0.16	0.68 ± 0.15
TE ACC	0.13 ± 0.01	0.63 ± 0.17	0.74 ± 0.01	0.72 ± 0.04	0.73 ± 0.01	0.73 ± 0.01	0.64 ± 0.14	0.61 ± 0.11

Table 6: MNIST $\lambda = 5$.

	NONE	OBS	FULL	DR	DR+	REG	IPCW	IPCW+
TR MMD	2.05 ± 0.18	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01	0.00 ± 0.01	0.01 ± 0.01	0.01 ± 0.02
TR ACC	0.9 ± 0.01	0.66 ± 0.15	0.75 ± 0.01	0.65 ± 0.14	0.65 ± 0.13	0.75 ± 0.01	0.71 ± 0.08	0.60 ± 0.12
TE ACC	0.13 ± 0.01	0.65 ± 0.15	0.75 ± 0.01	0.73 ± 0.02	0.70 ± 0.09	0.75 ± 0.01	0.55 ± 0.3	0.60 ± 0.12

B Textured mnist details

We adapt the code from [5] https://github.com/lisa-lab/pylearn2/blob/master/pylearn2/scripts/datasets/make_mnistplus.py.

We correlate MNIST digits with textures from the Brodatz dataset (Figure 2).

We follow a similar setup to colored MNIST [1]: because $Y|X$ is essentially determinstic, even strong spurious correlations may be ignored by a model on MNIST. To get $Y|X$ closer to what may be expected in real data, we flip the label with 25% chance. Letting X only predict Y with 75% chance means a strong spurious correlation with better predictive power can be used (a color or in this case a background).

The missingness distribution is chosen to be a proxy for Z by measuring the average pixel intensity of X. Let μ_X be the mean pixel value of a 28x28 MNIST image. Then $Q = \mathbb{1}[Y = 1]\,\mathbb{1}[\mu_X < 0.3]$ and $\Delta \sim \mathcal{B}(Q + .2\overline{Q})$. The choice of $\mathbb{1}[\mu_X < 0.3]$ is because this is a proxy for the value of Z (nuisance determines whether the image is light or dark grey).

C Related Work

Motivated by fairness, [25] study a related problem of optimizing invariance-inducing objectives when the protected group label (analogous to our nuisance variables) are noisy. Given bounds on the level of label noise, this work proposes a optimizing an objective based on the distributionally robust optimization framework [13]. Additionally, if given a small amount of true labels the authors suggest fitting a model to de-noise the noisy group labels and re-weight examples in the objective, which is

similar in spirit to our work. In our approach, however, we exploit structural assumptions about the missingness process to build a doubly-robust estimator of the MMD penalty used during optimization.

At the time of this submission, there is an OpenReview entry for an ICLR submission titled "Invariant Learning With Partial Group Labels" (https://openreview.net/pdf?id=sWbXSWzHPa). In this work the authors maximize over unknown group assignments to get worst-case versions of invariance objectives. As opposed to the DR estimator presented here, in that work the objective still depends on the missingness distribution.

D Invariant Predictor

Satisfying independence $h_X \perp\!\!\!\perp Z \mid Y$ means $P_{train}(h_X|Y,Z) = P_{train}(h_X|Y)$ and the graphical assumptions on the anti-causal family mean $P_{train}(h_X|Y,Z) = P(h_X|Y,Z)$ in any member of the family. Combined this means $P_{train}(h_X|Y,Z) = P(h_X|Y)$ in any member of the family when the model $P_{train}(Y|h_X)$ satisfies $h_X \perp\!\!\!\perp Z|Y$. Consider test set performance $\mathbb{E}_{P_{test}(Y,X)}[\log P_{train}(Y|h_X)]$. By the assumption on the family, by Bayes, and by satisfying the independence constraint:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{P_{test}(Y,X)}[\log P_{train}(Y|h_X)] &= \mathop{\mathbb{E}}_{P_{test}(Y,X)}\left[\log \frac{P_{train}(h_X|Y)P(Y)}{P_{train}(h_X)}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(Y,X,Z)}\left[\log \frac{P_{train}(h_X|Y,Z)P(Y)}{\mathbb{E}_{P(Y)}\left[P_{train}(h_X|Y,Z)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(Y,h_X,Z)}\left[\log \frac{P_{train}(h_X|Y,Z)P(Y)}{\mathbb{E}_{P(Y)}\left[P_{train}(h_X|Y,Z)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(Y,h_X,Z)}\left[\log \frac{P(h_X|Y,Z)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y,Z)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(Y,h_X,Z)}\left[\log \frac{P(h_X|Y)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(Y,h_X)}\left[\log \frac{P(h_X|Y)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P_{test}(h_X|Y)P_{test}(Y)}\left[\log \frac{P(h_X|Y)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P(h_X|Y)P(Y)}\left[\log \frac{P(h_X|Y)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y)\right]}\right] \\
&= \mathop{\mathbb{E}}_{P(h_X,Y)}\left[\log \frac{P(h_X|Y)P(Y)}{\mathbb{E}_{P(Y)}\left[P(h_X|Y)\right]}\right]
\end{aligned}
$$

The last quantity does not depend on any specific $P_D(Z|Y)$ so performance of the is the same on all P_{test} in $\mathcal{F}$.

E Restricting to Observed Data is Bad

Proposition. There exists $h_X^\star$ that satisfies $h_X^\star \perp\!\!\!\perp Z|Y = y$ but not $h_X^\star \perp\!\!\!\perp Z|Y = y, \Delta = 1$

This means restricting the data would wrongly discard a solution. It suffices to illustrate this even when Z, Y are not correlated. Consider

$$Y \sim \text{Bern}(0.5), \quad Z \sim \mathcal{N}(0, \sigma_Z^2) \quad X = [\mathcal{N}(Y - Z, \sigma_X^2), Y + Z]$$

Let $h_X^\star = (X_1 + X_2)/2$ which has the distribution $\mathcal{N}(Y, \sigma_X^2/4)$. Let $\hat{Z} = -1.0(X_1 - X_2)/2$ which has the distribution $\mathcal{N}(Z, \sigma_X^2/4)$.

In this process $\hat{Z}$ is correlated with Z even when conditioning on Y due to the information from Z in X. In particular the definition of $\hat{Z}$ removes the information from Y.

Dropping the constants, we therefore need a distribution $\Delta|X, Y$ such that $X_1 - X_2$ is dependent with $X_1 + X_2$ given $Y, \Delta = 1$. This happens when, within each $Y = y$, the distribution $\Delta|X, Y = y$ computes functions that mix $X_1 - X_2$ and $X_1 + X_2$.

For example, let $A = (X_1 + X_2)/2$ and $B = -(X_1 - X_2)/2$. Let $Q = \mathbb{1}[A < 0.5]\,\mathbb{1}[B > 0.5]$ and let $\Delta \sim \mathcal{B}(Q + .2(1 - Q))$. Then h_X is not independent of Z given $Y, \Delta = 1$.

F IP and Outcome estimators

We review estimation of $\mathbb{E}[Z]$ under missingness. Two pieces of the data generation process can help, the missingness process G_X and the conditional expectation m_X of the missing variable:

$$G_X \triangleq \mathbb{E}[\Delta \mid X], \qquad m_X \triangleq \mathbb{E}[Z \mid X]$$

Inverse-weighting estimators use G_X [9, 3, 18, 23, 8]

$$\begin{aligned}
\mathbb{E}[Z] &= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X}[Z] \\
&= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X}\left[\frac{\mathbb{E}[\Delta|X]}{\mathbb{E}[\Delta|X]} Z\right] \\
&= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X} \mathop{\mathbb{E}}_{\Delta|X}\left[\frac{\Delta Z}{\mathbb{E}[\Delta|X]}\right] \\
&= \mathop{\mathbb{E}}_{XZ\Delta}\left[\frac{\Delta Z}{\mathbb{E}[\Delta|X]}\right] \\
&= \mathop{\mathbb{E}}_{X\Delta Z}\left[\frac{\Delta Z}{G_X}\right] \\
&= \mathop{\mathbb{E}}_{X\Delta Z}\left[\frac{\Delta \tilde{Z}}{G_X}\right]
\end{aligned} \tag{7}$$

This means we can estimate $\mathbb{E}[Z]$ provided that (1) ignorability and positivity hold and (2) G_X is known. G_X can be estimated by regressing Δ on X. Alternatively, standardization estimators use m_X [19, 22, 20, 14, 11, 8]:

$$\mathbb{E}[Z] = \mathop{\mathbb{E}}_{X}\Big[\mathbb{E}[Z|X]\Big] = \mathop{\mathbb{E}}_{X}\Big[\mathbb{E}[Z|X, \Delta = 1]\Big] = \mathop{\mathbb{E}}_{X}[m_X] \tag{8}$$

m_X can be estimated by regressing $\tilde{Z}$ on X where $\Delta = 1$ (by ignorability).

G DR estimator of Mean of Z

The inverse weighting and regression estimators can be combined. Equation (2) defines the DR estimator of $\mathbb{E}[Z]$ by

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{\Delta \tilde{Z}}{G_X} - \frac{\Delta - G_X}{G_X} m_X\right]$$

Let us re-write this expectation until we see it equals $\mathbb{E}[Z]$ when G or m are correct.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\Delta \tilde{Z}}{G_X} - \frac{\Delta - G_X}{G_X} m_X\right] \\
&= \mathbb{E}\left[\frac{\Delta Z}{G_X} - \frac{\Delta - G_X}{G_X} m_X\right] \\
&= \mathbb{E}\left[Z + \frac{\Delta Z}{G_X} - Z - \frac{\Delta - G_X}{G_X} m_X\right] \\
&= \mathbb{E}\left[Z + \frac{\Delta Z}{G_X} - \frac{G_X}{G_X} Z - \frac{\Delta - G_X}{G_X} m_X\right] \\
&= \mathbb{E}\left[Z + \frac{\Delta - G_X}{G_X} Z - \frac{\Delta - G_X}{G_X} m_X\right] \\
&= \mathbb{E}\left[Z + \frac{\Delta - G_X}{G_X} (Z - m_X)\right] \\
&= \mathbb{E}\left[Z\right] + \mathbb{E}\left[\frac{\Delta - G_X}{G_X} (Z - m_X)\right]
\end{aligned}
$$

The first term is what we want, so we just have to check if the second term is 0 when either G or m are correct. If G is correct (regardless of m) then:

$$
\begin{aligned}
\mathbb{E}\left[\frac{\Delta - G_X}{G_X} (Z - m_X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\Delta - G_X}{G_X} (Z - m_X) \middle| X, Z\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[\Delta | X, Z] - G_X}{G_X} (Z - m_X)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[\Delta | X] - G_X}{G_X} (Z - m_X)\right] \\
&= \mathbb{E}\left[\frac{G_X - G_X}{G_X} (Z - m_X)\right] = 0
\end{aligned}
$$

When m is correct (regardless of G):

$$
\begin{aligned}
\mathbb{E}\left[\frac{\Delta - G_X}{G_X} (Z - m_X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\Delta - G_X}{G_X} (Z - m_X) \middle| X, \Delta\right]\right] \\
&= \mathbb{E}\left[\frac{\Delta - G_X}{G_X} (\mathbb{E}[Z | X, \Delta] - m_X) \middle| X, \Delta\right] \\
&= \mathbb{E}\left[\frac{\Delta - G_X}{G_X} (\mathbb{E}[Z | X, \Delta] - \mathbb{E}[Z | X, \Delta = 1]) \middle| X, \Delta\right] \\
&= \mathbb{E}\left[\frac{\Delta - G_X}{G_X} (\mathbb{E}[Z | X] - \mathbb{E}[Z | X]) \middle| X, \Delta\right] = 0
\end{aligned}
$$

H MMD estimators under missingness

H.1 inverse-weighted estimator

Here we start at the target and derive the estimator.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{P(X|Z=1)P(X'|Z'=1)}\Big[k_{XX'}\Big] \\
&= \int_{X,X'} kP(X|Z=1)P(X'|Z'=1)dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1,X)P(Z'=1,X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1|X)P(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} k\,\mathbb{E}(Z=1|X)\,\mathbb{E}(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[k\cdot Z\cdot Z'\Big] \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[\frac{\mathbb{E}[\Delta|X]\,\mathbb{E}[\Delta'|X']}{\mathbb{E}[\Delta|X]\,\mathbb{E}[\Delta'|X']}k\cdot Z\cdot Z'\Big] \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'}{G_X G_{X'}}k\cdot Z\cdot Z'\Big]
\end{aligned}
$$

H.2 standardization estimator

Here we start at the target and derive the estimator.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{P(X|Z=1)P(X'|Z'=1)}\Big[k_{XX'}\Big] \\
&= \int_{X,X'} kP(X|Z=1)P(X'|Z'=1)dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1,X)P(Z'=1,X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1|X)P(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} k\,\mathbb{E}(Z=1|X)\,\mathbb{E}(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{X,X'}\Big[m_X\cdot m_{X'}\cdot k\Big]
\end{aligned}
$$

H.3 DR estimator

Let us prove the equality for $\mathbb{E}_{X|Z=1,X'|Z'=1}[k_{XX'}]$. The proof is analogous for the $(Z=0,Z'=0)$ and $(Z=1,Z'=0)$ cases. Here we start at the estimator and derive the target.

$$
\begin{aligned}
&\frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\frac{1}{N(N-1)}\sum_{i\neq j}\Big[\frac{\Delta_{ij}\tilde{Z}_{ij}}{G_{ij}}k_{ij}-\frac{\Delta_{ij}-G_{ij}}{G_{ij}}m_{ij}k_{ij}\Big]\\
&\approx \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'\tilde{Z}\tilde{Z}'}{G_XG_{X'}}k-\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}m_Xm_{X'}k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'ZZ'}{G_XG_{X'}}k-\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}m_Xm_{X'}k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k+\frac{\Delta\Delta'}{G_XG_{X'}}ZZ'k-ZZ'k-\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}m_Xm_{X'}k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k+\frac{\Delta\Delta'}{G_XG_{X'}}ZZ'k-\frac{G_XG_{X'}}{G_XG_{X'}}ZZ'k-\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}m_Xm_{X'}k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k+\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}ZZ'k-\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}m_Xm_{X'}k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k+\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k\Big]+\frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]
\end{aligned}
$$

Our estimator equals two terms. We first show that the first term equals the desired quantity, and then show the second term equals 0 when either auxiliary model is correct.

$$
\begin{aligned}
\frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[ZZ'k\Big] &= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{X,X'}\Big[k\,\mathbb{E}[Z,Z'|X,X']\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\mathop{\mathbb{E}}_{X,X'}\Big[kP(Z=1,Z'=1|X,X')\Big]\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'}kP(Z=1,Z'=1|X,X')P(X,X')dXdX'\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'}kP(Z=1|X)P(Z'=1|X)P(X)P(X')dXdX'\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'}kP(Z=1,X)P(Z'=1,X')dXdX'\\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'}kP(Z=1,X)P(Z'=1,X')dXdX'\\
&= \int_{X,X'}kP(X|Z=1)P(X'|Z'=1)dXdX'\\
&= \mathop{\mathbb{E}}_{P(X|Z=1)P(X'|Z'=1)}\Big[k\Big]
\end{aligned}
$$

That's the expectation we want missing just the $P(Z=1)$ constants, so now we should show the next term is 0 when either m or G are correct. When G correct:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big] &= \mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[\frac{\mathbb{E}[\Delta\Delta'|X,X',Y,Z']-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[\frac{\mathbb{E}[\Delta\Delta'|X,X']-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[\frac{\mathbb{E}[\Delta|X]\,\mathbb{E}[\Delta'|X']-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,Z\\X',Z'}}\Big[\frac{G_XG_{X'}-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]=0
\end{aligned}$$

Likewise, when m correct:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{X,\Delta,Z\\X',\Delta',Z'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(ZZ'-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,\Delta\\X',\Delta'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(\mathbb{E}[ZZ'|X,X',\Delta,\Delta']-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,\Delta\\X',\Delta'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(\mathbb{E}[Z|X,\Delta]\,\mathbb{E}[Z'|X',\Delta']-m_Xm_{X'}\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,\Delta\\X',\Delta'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(\mathbb{E}[Z|X,\Delta]\,\mathbb{E}[Z'|X',\Delta']-\mathbb{E}[Z|X,\Delta=1]\,\mathbb{E}[Z'|X',\Delta'=1]\Big)k\Big]\\
&= \mathop{\mathbb{E}}_{\substack{X,\Delta\\X',\Delta'}}\Big[\frac{\Delta\Delta'-G_XG_{X'}}{G_XG_{X'}}\Big(\mathbb{E}[Z|X]\,\mathbb{E}[Z'|X']-\mathbb{E}[Z|X]\,\mathbb{E}[Z'|X']\Big)k\Big]=0
\end{aligned}$$

The proof for the other two terms is analogous but with using $\overline{Z}=(1-Z)$ instead of Z and $\overline{m}=1-m$ when conditioning on $Z=0$.

I kernel mmd between joint and product of marginals

The above formulation of mmd between $h_X|Z=1$ and $h_X|Z=0$ relied on optimizing with respect to h only: $P(Z)$ is constant in the optimization so the distance between conditionals specifies the distance between the product of marginals and joint and thus the dependence. However, considering the more general case of MMD between $P(h_X,Z)$ and $P(h_X)P(Z)$ has the advantage that is not necessary to consider a finite set of conditioning values for Z. That means the MMD can be extended to continuous nuisance Z. The more general formulation is:

$$\begin{aligned}
&\mathbb{E}_{\substack{(X,Z)\sim P(X,Z)\\(X',Z')\sim P(X,Z)}}\Big[k\Big(X::Z,X'::Z'\Big)\Big]+\mathbb{E}_{\substack{(X,Z)\sim P(X)P(Z)\\(X',Z')\sim P(X')P(Z')}}\Big[k\Big(X::Z,X'::Z'\Big)\Big]\\
&\qquad-2\mathbb{E}_{\substack{(X,Z)\sim P(X,Z)\\(X',Z')\sim P(X')P(Z')}}\Big[k\Big(X::Z,X'::Z'\Big)\Big]
\end{aligned}$$

This leads to the following estimator:

$$\mathbb{E}_{\substack{P(X,Z)\\P(X',Z')}}\Big[k\Big(X::Z,X'::Z'\Big)\Big]=\mathbb{E}\Big[\frac{\Delta\Delta'k(X::Z,X'::Z')}{G_{XX'}}-\frac{\Delta\Delta'-G_{XX'}}{G_{XX'}}\,\mathbb{E}[k|X,X']\Big]$$

and

$$
\begin{aligned}
&\mathbb{E}_{\substack{P(X)P(Z)\\P(X')P(Z')}}\left[k\Big(X{::}Z, X'{::}Z'\Big)\right] \\
&\qquad = \mathbb{E}_{\substack{P(X_1)P(X_2,Z_2)\\P(X_3)P(X_4,Z_4)}}\left[k\Big(X_1{::}Z_2, X_3{::}Z_4\Big)\right] \\
&\qquad = \mathbb{E}\left[\frac{\Delta\Delta' k(X_1{::}Z_2, X_3{::}Z_4)}{G_{X_1X_3}} - \frac{\Delta\Delta' - G_{X_1X_3}}{G_{X_1X_3}}\,\mathbb{E}[k(X_1{::}Z_2, X_3{::}Z_4)|X_1, X_3]\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_{\substack{P(X,Z)\\P(X')P(Z')}}\left[k\Big(X{::}Z, X'{::}Z'\Big)\right] \\
&\qquad = \mathbb{E}_{\substack{P(X_1,Z_1)\\P(X_2)P(X_3,Z_3)}}\left[k\Big(X_1{::}Z_1, X_2{::}Z_3\Big)\right] \\
&\qquad = \mathbb{E}\left[\frac{\Delta\Delta' k(X_1{::}Z_1, X_2{::}Z_3)}{G_{X_1X_3}} - \frac{\Delta\Delta' - G_{X_1X_3}}{G_{X_1X_3}}\,\mathbb{E}[k(X_1{::}Z_1, X_2{::}Z_3)|X_1, X_3]\right]
\end{aligned}
$$

The challenging part of applying this estimator is that now instead of one function m_X we have three functions, each of which estimates the mean of k under a different sampling distributions. Moreover, these conditional expectations depend on the current representation h_X.