Learning Invariant Representations with Missing Data

Mark Goldstein* **Jörn-Henrik Jacobsen[1]** **Olina Chau[1]** **Adriel Saporta[1]**

Aahlad Puli[2] **Rajesh Ranganath[2]** **Andrew C. Miller[1]**

[1]Apple [2]New York University

Abstract

Spurious correlations allow flexible models to predict well during training but poorly on related test populations. Recent work has shown that models that satisfy particular independencies involving correlation-inducing *nuisance* variables have guarantees on their test performance. Enforcing such independencies requires nuisances to be observed during training. However, 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. Here we derive MMD estimators used for invariance objectives under missing nuisances. On simulations and clinical data, optimizing through these estimates achieves test performance similar to using estimators that make use of the full data.

1 Introduction

Spurious correlations allow models that predict well on training data to have worse than chance performance on related populations at test time [9, 27, 35, 22, 13, 32]. For example, diabetes is associated with high body mass index (BMI) in the United States. However, in India and Taiwan, diabetes also frequently co-occurs with low and average BMI [37]. Due to their shifting relationship with the label, nuisance variables (e.g., BMI) can cause models to exploit correlations in training data, leading them to generalize poorly on test sets of interest.

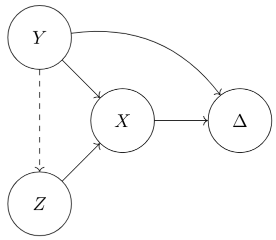
Invariant prediction methods are designed to improve performance on a range of test distributions when training data exhibits spurious correlations [26, 3]. We focus on methods that enforce independencies between the model and nuisance given some assumed causal structure [22, 35, 27]. These methods require 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 report other correlated conditions (e.g., hypertension, depression) or demographics (e.g., gender).

To improve generalization on a range of test distributions, it is necessary to handle missingness appropriately. However, extending invariant prediction methods to handle missing data is not straightforward. This difficulty stems from the invariant method's optimization objective, which usually includes a measure of dependence — e.g., Maximum Mean Discrepancy (MMD) or Mutual Information (MI) — that requires a sample from the fully-observed data to estimate consistently.

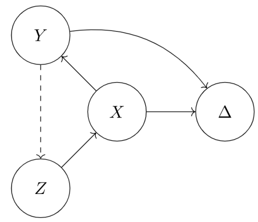
We propose MMD estimators for measuring nuisance-model dependence under missingness. First, we show that enforcing independence on only the nuisance-observed data does not imply independence on the full data, and vice versa. Next, we derive three estimators, including one that is *doubly-robust*: it is consistent when either the nuisance or missingness can be consistently modeled [4]. Using

*Work done while at Apple. Correspondence to <goldstein@nyu.edu> and <acmiller@apple.com>.

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).



(a) Anti-Causal



(b) Causal

Figure 1: Generative processes we consider in this work. The $Y \to Z$ edge 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.

simulations, a semi-simulation using textured MNIST and clinical data from MIMIC, we show that the estimators perform close to ground-truth estimation with no missingness and that they improve test accuracy relative to MMD computation using only the data with nuisances observed.

2 Notation and background

Notation. Let X denote features. Let Y be a label such as disease status. Let Z denote the nuisance, e.g., another disease correlated with Y, demographics, or image backgrounds. 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. We write functions as $f_X = f(X)$ to avoid excess parentheses. Let $h_X = h(X)$ denote a model to predict Y. Let $Y \sim \mathcal{B}(p)$ denote $Y \sim$ Bernoulli(p).

Assumptions and scope. The estimators require ignorability, $Z \perp\!\!\!\perp \Delta | X, Y$ [15]. Some distributions that satisfy this are shown in Figure 1. We require positivity $0 < \epsilon \leq P(\Delta = 1 | X, Y)$ to observe Z appropriately. While we focus on the graph in Figure 1(a) with binary Z, the presented method can extend to continuous Z and to other graphs (e.g., Figure 1(b)).

Modeling under spurious correlations. 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 distribution is expected to have a different (Y, Z) relationship from the training distribution. For example, define a family of distributions, indexed by D, that varies along $P(Z \mid Y)$:

$$\mathcal{F} = \{P_D(X, Y, Z) = P(Y) P_D(Z|Y) P(X|Y, Z)\}.$$

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 (Appendix C). When it is possible to anticipate and observe nuisances during training, enforcing certain independence constraints [22, 35, 27] helps guarantee performance regardless of the nuisance-label relationship. For example, for this choice of $\mathcal{F}$, maximum likelihood estimation for $Y|h_X$ while enforcing the constraint $h_X \perp\!\!\!\perp Z \mid Y = y$ for $y \in \{0, 1\}$ implies equal performance on all members, and better than chance on all members (Appendix D).

Missingness. When Z is subject to missingness, two parts of the data distribution help estimate functionals like $\mathbb{E}[Z]$: the missingness process $G_X \triangleq \mathbb{E}[\Delta \mid X, Y]$ and conditional expectation $m_X \triangleq \mathbb{E}[Z \mid X, Y]$. We review estimators that use either G_X [16, 5, 29] or m_X [30, 33] in Appendix H. The *doubly-robust* (DR) estimator [28, 4, 18] combines both by noting the equality:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{\Delta \tilde{Z}}{G_X} - \frac{\Delta - G_X}{G_X} m_X\right]. \tag{1}$$

Replacing G_X or m_X with models, Monte Carlo estimates of the right side of Equation (1) are consistent for $\mathbb{E}[Z]$ when, for all X, *either* G_X or m_X are consistently estimated (Appendix I).

3 Invariant representations with missing data

In the anti-causal setting, [35] enforce $h_X \perp\!\!\!\perp Z | Y = y$ for $y \in \{0, 1\}$ by minimizing the MMD:

$$\max_h \quad \log p(y|h_X) - \lambda \cdot \sum_{y \in \{0,1\}} \text{MMD}\left(p(h_X | Z = 1, Y = y), p(h_X | Z = 0, Y = y)\right). \tag{2}$$

First, we demonstrate what can go wrong when enforcing this MMD penalty only on samples where Z is observed. We then derive estimators of the full-data MMD under missingness.

3.1 Failures of restricting to observed data

Restricting computation to data with non-missing Z enforces $h_X \perp\!\!\!\perp Z|Y = y, \Delta = 1$ instead of $h_X \perp\!\!\!\perp Z|Y = y$. We show that these conditions do not imply each other.

Proposition 1. There exist distributions on (X, Y, Δ, Z) such that

$$\exists h^\star_X \quad \text{s.t.} \quad h^\star_X \perp\!\!\!\perp Z|Y = y, \quad \text{but} \quad h^\star_X \not\perp\!\!\!\perp Z|Y = y, \Delta = 1$$

and there exist distributions on (X, Y, Δ, Z) such that

$$\exists h^\star_X \quad \text{s.t.} \quad h^\star_X \perp\!\!\!\perp Z|Y = y, \Delta = 1 \quad \text{but} \quad h^\star_X \not\perp\!\!\!\perp Z|Y = y$$

The proof is in Appendix G. This implies that (1) optimizing the observed-only MMD can discard a solution to the full-data MMD, and (2) using the observed-only data may lead one to believe a model is invariant when it is not. This means one must enforce independence on the full data.

3.2 MMD estimation under missingness

We present estimators of the full-data unconditional MMD. One can compute the MMD conditional on $Y = y$ simply by restricting samples to those with $Y = y$. For a kernel k and $k_{XX'} \triangleq k(h_X, h_{X'})$,

$$\text{MMD}\big(p(h_X|Z = 1), p(h_X|Z = 0)\big) = \mathop{\mathbb{E}}_{\substack{X|Z=1\\X'|Z'=1}} k_{XX'} + \mathop{\mathbb{E}}_{\substack{X|Z=0\\X'|Z'=0}} k_{XX'} - 2 \mathop{\mathbb{E}}_{\substack{X|Z=1\\X'|Z'=0}} k_{XX'}\,. \tag{3}$$

Estimation is challenging due to missingness in the conditioning set. For $b \in \{0, 1\}$, let $N(b, b') = P(Z = b)P(Z' = b')$ and let $Z_1 \triangleq Z$ and $Z_0 \triangleq 1 - Z$. The dependence on Z can be re-written:

$$\mathop{\mathbb{E}}_{\substack{X|Z=b\\X'|Z'=b'}} k_{XX'} = \frac{1}{N(b, b')} \mathbb{E}\left[k_{XX'} \cdot Z_b \cdot Z_{b'}\right]. \tag{4}$$

Under no missingness, each expectation could be estimated with Monte Carlo. We now develop three MMD estimators. We derive simpler G_X-based and m_X-based estimators in Equations (10) and (11) (Appendix J). Here we combine them. Let $m_{X1} \triangleq m_X$, $m_{X0} \triangleq 1 - m_X$, and $G_{XX'} \triangleq G_X G_{X'}$.

Proposition 2. (DR estimator). Assume positivity, ignorability, and $\forall X$, $G_X = \mathbb{E}[\Delta|X, Y]$ or $m_X = \mathbb{E}[Z|X, Y]$. Then,

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

The proof is in Appendix J. We can use any of Equations (5), (10) and (11) to estimate the terms in eq. (3). Each of Equations (5), (10) and (11) is a ratio of two expectations: the normalization constant $N(b, b')$ depends on $\mathbb{E}[Z]$ and must itself be estimated under missingness (e.g., with Equation (1)). The ratio of consistent estimates of these quantities is consistent by Weak Law of Large Numbers and Slutsky's theorem. We discuss estimation in practice, trade-offs among the three estimators, and variance in Appendix E. We review recent related work in Appendix A.

4 Experiments

We compare accuracy and MMD minimization using different estimators: NONE (MLE only, no MMD), FULL (MLE and MMD using data with Z fully-observed), OBS (MLE and observed-only MMD), DR (MLE and DR estimator, called DR+ when using true G_X), IP (MLE and re-weighted estimator, called IP+ when using true G_X), and REG (MLE and regression estimator).

We first compare these algorithms in a simulation study. We then use textured MNIST to show the utility of the proposed estimators on high-dimensional data. In quantitative tables, we show mean $\pm$ standard deviation over three seeds. We then predict hospital length of stay in the MIMIC dataset, and compare performance when demographic nuisances are subject to missingness. For the $Y|X$ predictive loss, we use negative Bernoulli log likelihood with logit equal to h_X.

Table 1: Simulation. $\lambda = 1$. NONE has highest MMD and lowest test accuracy. OBS improves over this. The DR and REG methods are able to bring the MMD close to 0.0 and attain best test accuracy.

	NONE	OBS	FULL	DR	DR+	REG
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
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
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

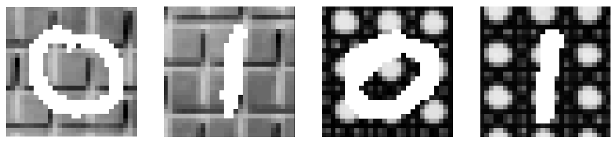


Figure 2: Textured MNIST with digits 0,1 on two textures from the Brodatz dataset.

Comparing MMDs. In all tables, the training set MMD for each method is computed using the ground-truth full-data MMD estimation method (see eq. (4)). This is also what the FULL model optimizes. True Z's are available in both simulated and real data as missingness is simulated. However, each model trains and validates the $\log p(Y|X) +$ MMD loss using its own estimation method.

4.1 Experiment 1: Simulation.

We set up strong (Y, Z) correlation. With $\overline{Y} = 1 - Y$, the training and validation sets are drawn:

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

The test set has the opposite relationship $Z \sim \mathcal{B}(.1Y + .9\overline{Y})$. Here $h_X^\star = (X_1 + X_2)/2$ predicts Y with smallest MSE among representations satisfying independence. We construct Δ to show the failure of computing MMD on the observed-only subset. For this, we use $\hat{Z} \triangleq -(X_1 - X_2)/2$, which is correlated with Z. We draw Δ conditional on $h_X^\star$ and $\hat{Z}$ (both are functions of X):

$$Q = \mathbb{1}\left[h_X^\star > 0.6\right] \cdot \mathbb{1}\left[\hat{Z} < 0.6\right], \quad \Delta \sim \mathcal{B}(Q + 0.2\overline{Q})\,.$$

This example construction leads to $h_X^\star \perp\!\!\!\perp Z | Y$ but $h_X^\star \not\perp\!\!\!\perp Z | Y, \Delta = 1$. For h, G_X and m_X we use small feed-forward neural networks (Appendix L.1).

Results. In Table 1, the DR estimators achieve indistinguishable performance to the full-data MMD, both in MMD and accuracy, and better than NONE and OBS. We include more results in Appendix F.

4.2 Experiment 2: Textured MNIST.

Following [11][2], we correlate MNIST digits 0 and 1 with two textures from the Brodatz dataset (Figure 2). The missingness is based on the average pixel intensity of X and its class. For h, G_X and m_X we use small convolutional networks. We include more details in Appendix L.2.

Results. In Table 2, NONE and OBS perform poorly on test. In contrast, the DR estimators — including the one with a learned G_X, m_X — achieve close to FULL's performance.

4.3 Experiment 3: Predicting length of stay in the ICU

We predict length of stay in the intensive care unit (ICU) in MIMIC [17][3] using demographics and first day labs/vitals among patients that stay at least one day. The prediction task is whether the stay is more than 2.5 days. To demonstrate that spurious correlations cause issues at deployment, we choose

[2] We adapt this repository (linked) to construct textured MNIST and will make our code available.

[3] The MIMIC critical-care database is available on Physionet [10].

Table 2: MNIST $\lambda = 1$. DR and REG estimators achieve close to full performance as measured by full MMD$= 0$ and high test accuracy. NONE and OBS perform poorly on test. OBS is notably high variance.

	NONE	OBS	FULL	DR	DR+	REG
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
TR ACC	0.90 ± 0.01	0.74 ± 0.03	0.76 ± 0.01	0.77 ± 0.0	0.76 ± 0.01	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	0.73 ± 0.01

Table 3: MIMIC $\lambda = 1$. REG estimator matches FULL's performance and improves upon OBS while DR does not, due to high objective variance during training (not shown in table).

	NONE	OBS	FULL	DR	REG
TR MMD	0.017 ± 0.02	0.002 ± 0.01	0.00 ± 0.00	0.009 ± 0.01	0.00 ± 0.00
TR ACC	0.71 ± 0.02	0.68 ± 0.01	0.70 ± 0.01	0.70 ± 0.01	0.71 ± 0.00
TE ACC	0.64 ± 0.00	0.64 ± 0.00	0.66 ± 0.00	0.62 ± 0.00	0.66 ± 0.01

$Z = 1$ to indicate the patient is recorded as white. While race may be correlated with health outcomes (e.g., due to unobserved socioeconomic factors [24]), it may not always be appropriate for a model to use this information [6]. The test set represents a new population with different outcome-demographic structure: we split the data so that the training/validation set has mostly samples with $Y \neq Z$ while the test set has mostly samples with $Y = Z$. We set non-male patients to have Z observed with probability 0.2. We include more details in Appendix M.

Results. In this real data setting with strong (Y, Z) correlation, the full-data MMD estimator reported in the table for all methods may have high variance. We focus on the attained accuracies. The REG estimator matches the ground-truth FULL estimator and performs better than OBS and DR. This is not unexpected, since it is possible for the REG estimator to be lower variance than DR when the true G_X is small or G_X is not modeled well [8], especially under strong (Y, Z) correlation (Appendix E).

5 Conclusion

We present estimators for the MMD that extend recent invariant prediction methods to missing data. Unlike prior estimators that only leverage data with nuisances observed, or consider worst-case estimation (see related work in Appendix A), the presented estimators of the full data objective are consistent when either auxiliary model can be learned. As we show in proposition 1, estimation of the full data objective is necessary to preserve the theoretical properties of invariant prediction methods. In the experiments, the DR and REG estimators are able to match full-data MMD performance and improve test accuracy relative to the OBS estimator. In practice, we recommend exploring the two simpler proposed estimators (REG and IP) in addition to the DR estimator and selecting the model based on the validation metric.

Moving forward, one limitation is that the full-data estimator — used as ground-truth MMD evaluation for the experiments — may itself have high variance on small datasets with strong nuisance-label correlation. Variance reduction is an important avenue both for optimizing and evaluating with the MMD using smaller batch sizes (in our experiments, batch sizes 1500 for MNIST and 4000 for MIMIC are large). Beyond variance reduction, it is a promising direction to apply the methodology in this work to the mutual information objective in [27], which sidesteps the choice of kernel and may be better suited for continuous and high dimensional nuisances.

Acknowledgments and Disclosure of Funding

The authors thank Scotty Fleming, Joe Futoma, Leon Gatys, Sean Jewell, Tayor Killian, and Guillermo Sapiro for feedback and discussions. This work was in part supported by NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, and NSF Award 1815633 SHF.

References

[1] F. Ahmed, Y. Bengio, H. van Seijen, and A. Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2020.

[2] Anonymous. Invariant learning with partial group labels. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=sWbXSWzHPa. under review.

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

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

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

[6] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

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

[8] M. Davidian. Double robustness in estimation of causal treatment effects, 2005.

[9] 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.

[10] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

[11] 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.

[12] 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.

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

[14] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

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

[16] 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.

[17] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[18] 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.

[19] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

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

[21] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[22] 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.

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

[24] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

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

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

[35] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] 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.

[37] 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 Related work

We focus on recent work in fairness and invariant prediction on missing group/environment labels. Motivated by fairness, [36] study a related problem of optimizing invariance-inducing objectives when the protected group label (analogous to our nuisance variables) is noisy. Given bounds on the level of label noise, this work proposes optimizing an objective based on the *distributionally robust optimization* framework [23]. 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.

[19] optimize worst-case-over-groups performance without known group labels. They rely on the assumption that groups are *computationally-identifiable* (i.e. that there exists some function on the data that labels their protected group membership) [14] and use a model to identify groups on which performance is worst. They pose an adversarial optimization between the group-labeling model — which searches for groups with poor performance — and the primary predictive model. Inspired by this work, [7] find worst-case group assignments based on an empirical risk minimization (ERM) model that maximizes invariance penalties and [1] illustrate that this objective performs well on a wide range of benchmarks. Relatedly, [21] run usual ERM training and then a second iteration of ERM that upweights the loss for datapoints on which the model performs badly. This identifies groups with bad model performance without explicit group labels. In both of these works, the groups could be seen either as a nuisance variable or as a confounder that correlates the label and some nuisance variable. However, in our setting (and in that of [22, 35, 27]), in exchange for being willing to make assumptions on the test distribution family, we do not need to observe samples with poor model performance at training time (and may not see any) to prevent sudden decreases in performance on held-out data at test time.

[2] too consider invariant learning with partial/missing group labels. They maximize over unknown group assignments to get worst-case (over group assignments) versions of invariance objectives. As opposed to the DR estimator presented here, in their work the objective still depends on the missingness distribution: the objective value can take values closer or further from 0 for the same model but different missingness distributions.

B Measuring dependence with MMDs

To enforce independencies it is necessary to measure dependence, i.e., to measure the distance between a joint distribution and the product of its marginals. The general Integral Probability Metrics (IPMs) class defines a metric on distributions. A special case, the kernel-based MMD [12], has a closed form. Let $X_1 \sim P, X_2 \sim Q$. Let X_j' be an independent sample identically distributed as X_j. For kernel k,

$$\text{MMD}(X_1, X_2) = \mathbb{E}[k(X_1, X_1')] + \mathbb{E}[k(X_2, X_2')] - 2\,\mathbb{E}[k(X_1, X_2)]\,. \tag{7}$$

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

C Out-of-Distribution example

Example. Let $\overline{Y} = 1 - Y$ and define P_{train} by:

$$Y \sim \mathcal{B}(0.5), \quad Z \sim \mathcal{B}(0.9Y + 0.1\overline{Y}), \quad X \sim [\mathcal{N}(Y, 10), \mathcal{N}(Z, 1)]$$

X_1 directly contains Y and X_2 contains information about Y through Z. 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 for $a \in \{0, 1\}$ we have $P_{test}(Z = a \mid Y = a) = 0.9(1 - a) + 0.1a$, then unlike in P_{train}, higher values correspond to label $Y = 1$.

D Invariant predictor

There are at least two distinct usages of the word *invariance* in the literature. For one usage, invariance to the nuisance usually implies independence of model and nuisance, potentially conditional on label. For the other usage, invariance refers to invariant risk, i.e., the risk is the same for all test distributions in some family. In some distribution families and for some independence constraints, these can imply each other. Here, we show that the particular conditional independence involving the nuisance studied in this work and originally from [22, 35, 27] implies invariant risk.

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:

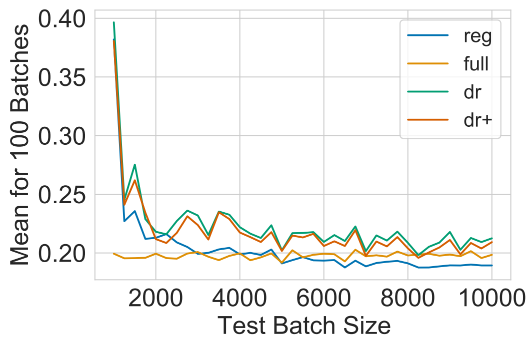
$$
\begin{aligned}
\mathbb{E}_{P_{test}(Y,X)}[\log P_{train}(Y|h_X)] &= \mathbb{E}_{P_{test}(Y,X)}\left[\log \frac{P_{train}(h_X|Y)P(Y)}{P_{train}(h_X)}\right] \\
&= \mathbb{E}_{P_{test}(Y,X,Z)}\left[\log \frac{P_{train}(h_X|Y,Z)P(Y)}{\mathbb{E}_{P(Y)}\left[P_{train}(h_X|Y,Z)\right]}\right] \\
&= \mathbb{E}_{P_{test}(Y,h_X,Z)}\left[\log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

The last quantity does not depend on any specific $P_D(Z|Y)$. This means that performance of the $P_{train}(Y|h_X)$ model, when the independence is satisfied, is the same on all P_{test} in $\mathcal{F}$.

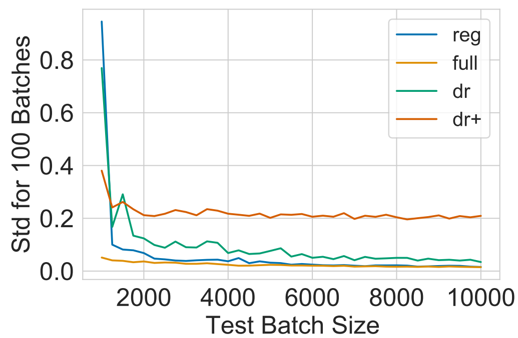
E Estimation in practice

E.1 Splitting samples

For a given batch, we use $1/4$ of the samples for the normalization term and $3/4$ for the main term, though this number may be changed. Further, the main term of any of the three estimators is defined on a pair of independent samples, i.e. it is a *U-statistic*. There are two ways to estimate such expectations. One option is to further break the samples left for the main term in half into two batches S_1 and S_2 and then compute on all pairs $i \in S_1, j \in S_2$. The alternative, which has slightly higher sampler efficiency and is the method we use, is to compute on all pairs of samples and then leave out any diagonal terms $k(X_i, X_i)$ from the average.



(a) Means



(b) Std. Dev.'s

Figure 3: Figure 3(a): Mean of 100 MMD estimates at each batch size. Figure 3(b): Standard Deviation of 100 MMD estimates at each batch size

E.2 Trade-offs among the 3 proposed estimators

For large samples, DR estimates with correct G_X and correct m_X are lower variance than the regression with correct m_X, and lower variance than re-weighting with correct G_X. Even when m_X is mis-specified but G_X is correct, the DR estimator may still be lower variance than the re-weighting estimator with correct G_X alone. However, the DR estimator with correct m_X but mis-specified G_X may be *higher variance* than the regression estimator with correct m_X [8]. For this reason, when the missingness model G_X is wrong, the regression estimator may out-perform the DR estimator even in large sample sizes.

The variance of the DR and re-weighting estimators comes from two distinct places. One is general to missingness: small observation probabilities G_X in the denominator. The other reason is general Monte Carlo error: we need individual samples of $\tilde{Z}$ in the numerator. This is *especially a problem in the spurious correlation setting*: Y and Z are possibly strongly correlated. We need to compute the MMD conditional on $Y = y$ which involves, for each $Y = y$, expectations using samples where $Z = 1$ and where $Z = 0$, but we may have very few samples for one of these Z values. This second source of variance *also applies to estimates of the full-data* MMD under no missingness (eq. (4)). We compare the mean and variance of these estimators empirically in Appendix E.3.

E.3 Empirical investigation of variance

As discussed, when $\mathbb{E}[\Delta|X]$ small, or (Y, Z) highly correlated, or both, all estimators will be high variance. We train a model on the experiment 1 simulation using the NONE method and then study the mean and variance of DR, DR+ (to study the effect of using the true G_X), REG (since it yielded better performance on MIMIC) and FULL (since this method is used to report the MMDs in the tables). In this simulation, we are free to generate as many large batches of samples as needed. Keeping the model fixed, for each batch size between 1000 and $10,000$ incrementing by 250 we draw 100 new batches of that size and estimate the MMD using each method. For each method, we report the mean (fig. 3(a)) and standard deviation (fig. 3(b)) of these estimates.

Notably, we cannot compute an actual ground-truth for the MMD of this model, but we could take the mean of the FULL estimate (no missingness) at the largest sample size of $10,000$ samples. This is about 0.2. We see that the regression estimator stays closer to this number for all sample sizes relative to the DR methods. Interesting, for standard deviation, we see that the DR estimator is more well-behaved than the DR+ estimator that uses the true G_X. This has also been observed for learned versus true propensity scores in treatment effect estimation and usually results from models learning less extreme probabilities than the true ones, trading some bias. In this case, there is not a substantial difference in estimated weights or in bias, but there is a large difference in variance. More investigation is required.

The main take-away from both plots is that the regression method seems more stable than DR and that G_X may be the part of the DR estimator that is not being learned well. On the other hand the DR estimator may possibly be safer when it is unknown if it is easier to estimate G_X or m_X. We recommend using all 3 of the proposed estimators and comparing validation objectives.

F Full experiments

Table 4: Simulation. $\lambda = 1$.

	NONE	OBS	FULL	DR	DR+	REG	IP	IP+
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 5: Simulation. $\lambda = 5$.

	NONE	OBS	FULL	DR	DR+	REG	IP	IP+
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 6: MNIST $\lambda = 1$.

	NONE	OBS	FULL	DR	DR+	REG	IP	IP+
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 7: MNIST $\lambda = 5$.

	NONE	OBS	FULL	DR	DR+	REG	IP	IP+
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

G Failures of restricting to observed data

Proposition. There exist distributions on (X, Y, Δ, Z) such that

$$\exists h_X^\star \quad \text{s.t.} \quad h_X^\star \perp\!\!\!\perp Z | Y = y, \quad \text{but} \quad h_X^\star \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$$

and there exist distributions on (X, Y, Δ, Z) such that

$$\exists h_X^\star \quad \text{s.t.} \quad h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1 \quad \text{but} \quad h_X^\star \not\perp\!\!\!\perp Z | Y = y$$

First direction. There exist distributions on (X, Y, Δ, Z) such that

$$\exists h_X^\star \quad \text{s.t.} \quad h_X^\star \perp\!\!\!\perp Z | Y = y, \quad \text{but} \quad h_X^\star \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$$

It suffices to illustrate this even when Z, Y are not correlated. Consider

$$Y \sim \mathcal{N}(0, 1), \quad Z \sim \mathcal{N}(0, \sigma_Z^2), \quad \epsilon_X \sim \mathcal{N}(0, \sigma_X^2), \quad X = [Y - Z + \epsilon_X, Y + Z]$$

For $h_X^\star = (X_1 + X_2)$, we first show $h_X^\star \perp\!\!\!\perp Z | Y = y$. We have

$$h_X^\star | Y \sim \mathcal{N}(2Y, \sigma_X^2)$$

and in particular $h_X^\star = 2Y + \epsilon_X$. Given $Y = y$, the only randomness in $h_X^\star$ is due ϵ_X. But ϵ_X is independent of the joint variable (Z, Y) meaning $\epsilon_X \perp\!\!\!\perp Z | Y = y$ and therefore $h_X^\star \perp\!\!\!\perp Z | Y = y$.

We now construct $\Delta | (X, Y)$ such that $h_X^\star \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$. Let

$$\Delta = \text{OR}\Big(\mathbb{1}\left[X_1 + X_2 < 0\right], \mathbb{1}\left[X_2 - Y < 0\right]\Big).$$

Checking the condition

$$h_X^\star \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$$

(using definition of $h_X^\star$) is equivalent to checking

$$(X_1 + X_2) \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$$

(using definition of Δ) is equivalent to checking

$$(X_1 + X_2) \not\perp\!\!\!\perp Z | Y = y, \text{OR}\Big(\mathbb{1}\left[X_1 + X_2 < 0\right], \mathbb{1}\left[X_2 - Y < 0\right]\Big) = 1$$

(using defintion of X_2) is equivalent to checking

$$(X_1 + X_2) \not\perp\!\!\!\perp Z | Y = y, \text{OR}\Big(\mathbb{1}\left[X_1 + X_2 < 0\right], \mathbb{1}\left[Z < 0\right]\Big) = 1$$

To check that, we need to check if the distribution of $(X_1 + X_2) | Y = y, \Delta = 1$ changes when conditioning on different events involving the random variable Z. For example, $\mathbb{1}\left[Z < 0\right]$ and $\mathbb{1}\left[Z \geq 0\right]$:

1. $(X_1 + X_2) \quad | \quad Y = y, \text{OR}\Big(\mathbb{1}\left[X_1 + X_2 < 0\right], \mathbb{1}\left[Z < 0\right]\Big) = 1, \mathbb{1}\left[Z < 0\right] = 1$
2. $(X_1 + X_2) \quad | \quad Y = y, \text{OR}\Big(\mathbb{1}\left[X_1 + X_2 < 0\right], \mathbb{1}\left[Z < 0\right]\Big) = 1, \mathbb{1}\left[Z \geq 0\right] = 1.$

We can show these two conditional variables differ in distribution simply by showing they differ in support. The first conditional variable can be full support because the event $\mathbb{1}\left[Z < 0\right]$ satisfies one of the OR conditions leaving the other condition $\mathbb{1}\left[X_1 + X_2 < 0\right] = \mathbb{1}\left[h_X^\star < 0\right]$ free to take either value. However, the second conditional variable needs $X_1 + X_2 = h_X^\star < 0$ because $\mathbb{1}\left[Z < 0\right]$ is not satisfied (since we condition on $\mathbb{1}\left[Z \geq 0\right] = 1$) but the OR has to be 1. These different supports imply the distributions differ. That the variables differ on two non-measure zero sets is enough to show dependence. Then $(X_1 + X_2) \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$ which means $h_X^\star \not\perp\!\!\!\perp Z | Y = y, \Delta = 1$.

Second direction. There exist distributions on (X, Y, Δ, Z) such that

$$\exists h_X^\star \quad \text{s.t.} \quad h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1 \quad \text{but} \quad h_X^\star \not\!\perp\!\!\!\perp Z | Y = y$$

Let the data be drawn as

$$Y \sim \mathcal{N}(0,1), \quad Z \sim \mathcal{B}(0.5), \quad X = [Y - Z, Y + Z]$$

Let $h_X^\star = \mathbb{1}[X_1 \geq 0]$. We first show $h_X^\star \not\!\perp\!\!\!\perp Z | Y = y$. We have

$$\begin{aligned} h_X^\star &= \mathbb{1}[X_1 \geq 0] \\ &= \mathbb{1}[Y - Z \geq 0] \end{aligned}$$

Given $Y = y$, we ask if the random variable $\mathbb{1}[y - Z \geq 0]$ is independent of Z. To show dependence, we show that the random variable $\mathbb{1}[y - Z \geq 0]$ changes in distribution when Z takes on its two values:

1. $\mathbb{1}[y - Z \geq 0] | Y = y, Z = 0$
2. $\mathbb{1}[y - Z \geq 0] | Y = y, Z = 1$

Suppose $y \in (0, 1)$. When $Z = 0$ we have that $\mathbb{1}[y - Z \geq 0] = 1$ with probability one. When $Z = 1$, we have $\mathbb{1}[y - Z \geq 0] = 0$ with probability one. Therefore the variables are dependent.

We now let $\Delta = \mathbb{1}[X_1 \geq 0] = \mathbb{1}[Y - Z \geq 0]$ and show $h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1$. Note that $\Delta(X, Y) = h_X^\star$. We ask whether

$$\mathbb{1}[Y - Z \geq 0] \perp\!\!\!\perp Z | Y = y, \mathbb{1}[Y - Z \geq 0]$$

The conditioning set fully determines the variable $\mathbb{1}[Y - Z \geq 0]$ meaning it is a constant and is therefore independent of Z. Therefore $h_X^\star \perp\!\!\!\perp Z | Y = y, \Delta = 1$ as desired.

H 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:

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

Inverse-weighting estimators use G_X [16, 5, 29, 34, 15]

$$\begin{aligned} \mathbb{E}[Z] &= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X} [Z] \\ &= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X} \left[\frac{\mathbb{E}[\Delta|X]}{\mathbb{E}[\Delta|X]} Z \right] \\ &= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{Z|X} \mathop{\mathbb{E}}_{\Delta|X} \left[\frac{\Delta Z}{\mathbb{E}[\Delta|X]} \right] \\ &= \mathop{\mathbb{E}}_{XZ\Delta} \left[\frac{\Delta Z}{\mathbb{E}[\Delta|X]} \right] \\ &= \mathop{\mathbb{E}}_{X\Delta Z} \left[\frac{\Delta Z}{G_X} \right] \\ &= \mathop{\mathbb{E}}_{X\Delta Z} \left[\frac{\Delta \tilde{Z}}{G_X} \right] \end{aligned} \tag{8}$$

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 [30, 33, 31, 25, 20, 15]:

$$\mathbb{E}[Z] = \mathop{\mathbb{E}}_{X} \Big[\mathbb{E}[Z|X] \Big] = \mathop{\mathbb{E}}_{X} \Big[\mathbb{E}[Z|X, \Delta = 1] \Big] = \mathop{\mathbb{E}}_{X} [m_X] \tag{9}$$

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

I DR estimator of mean of Z

The inverse weighting and regression estimators can be combined. Equation (1) 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):

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

J MMD estimators under missingness

J.1 Statement of G_x and m_x based estimators of the MMD

Proposition. (G_X-based re-weighted estimator) Assume positivity, ignorability, and, for each X, $G_X = \mathbb{E}[\Delta|X,Y]$. Then,

$$\underset{\substack{X|Z=b \\ X'|Z'=b'}}{\mathbb{E}} \left[k_{XX'}\right] = \frac{1}{N(b,b')}\,\mathbb{E}\left[\frac{\Delta\Delta'\tilde{Z}_b\tilde{Z'}_{b'}}{G_{XX'}}k_{XX'}\right]. \tag{10}$$

Proposition. (m_X-based regression estimator) Assume ignorability, and, for each X, $m_X = \mathbb{E}[Z|X,Y]$. Then,

$$\underset{\substack{X|Z=b \\ X'|Z'=b'}}{\mathbb{E}} \left[k_{XX'}\right] = \frac{1}{N(b,b')}\,\mathbb{E}\left[m_{Xb}\cdot m_{X'b'}\cdot k_{XX'}\right]. \tag{11}$$

J.2 Deriving the G_X-based re-weighted estimator

Here we start at the target quantity and derive the estimator. We give the derivation for $Z=1, Z'=1$. The other cases are analogous.

$$\begin{aligned}
&\underset{P(X|Z=1)P(X'|Z'=1)}{\mathbb{E}} \left[k_{XX'}\right] \\
&= \int_{X,X'} kP(X|Z=1)P(X'|Z'=1)dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1,X)P(Z'=1,X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} kP(Z=1|X)P(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\int_{X,X'} k\,\mathbb{E}(Z=1|X)\,\mathbb{E}(Z'=1|X)P(X)P(X')dXdX' \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\underset{\substack{X,Z \\ X',Z'}}{\mathbb{E}}\left[k\cdot Z\cdot Z'\right] \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\underset{\substack{X,Z \\ X',Z'}}{\mathbb{E}}\left[\frac{\mathbb{E}[\Delta|X]\,\mathbb{E}[\Delta'|X']}{\mathbb{E}[\Delta|X]\,\mathbb{E}[\Delta'|X']}k\cdot Z\cdot Z'\right] \\
&= \frac{1}{P(Z=1)}\frac{1}{P(Z'=1)}\underset{\substack{X,\Delta,Z \\ X',\Delta',Z'}}{\mathbb{E}}\left[\frac{\Delta\Delta'}{G_XG_{X'}}k\cdot Z\cdot Z'\right]
\end{aligned}$$

J.3 Deriving the m_X-based standardization estimator

Here we start at the target quantity and derive the estimator. We give the derivation for $Z = 1, Z' = 1$. The other cases are analogous.

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

J.4 Deriving the DR estimator

Here we start at the estimator and derive the target quantity. We give the derivation for $Z = 1, Z' = 1$. The other cases are analogous.

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

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 $Z=0$.

K kernel mmd between joint and product of marginals

Continuous nuisances. In this work we study binary nuisance. We can instead measure the MMD between joint $p(h_X, Z)$ and product of marginals $p(h_X)P(Z)$, which allows for continuous nuisance.

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. Let $X :: Z$ denote the concatenation of X and Z. The more general formulation is:

$$\mathbb{E}_{\substack{(X,Z)\sim P(X,Z)\\(X',Z')\sim P(X,Z)}}\left[k\Big(X{::}Z, X'{::}Z'\Big)\right] + \mathbb{E}_{\substack{(X,Z)\sim P(X)P(Z)\\(X',Z')\sim P(X')P(Z')}}\left[k\Big(X{::}Z, X'{::}Z'\Big)\right] \\ - 2\mathbb{E}_{\substack{(X,Z)\sim P(X,Z)\\(X',Z')\sim P(X')P(Z')}}\left[k\Big(X{::}Z, X'{::}Z'\Big)\right]$$

This leads to the following estimator:

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

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 distribution. Moreover, these conditional expectations depend on the current representation h_X. This means they must be updated each time h changes.

L Experimental Details

L.1 Simulation details

for h we use a small feed-forward neural network

```
class SmallNet(nn.Module):
  def __init__(self, D_in,D_hid,D_out):
      super(SmallNet, self).__init__()
      self.fc1 = nn.Linear(D_in,D_hid)
      self.fc2 = nn.Linear(D_hid,D_hid)
      self.fc3 = nn.Linear(D_hid,D_out)
  def forward(self,h):
      h = RELU(self.fc1(h))
      h = RELU(self.fc2(h))
      return self.fc3(h)
```

We use D_hid $= 128$. For G_X and m_X we use a similar model but that also takes y as input:

```
class SmallAuxNet(nn.Module):
  def __init__(self, D_in,D_hid,D_out):
      super(SmallAuxNet, self).__init__()
      self.fc1 = nn.Linear(D_in+1,D_hid)
      self.fc2 = nn.Linear(D_hid,D_out)
  def forward(self,x,y):
      h = torch.cat([x,y.unsqueeze(-1)],dim=-1)
      h = RELU(self.fc1(h))
      return self.fc2(h)
```

L.2 MNIST details

Data details. Following [11][4], we correlate MNIST digits 0 and 1 with two textures from the Brodatz dataset (Figure 2). This is an example of invariance to image backgrounds when not all background labels are available. We follow a similar setup to colored MNIST [3]: because $Y|X$ is essentially deterministic, even strong spurious correlations may be ignored by a model on MNIST. To push $Y|X$ closer to what may be expected in noisier 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 is based on the average pixel intensity of X and its class. Let μ_X be the mean pixel value of a 28x28 MNIST image. We set

$$Q = \mathbb{1}\left[Y = 1\right] \cdot \mathbb{1}\left[\mu_X < 0.3\right], \quad \Delta \sim \mathcal{B}(Q + .2\overline{Q}).$$

The choice of Q is correlated with Z through whether the image is light or dark grey. Similar to proposition 1 and experiment 1, this means subsetting on $\Delta = 1$ may not imply independence on the full population and may throw away solutions that do.

Model details. For h we use a small convolutional neural network

```
class MNISTNet(nn.Module):
    def __init__(self,args):
        super(MNISTNet,self).__init__()
        self.args = args
        self.conv1 = nn.Conv2d(1,6,5)
        self.conv2 = nn.Conv2d(6,16,5)
        self.pool = nn.MaxPool2d(2,2)
        self.fc1 = nn.Linear(256,64)
        self.fc2 = nn.Linear(64,64)
        self.fc3 = nn.Linear(64,1)
    def forward(self,x):
```

[4]We adapt this repository (linked) to construct textured MNIST and will make our code available.

```
        h = self.pool(self.conv1(x).relu())
        h = self.pool(self.conv2(h).relu())
        h = torch.flatten(h,1)
        h = self.fc1(h).relu()
        h = self.fc2(h).relu()
        return self.fc3(h)
```

For G_X and m_X we use a similar model but that also take y as input. We include y in the computation after the convolutions.

```
    class NumAuxNet(nn.Module):
    def __init__(self,args):
        super(NumAuxNet,self).__init__()
        self.args = args
        self.conv1 = nn.Conv2d(1,6,5)
        self.conv2 = nn.Conv2d(6,16,5)
        self.pool = nn.MaxPool2d(2,2)
        self.fc1 = nn.Linear(256+1,64)
        self.fc2 = nn.Linear(64,64)
        self.fc3 = nn.Linear(64,1)
    def forward(self,x,y):
        h = self.pool(self.conv1(x).relu())
        h = self.pool(self.conv2(h).relu())
        h = torch.flatten(h,1)
        h = torch.cat([h,y.unsqueeze(-1)],dim=-1)
        h = self.fc1(h).relu()
        h = self.fc2(h).relu()
        return self.fc3(h)
```

M MIMIC Data Details

We use the following SQL script to extract the dataset from the MIMIC-IV dataset following the instructions on Physionet for querying on Google Big-Query. This could also be done in a local MIMIC-IV database. We are grateful to the MIMIC team especially for creating the MIMIC-DERIVED tables which include aggregated vitals/labs measured during the first day of ICU stay.

```
SELECT
-- ids
pat.subject_id as subject_id, adm.hadm_id as hadm_id,icu.stay_id as stay_id,
-- demographics
CASE WHEN pat.gender="M" THEN 1 ELSE 0 END as is_male,
CASE WHEN adm.ethnicity="WHITE" THEN 1 ELSE 0 END as is_white,
icu_detail.admission_age as age,
-- weight height
fdw.weight , fdh.height ,
-- LOS
icu.los as los_icu_days,
-- death
adm.hospital_expire_flag as expire_flag,
-- vitals labs min max mean over first day
vitals.*, labs.*, sofa.*, bg.*,
FROM `physionet-data.mimic_core.patients` pat
INNER JOIN
    `physionet-data.mimic_core.admissions` adm
        on pat.subject_id=adm.subject_id
INNER JOIN
    `physionet-data.mimic_icu.icustays` icu
      on adm.subject_id=icu.subject_id
      and
```

```
      adm.hadm_id=icu.hadm_id
INNER JOIN
     `physionet-data.mimic_derived.first_day_height` fdh
        on
        adm.subject_id = fdh.subject_id  and icu.stay_id = fdh.stay_id
INNER JOIN
     `physionet-data.mimic_derived.first_day_weight` fdw
        on
        adm.subject_id = fdw.subject_id  and icu.stay_id = fdw.stay_id
INNER JOIN
     `physionet-data.mimic_derived.icustay_detail` icu_detail
        on
        adm.subject_id=icu_detail.subject_id
        and
        adm.hadm_id=icu_detail.hadm_id
        and
        icu.stay_id=icu_detail.stay_id
INNER JOIN
    `physionet-data.mimic_derived.first_day_bg` bg
    on
     adm.subject_id=bg.subject_id
        and
        icu.stay_id = bg.stay_id
INNER JOIN
    `physionet-data.mimic_derived.first_day_sofa` sofa
    on
        adm.subject_id=sofa.subject_id
        and
        adm.hadm_id=sofa.hadm_id
        and
        icu.stay_id=sofa.stay_id
INNER JOIN
     `physionet-data.mimic_derived.first_day_vitalsign` vitals
     on
    adm.subject_id=vitals.subject_id
    and
    icu.stay_id=vitals.stay_id
INNER JOIN
 `physionet-data.mimic_derived.first_day_lab` labs
     on
    adm.subject_id=labs.subject_id
    and
    icu.stay_id=labs.stay_id
where icu_detail.los_icu > 1
    and pat.gender is not null
    and adm.ethnicity is not null
    and adm.ethnicity != "UNABLE TO OBTAIN"
    and adm.ethnicity != "UNKNOWN"
```