

## Appendix A. Invariant predictor

Here, we show that  $h_X \perp\!\!\!\perp Z | Y = y$  for all  $y$  implies invariant risk in  $\mathcal{F}$ . The proof uses the following insights

- Satisfying independence  $h_X \perp\!\!\!\perp Z | Y$  means  $P_{train}(h_X | Y, Z) = P_{train}(h_X | Y)$ .
- The assumptions on  $\mathcal{F}$  mean  $P_{train}(h_X | Y, Z) = P(h_X | Y, Z)$  in any member of the family.
- Combined this means  $P_{train}(h_X | Y, Z) = P_D(h_X | Y)$  for any  $P_D \in \mathcal{F}$  when the model  $P_{train}(Y | h_X)$  satisfies  $h_X \perp\!\!\!\perp Z | Y$ .

**Proposition** *Suppose model  $h_X$  satisfies  $h_X \perp\!\!\!\perp Z | Y$  on any  $P_D \in \mathcal{F}$ . Then for all  $P_{D'} \in \mathcal{F}$ ,  $\mathbb{E}_{P_D}[\log p_D(Y | h_X)] = \mathbb{E}_{P_{D'}}[\log p_D(Y | h_X)]$ .*

**Proof** 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:

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

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  $\mathcal{F}$ .  $\blacksquare$

## Appendix B. Estimation in practice

### B.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.

### B.2. Trade-offs among the 3 proposed estimators

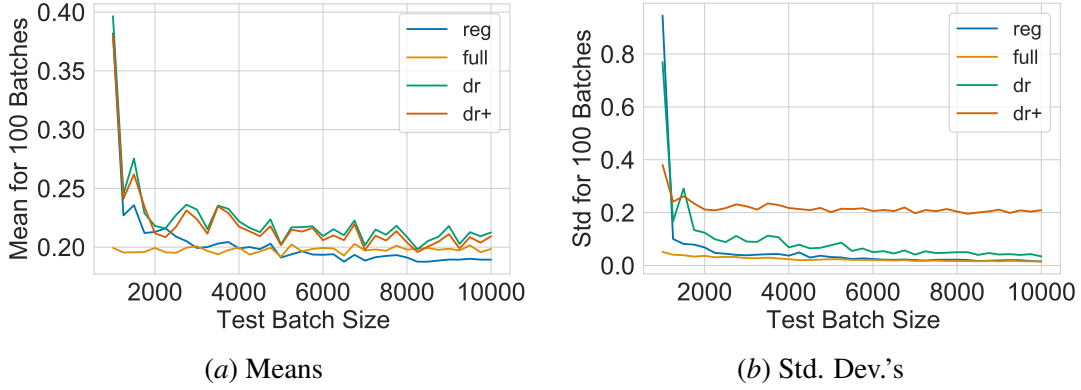
For large samples, DR estimates with correct  $G_W$  and correct  $m_W$  are lower variance than the regression with correct  $m_W$ , and lower variance than re-weighting with correct  $G_W$ . Even when  $m_W$  is mis-specified but  $G_W$  is correct, the DR estimator may still be lower variance than the re-weighting estimator with correct  $G_W$  alone. However, the DR estimator with correct  $m_W$  but mis-specified  $G_W$  may be *higher variance* than the regression estimator with correct  $m_W$  (Davidian, 2005). For this reason, when the missingness model  $G_W$  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  $G_W$  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. (6)). We compare the mean and variance of these estimators empirically in Appendix B.3.

### B.3. Empirical investigation of variance

As discussed, when  $\mathbb{E}[\Delta|X, Y]$  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  $G_W$ ), 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  $G_W$ . This has also been observed for learned versus true propensity scores in treatment effect estimation and usually results from models



**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

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  $G_W$  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  $G_W$  or  $m_W$ . We recommend using all 3 of the proposed estimators and comparing validation objectives.

## Appendix C. Full experiments

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

|        | NONE            | OBS             | FULL            | DR              | DR+             | REG             | IP              | IP+             |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| TR MMD | $0.21 \pm 0.04$ | $0.05 \pm 0.04$ | $0.00 \pm 0.01$ | $0.00 \pm 0.01$ | $0.00 \pm 0.01$ | $0.01 \pm 0.00$ | $0.00 \pm 0.00$ | $0.00 \pm 0.01$ |
| TR ACC | $0.89 \pm 0.00$ | $0.87 \pm 0.00$ | $0.86 \pm 0.01$ | $0.85 \pm 0.01$ | $0.84 \pm 0.02$ | $0.86 \pm 0.00$ | $0.84 \pm 0.01$ | $0.84 \pm 0.01$ |
| TE ACC | $0.67 \pm 0.02$ | $0.77 \pm 0.02$ | $0.80 \pm 0.01$ | $0.81 \pm 0.02$ | $0.81 \pm 0.01$ | $0.79 \pm 0.02$ | $0.82 \pm 0.02$ | $0.81 \pm 0.00$ |

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

|        | NONE            | OBS             | FULL            | DR              | DR+             | REG             | IP              | IP+             |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| TR MMD | $0.21 \pm 0.04$ | $0.03 \pm 0.02$ | $0.00 \pm 0.01$ | $0.00 \pm 0.00$ | $0.00 \pm 0.0$  | $0.00 \pm 0.01$ | $0.00 \pm 0.0$  | $0.00 \pm 0.0$  |
| TR ACC | $0.89 \pm 0.0$  | $0.85 \pm 0.02$ | $0.84 \pm 0.01$ | $0.82 \pm 0.01$ | $0.78 \pm 0.06$ | $0.84 \pm 0.00$ | $0.81 \pm 0.02$ | $0.81 \pm 0.03$ |
| TE ACC | $0.67 \pm 0.02$ | $0.78 \pm 0.02$ | $0.83 \pm 0.01$ | $0.82 \pm 0.02$ | $0.77 \pm 0.04$ | $0.82 \pm 0.01$ | $0.81 \pm 0.02$ | $0.80 \pm 0.01$ |

**Table 6:** MNIST  $\lambda = 1$ .

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

**Table 7:** MNIST  $\lambda = 5$ .

|        | NONE            | OBS             | FULL            | DR              | DR+             | REG             | IP              | IP+             |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| TR MMD | $2.05 \pm 0.18$ | $0.01 \pm 0.02$ | $0.00 \pm 0.00$ | $0.00 \pm 0.00$ | $0.00 \pm 0.01$ | $0.00 \pm 0.01$ | $0.01 \pm 0.01$ | $0.01 \pm 0.02$ |
| TR ACC | $0.9 \pm 0.01$  | $0.66 \pm 0.15$ | $0.75 \pm 0.01$ | $0.65 \pm 0.14$ | $0.65 \pm 0.13$ | $0.75 \pm 0.01$ | $0.71 \pm 0.08$ | $0.60 \pm 0.12$ |
| TE ACC | $0.13 \pm 0.01$ | $0.65 \pm 0.15$ | $0.75 \pm 0.01$ | $0.73 \pm 0.02$ | $0.70 \pm 0.09$ | $0.75 \pm 0.01$ | $0.55 \pm 0.3$  | $0.60 \pm 0.12$ |

**Appendix D. Failures of restricting to observed data**

**Proposition** *There exist distributions on  $(X, Y, \Delta, Z)$  such that*

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

*and there exist distributions on  $(X, Y, \Delta, Z)$  such that*

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

**First direction.** There exist distributions on  $(X, Y, \Delta, Z)$  such that

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

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

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

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

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

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

Checking the condition

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

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

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

(using definition of  $\Delta$ ) is equivalent to checking

$$(X_1 + X_2) \perp\!\!\!\perp Z | Y = y, \text{OR} \left( \mathbb{1}[X_1 + X_2 < 0], \mathbb{1}[X_2 - Y < 0] \right) = 1$$

(using definition of  $X_2$ ) is equivalent to checking

$$(X_1 + X_2) \perp\!\!\!\perp Z | Y = y, \text{OR} \left( \mathbb{1}[X_1 + X_2 < 0], \mathbb{1}[Z < 0] \right) = 1$$

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,  $\mathbb{1}[Z < 0]$  and  $\mathbb{1}[Z \geq 0]$ :

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

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  $\mathbb{1}[Z < 0]$  satisfies one of the OR conditions leaving the other condition  $\mathbb{1}[X_1 + X_2 < 0] = \mathbb{1}[h_X^* < 0]$  free to take either value. However, the second conditional variable needs  $X_1 + X_2 = h_X^* < 0$  because  $\mathbb{1}[Z < 0]$  is not satisfied (since we condition on  $\mathbb{1}[Z \geq 0] = 1$ ) 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  $(X_1 + X_2) \perp\!\!\!\perp Z | Y = y, \Delta = 1$  which means  $h_X^* \perp\!\!\!\perp Z | Y = y, \Delta = 1$ .

**Second direction.** There exist distributions on  $(X, Y, \Delta, Z)$  such that

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

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  $h_X^* = \mathbb{1}[X_1 \geq 0]$ . We first show  $h_X^* \perp\!\!\!\perp Z | Y = y$ . We have

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

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  $h_X^* \perp\!\!\!\perp Z | Y = y, \Delta = 1$ . Note that  $\Delta(X, Y) = h_X^*$ . 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  $h_X^* \perp\!\!\!\perp Z | Y = y, \Delta = 1$  as desired.

## Appendix E. 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  $G_W$  and the conditional expectation  $m_W$  of the missing variable:

$$G_W \triangleq \mathbb{E}[\Delta \mid X, Y], \quad m_W \triangleq \mathbb{E}[Z \mid X, Y]$$

Inverse-weighting estimators use  $G_W$  (Horvitz and Thompson, 1952; Binder, 1983; Robins et al., 1994; Van der Laan and Robins, 2003; Hernan and Robins, 2021)

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}_X \mathbb{E}_{Z|X} [Z] \\ &= \mathbb{E}_X \mathbb{E}_{Z|X} \left[ \frac{\mathbb{E}[\Delta|X]}{\mathbb{E}[\Delta|X]} Z \right] \\ &= \mathbb{E}_X \mathbb{E}_{Z|X} \mathbb{E}_{\Delta|X} \left[ \frac{\Delta Z}{\mathbb{E}[\Delta|X]} \right] \\ &= \mathbb{E}_{XZ\Delta} \left[ \frac{\Delta Z}{\mathbb{E}[\Delta|X]} \right] \\ &= \mathbb{E}_{X\Delta Z} \left[ \frac{\Delta Z}{G_W} \right] \\ &= \mathbb{E}_{X\Delta Z} \left[ \frac{\Delta \tilde{Z}}{G_W} \right] \end{aligned} \tag{11}$$

This means we can estimate  $\mathbb{E}[Z]$  provided that (1) ignorability and positivity hold and (2)  $G_W$  is known.  $G_W$  can be estimated by regressing  $\Delta$  on  $X$ . Alternatively, standardization estimators use  $m_W$  (Rubin, 1976; Schafer, 1997; Rubin, 2004; Pearl, 2009; Little and Rubin, 2019; Hernan and Robins, 2021):

$$\mathbb{E}[Z] = \mathbb{E}_X \left[ \mathbb{E}[Z|X] \right] = \mathbb{E}_X \left[ \mathbb{E}[Z|X, \Delta = 1] \right] = \mathbb{E}[m_W] \tag{12}$$

The equality between the middle two terms means that  $m_W$  can be estimated by regressing  $\tilde{Z}$  on  $X$  just on those samples where  $\Delta = 1$ . The equality follows from the ignorability assumption  $Z \perp\!\!\!\perp \Delta \mid X, Y$ .

## Appendix F. DR estimator of mean of Z

The inverse weighting and regression estimators can be combined. Equation (3) defines the DR estimator of  $\mathbb{E}[Z]$  by

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

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

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

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:

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

When  $m$  is correct (regardless of  $G$ ):

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

## Appendix G. Deriving MMD estimators under missingness

### G.1. Deriving the $G_W$ -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.

$$\begin{aligned}
& \mathbb{E}_{P(X|Z=1)P(X'|Z'=1)} [k_{XX'}] \\
&= \int_{X, X'} k P(X|Z=1) P(X'|Z'=1) dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k P(Z=1, X) P(Z'=1, X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k P(Z=1|X) P(Z'=1|X) P(X) P(X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k \mathbb{E}(Z=1|X) \mathbb{E}(Z'=1|X) P(X) P(X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \mathbb{E}_{\substack{X, Z \\ X', Z'}} [k \cdot Z \cdot Z'] \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \mathbb{E}_{\substack{X, Z \\ X', Z'}} \left[ \frac{\mathbb{E}[\Delta|X] \mathbb{E}[\Delta'|X']}{\mathbb{E}[\Delta|X] \mathbb{E}[\Delta'|X']} k \cdot Z \cdot Z' \right] \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \mathbb{E}_{\substack{X, \Delta, Z \\ X', \Delta', Z'}} \left[ \frac{\Delta \Delta'}{G_W G_{W'}} k \cdot Z \cdot Z' \right]
\end{aligned}$$

### G.2. Deriving the $m_W$ -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.

$$\begin{aligned}
& \mathbb{E}_{P(X|Z=1)P(X'|Z'=1)} [k_{XX'}] \\
&= \int_{X, X'} k P(X|Z=1) P(X'|Z'=1) dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k P(Z=1, X) P(Z'=1, X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k P(Z=1|X) P(Z'=1|X) P(X) P(X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \int_{X, X'} k \mathbb{E}(Z=1|X) \mathbb{E}(Z'=1|X) P(X) P(X') dX dX' \\
&= \frac{1}{P(Z=1)} \frac{1}{P(Z'=1)} \mathbb{E}_{X, X'} [m_W \cdot m_{W'} \cdot k]
\end{aligned}$$



### G.3. 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.

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

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.

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

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:

$$\begin{aligned}
 \mathbb{E}_{\substack{X,\Delta,Z \\ X',\Delta',Z'}} \left[ \frac{\Delta\Delta' - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] &= \mathbb{E}_{\substack{X,Z \\ X',Z'}} \left[ \frac{\mathbb{E}[\Delta\Delta'|X, X', Y, Z'] - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] \\
 &= \mathbb{E}_{\substack{X,Z \\ X',Z'}} \left[ \frac{\mathbb{E}[\Delta\Delta'|X, X'] - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] \\
 &= \mathbb{E}_{\substack{X,Z \\ X',Z'}} \left[ \frac{\mathbb{E}[\Delta|X] \mathbb{E}[\Delta'|X'] - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] \\
 &= \mathbb{E}_{\substack{X,Z \\ X',Z'}} \left[ \frac{G_W G_{W'} - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] = 0
 \end{aligned}$$

Likewise, when  $m$  correct:

$$\begin{aligned}
 & \mathbb{E}_{\substack{X, \Delta, Z \\ X', \Delta', Z'}} \left[ \frac{\Delta \Delta' - G_W G_{W'}}{G_W G_{W'}} (ZZ' - m_W m_{W'}) k \right] \\
 &= \mathbb{E}_{\substack{X, \Delta \\ X', \Delta'}} \left[ \frac{\Delta \Delta' - G_W G_{W'}}{G_W G_{W'}} \left( \mathbb{E}[ZZ'|X, X', \Delta, \Delta'] - m_W m_{W'} \right) k \right] \\
 &= \mathbb{E}_{\substack{X, \Delta \\ X', \Delta'}} \left[ \frac{\Delta \Delta' - G_W G_{W'}}{G_W G_{W'}} \left( \mathbb{E}[Z|X, \Delta] \mathbb{E}[Z'|X', \Delta'] - m_W m_{W'} \right) k \right] \\
 &= \mathbb{E}_{\substack{X, \Delta \\ X', \Delta'}} \left[ \frac{\Delta \Delta' - G_W G_{W'}}{G_W G_{W'}} \left( \mathbb{E}[Z|X, \Delta] \mathbb{E}[Z'|X', \Delta'] - \mathbb{E}[Z|X, \Delta = 1] \mathbb{E}[Z'|X', \Delta' = 1] \right) k \right] \\
 &= \mathbb{E}_{\substack{X, \Delta \\ X', \Delta'}} \left[ \frac{\Delta \Delta' - G_W G_{W'}}{G_W G_{W'}} \left( \mathbb{E}[Z|X] \mathbb{E}[Z'|X'] - \mathbb{E}[Z|X] \mathbb{E}[Z'|X'] \right) k \right] = 0
 \end{aligned}$$

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

## Appendix H. 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:

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

This leads to the following estimator:

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

and

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

and

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

The challenging part of applying this estimator is that now instead of one function  $m_W$  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.