Causal Effect Estimation from Observational and Interventional Data Through Matrix Weighted Linear Estimators (Supplementary Material)

Klaus-Rudolf Kladny[1,2] **Julius von Kügelgen**[2,3] **Bernhard Schölkopf**[1,2] **Michael Muehlebach**[2]

[1]Department of Computer Science, ETH Zürich, Switzerland
[2]Max Planck Institute for Intelligent Systems Tübingen, Germany
[3]Department of Engineering, University of Cambridge, United Kingdom

A PROOFS

A.1 PROPOSITION 4.1

Proof. We begin by observing that we can write $\mathbf{W}_{\mathrm{P}}^m$ as

$$\mathbf{W}_{\mathrm{P}}^m = \left(m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} \;+\; \frac{n}{m} n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}\right)^{-1} \left(m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}}\right). \tag{A}$$

We apply the strong law of large numbers to obtain that

$$m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} \xrightarrow{a.s.} \mathbf{Cov}(\mathbf{X}_{\mathrm{I}}) \quad \text{and} \quad n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}} \xrightarrow{a.s.} \mathbf{Cov}(\mathbf{X}_{\mathrm{O}}).$$

Due to the fact that $\lim_{m\to\infty} \frac{n(m)}{m} = c$ for some $c > 0$, we conclude

$$\mathbf{W}_{\mathrm{P}}^m \;\xrightarrow{a.s.}\; \mathbf{W}_\infty \;:=\; \left(\mathbf{Cov}(\mathbf{X}_{\mathrm{I}}) \;+\; c \cdot \mathbf{Cov}(\mathbf{X}_{\mathrm{O}})\right)^{-1} \mathbf{Cov}(\mathbf{X}_{\mathrm{I}}).$$

We observe that

$$(\mathbf{I} - \mathbf{W}_\infty) = \left(\mathbf{Cov}(\mathbf{X}_{\mathrm{I}}) + c \cdot \mathbf{Cov}(\mathbf{X}_{\mathrm{O}})\right)^{-1} c \cdot \mathbf{Cov}(\mathbf{X}_{\mathrm{O}}).$$

Since both covariance matrices are positive definite, so is $\mathbf{Cov}(\mathbf{X}_{\mathrm{I}}) + c \cdot \mathbf{Cov}(\mathbf{X}_{\mathrm{O}})$. We conclude that the smallest singular value of $\mathbf{I} - \mathbf{W}_\infty$ is strictly greater than 0. This means

$$\left|\left|\mathbb{E}[\widehat{\boldsymbol{\alpha}}_{\mathbf{W}_\infty}^m] - \boldsymbol{\alpha}\right|\right|_2^2 \;=\; ||\left(\mathbf{I}_p - \mathbf{W}_\infty\right) \boldsymbol{\Delta}||_2^2 \;\geq\; c'||\boldsymbol{\Delta}||_2^2,$$

for some fixed constant $c' > 0$. We obtain therefore

$$0 < \lim_{m\to\infty} \left|\left|\mathbb{E}[\widehat{\boldsymbol{\alpha}}_{\mathbf{W}_\infty}^m] - \boldsymbol{\alpha}\right|\right|_2^2 \leq \lim_{m\to\infty} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}_\infty}^m\right),$$

where we invoked Jensen's inequality. We see that $\mathbf{W}_\infty$ is constant and bounded. We note that almost sure convergence implies convergence in probability. We can thus apply Lemma B.1, which yields the desired result

$$0 < \lim_{m\to\infty} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}_\infty}^m\right) \leq \lim_{m\to\infty} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}_{\mathrm{P}}^m}^m\right).$$

□

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

A.2 PROPOSITION 4.2

Proposition 4.2. *Let* $\lim_{m\to\infty} \frac{n(m)}{m} = 0$*. Then, it holds that*

$$\lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{P}}\right) = 0.$$

Proof. Similar to the proof of Proposition 4.1, we employ the formulation of (A) and consider the term

$$\frac{n}{m} \cdot n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}.$$

We see that $\lim\limits_{m\to\infty} \frac{n(m)}{m} = 0$ and by the strong law of large numbers, $n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}} \xrightarrow{a.s.} \mathbf{Cov}(\mathbf{X}_{\mathrm{O}})$. Hence, we obtain that

$$\frac{n}{m} \cdot n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}} \xrightarrow{a.s.} \mathbf{0}.$$

By the continuous mapping theorem, we conclude that

$$\mathbf{W}^m_{\mathrm{P}} \xrightarrow{a.s.} \mathbf{I}_p,$$

and by Lemma B.2, this implies that

$$\lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathbf{W}^m_{\mathrm{P}}}\right) \;\leq\; \lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}\right) \;=\; 0.$$

□

A.3 PROPOSITION 4.3

Proof. We rewrite $\widehat{\mathbf{W}}^m_*$ as follows:

$$\widehat{\mathbf{W}}^m_* = \left(n^{-1}\left(n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}\right)^{-1} \hat{\sigma}^2_{Y|X} + \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Delta}}^\top + \epsilon \mathbf{I}_p\right)$$
$$\left(n^{-1}\left(n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}\right)^{-1} \hat{\sigma}^2_{Y|X} + m^{-1}\left(m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}}\right)^{-1} \hat{\sigma}^2_{Y|\mathrm{do}(X)} + \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Delta}}^\top + \epsilon \mathbf{I}_p\right)^{-1},$$

where we insert any almost surely converging estimators for $\boldsymbol{\Delta}$, $\sigma^2_{Y|X}$ and $\sigma^2_{Y|\mathrm{do}(X)}$ instead of their ground-truth values. By almost sure convergence of linear estimators individually, we see that this holds specifically for $\hat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}} - \widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}$. Also, we can use the strong law of large numbers to conclude almost sure convergence of $\hat{\sigma}^2_{Y|X}$ and $\hat{\sigma}^2_{Y|\mathrm{do}(X)}$.

We now show $\widehat{\mathbf{W}}^m_* \xrightarrow{a.s.} \mathbf{I}_p$: First, we see that

$$(cm)^{-1}\left(n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}\right)^{-1} \hat{\sigma}^2_{Y|X} \xrightarrow{a.s.} \mathbf{0} \quad \text{and} \quad m^{-1}\left(m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}}\right)^{-1} \hat{\sigma}^2_{Y|\mathrm{do}(X)} \xrightarrow{a.s.} \mathbf{0},$$

since $m^{-1}\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}}\; \hat{\sigma}^2_{Y|\mathrm{do}(X)}$ and $n^{-1}\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}}\; \hat{\sigma}^2_{Y|X}$ converge almost surely to constants and m^{-1} vanishes. Hence,

$$\widehat{\mathbf{W}}^m_* \xrightarrow{a.s.} \left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top + \epsilon \mathbf{I}_p\right)\left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top + \epsilon \mathbf{I}_p\right)^{-1} = \mathbf{I}_p.$$

□

A.4 THEOREM 4.4

Proof. We have that $\mathbf{I}_p$ is bounded in norm, almost surely. So we can apply Lemma B.2 to see that

$$\lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m_*}\right) \;\leq\; \lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}\right) \;=\; 0.$$

□

A.5 PROPOSITION 4.5

Proof. By Theorem 4.4, it suffices to show that $\widehat{\mathbf{W}}_{\ell^2}^m \xrightarrow{a.s.} \mathbf{I}_p$. Since the other quantities $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m)$, $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n)$ for estimating $\mathbf{W}_*^m$ remain unchanged compared to $\widehat{\mathbf{W}}_*^m$, it suffices to show that the modified computation of $\hat{\boldsymbol{\Delta}}_m$ we call $\hat{\boldsymbol{\Delta}}_m^{\ell^2}$ converges almost surely to the true $\boldsymbol{\Delta} = \boldsymbol{\alpha}_{\mathrm{I}} - \boldsymbol{\alpha}_{\mathrm{O}}$, where $\boldsymbol{\alpha}_{\mathrm{I}}$ and $\boldsymbol{\alpha}_{\mathrm{O}}$ are short-hand for $\mathbb{E}_{\text{int}}[Y|\mathbf{X} = \mathbf{x}]$ and $\mathbb{E}_{\text{obs}}[Y|\mathbf{X} = \mathbf{x}]$, respectively. We observe that $\hat{\boldsymbol{\Delta}}_m^{\ell^2}$ has a closed-form solution

$$
\begin{aligned}
\hat{\boldsymbol{\Delta}}_m^{\ell^2} &= -(\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} + \lambda_{\ell^2}\mathbf{I}_p)^{-1}\mathbf{X}_{\mathrm{I}}^\top(\mathbf{y}_{\mathrm{I}} - \mathbf{X}_{\mathrm{I}}\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n) && \text{(B)}\\
&= (\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} + \lambda_{\ell^2}\mathbf{I}_p)^{-1}\mathbf{X}_{\mathrm{I}}^\top\mathbf{X}_{\mathrm{I}}\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n - (\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} + \lambda_{\ell^2}\mathbf{I}_p)^{-1}\mathbf{X}_{\mathrm{I}}^\top\mathbf{y}_{\mathrm{I}}, && \text{(C)}
\end{aligned}
$$

since $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n$ is again a closed-form solution to an ordinary least squares problem. Considering the first term in (C), we conclude almost sure convergence with respect to $\boldsymbol{\alpha}_{\mathrm{I}}$ (it is simply the ridge regression solution on the interventional data, which is well-known to converge almost surely for fixed λ_{ℓ^2}). The second term satisfies

$$
(\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}} + \lambda_{\ell^2}\mathbf{I}_p)^{-1}\mathbf{X}_{\mathrm{I}}^\top\mathbf{X}_{\mathrm{I}} \xrightarrow{a.s.} \mathbf{I}_p \quad \text{and} \quad \widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n \xrightarrow{a.s.} \boldsymbol{\alpha}_{\mathrm{O}}.
$$

This leads to the desired conclusion. □

B ADDITIONAL LEMMAS

Lemma B.1. *Let $\widehat{\mathbf{W}}^m - \mathbf{W}^m \xrightarrow{P} \mathbf{0}$ [1] and let there exist $c > 0$, $m' \in \mathbb{N}$, such that $||\mathbf{W}^m||_2 \leq c$, for all $m \geq m'$, almost surely. Then, it holds that*

$$
\lim_{m\to\infty} MSE\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m\right) \;\leq\; \lim_{m\to\infty} MSE\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m\right),
$$

where $\xrightarrow{P}$ denotes convergence in probability.

Proof. We derive a lower bound on $\text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m\right)$ by using the formulation

$$
\begin{aligned}
\text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m\right) \;&= \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\}\; ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \boldsymbol{\alpha}||_2^2\right] \;+ \\
&\quad \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right\}\; ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \boldsymbol{\alpha}||_2^2\right], \quad \forall \epsilon > 0.
\end{aligned} \tag{D}
$$

We bound the second summand of (D) from below by zero. For the first summand, we use reverse triangle inequality, which yields

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \boldsymbol{\alpha}||_2^2\right] \;=\; \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m - (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m)||_2^2\right] \\
\geq\;\; & \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m - \boldsymbol{\alpha}||_2^2\right] - 2\sqrt{\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m||_2^2\right]\; \mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m - \boldsymbol{\alpha}||_2^2\right]} + \\
& \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m||_2^2\right] \\
\geq\;\; & \text{MSE}(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m) - \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m - \boldsymbol{\alpha}||_2^2\right] - \\
& 2\sqrt{\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^m - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m||_2^2\right]\; \mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^m - \boldsymbol{\alpha}||_2^2\right]}.
\end{aligned} \tag{E}
$$

For any constant $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{p\times p}$, we rewrite

$$
\begin{aligned}
\mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}'}^m - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}}^m||_2^2\right] \;&=\; \mathbb{E}\left[||(\mathbf{W}' - \mathbf{W})\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m \;+\; (\mathbf{W} - \mathbf{W}')\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n||_2^2\right] \\
&\leq\; 2\left(||\mathbf{W} - \mathbf{W}'||_2^2 \text{Tr}\left(\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m \widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m\,\top}\right]\right) \;+\; ||\mathbf{W} - \mathbf{W}'||_2^2 \text{Tr}\left(\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n \widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n\,\top}\right]\right)\right) \\
&=\; 2||\mathbf{W} - \mathbf{W}'||_2^2\left[\left(||\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m\right]||_2^2 \;+\; \text{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m\right)\right)\right) \;+\; \left(||\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n\right]||_2^2 \;+\; \text{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n\right)\right)\right)\right],
\end{aligned}
$$

[1] We note that $\mathbf{W}^m$ may be random.

where we have used Young's inequality in the first step. We see that both $||\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}\right]||_2^2$ and $||\mathbb{E}\left[\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}\right]||_2^2$ remain bounded $\forall m$, while $\mathrm{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}\right)\right)$ and $\mathrm{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}\right)\right)$ decrease monotonically in m. Hence, we conclude that for any $\epsilon' > 0$, there exists an $\epsilon > 0$ such that

$$\mathbb{E}\left[\left|\left|\widehat{\boldsymbol{\alpha}}_{\mathbf{W}'}^{m} - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}}^{m}\right|\right|_2^2\right] \leq \epsilon',\ \forall m \in \mathbb{N} \text{ and } \forall \mathbf{W}, \mathbf{W}' \in \mathbb{R}^{p\times p} \text{ s.t. } ||\mathbf{W} - \mathbf{W}'||_2 \leq \epsilon. \tag{F}$$

Since $||\mathbf{W}^m||_2 \leq c$ for all $m \geq m'$, we have that $||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2$ is also bounded by some constant $c' > 0$, for all $m \geq m'$, almost surely. We now fix an $\epsilon' > 0$ and choose a corresponding ϵ such that (F) holds. We then conclude from (E) that

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m}\right) \geq\ & \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \boldsymbol{\alpha}||_2^2\right] \\
\geq\ & \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) - 2\sqrt{\epsilon'\,\mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2\right]} - P\left(||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right) c' \\
\geq\ & \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) - 2\sqrt{\epsilon' c'} - P\left(||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right) c',
\end{aligned}$$

for all $m \geq m'$. Thus, we conclude

$$\varliminf_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m}\right) \ \geq\ \varliminf_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) - 2\sqrt{\epsilon' c'}.$$

We can repeat this procedure for any $\epsilon' > 0$ and therefore conclude

$$\varliminf_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m}\right) \ \geq\ \varliminf_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right),$$

which is the desired result. □

Lemma B.2. *Let $\widehat{\mathbf{W}}^m - \mathbf{W}^m \xrightarrow{a.s.} \mathbf{0}$ and let there exist some $c > 0$, $m' \in \mathbb{N}$, such that $||\mathbf{W}^m||_2 \leq c, \forall m \geq m'$, almost surely. Then, it holds that*

$$\varlimsup_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m}\right) \ \leq\ \varlimsup_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right).$$

Proof. We again employ the formulation from (D), but this time to construct an upper bound. For the first term of (D), we see that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\}\ ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \boldsymbol{\alpha}||_2^2\right] \ =\ \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\}\ ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} + \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2\right] \\
\leq\ & \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) \ +\ 2\sqrt{\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}||_2^2\right]\ \mathbb{E}[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2]} + \\
& \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}||_2^2\right],
\end{aligned} \tag{G}$$

by triangle inequality and the Cauchy-Schwarz inequality. Since for $m \geq m'$ it holds that $||\mathbf{W}^m||_2 \leq c$, almost surely, there exists a constant $c' > 0$ such that $\mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2\right] \ \leq\ c'$, for all $m \geq m'$. This is true because the two estimators $\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}$ have both bounded mean squared error for any sample size m.

Analogously to the proof for Lemma B.1, we now fix an $\epsilon' > 0$ and choose a corresponding ϵ such that (F) holds. For $m \geq m'$, we then conclude from (G) that

$$\begin{aligned}
& \mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 \leq \epsilon\right\} ||\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}^m}^{m} - \boldsymbol{\alpha}||_2^2\right] \\
\leq\ & \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) + 2\sqrt{\epsilon'\,\mathbb{E}\left[||\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m} - \boldsymbol{\alpha}||_2^2\right]} + \epsilon' \\
\leq\ & \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{W}^m}^{m}\right) + 2\sqrt{\epsilon' c'} + \epsilon'.
\end{aligned} \tag{H}$$

This bounds the first term of (D). For the second term of (D), we use almost sure convergence of $\widehat{\mathbf{W}}^m - \mathbf{W}^m$. Since $\mathbf{W}^m$ is bounded in the limit, almost surely, so is $\widehat{\mathbf{W}}^m$. Formally, $||\widehat{\mathbf{W}}^m||_2 \leq c'', \forall m \geq m'$ for some $m' \in \mathbb{N}$, almost surely.

We use this to bound $||\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m} - \boldsymbol{\alpha}||_2^2 < c'''$ for all $m \geq m'$, almost surely, for some $c''' > 0$. Now, we apply iterated expectations to the second term of (D) to see that for all $m \geq m'$

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right\}||\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m} - \boldsymbol{\alpha}||_2^2\right] &= \mathbb{E}_{\widehat{\mathbf{W}}^m}\left[\mathbb{1}\left\{||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right\}\mathbb{E}_{\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m}|\widehat{\mathbf{W}}^m}\left[||\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m} - \boldsymbol{\alpha}||_2^2\right]\right] \\
&\leq \mathrm{P}\left(||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right)c''',
\end{aligned} \tag{I}$$

almost surely. Now, we can combine the inequalities (H) and (I) to obtain

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m}\right) \leq \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathbf{W}^m}\right) + 2\sqrt{\epsilon' c'} + \epsilon' + \mathrm{P}\left(||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right)c''',$$

for all $m \geq m''$. Almost sure convergence implies consistency of $\widehat{\mathbf{W}}^m - \mathbf{W}^m$ with respect to $\mathbf{0}$, so we see that $\mathrm{P}\left(||\widehat{\mathbf{W}}^m - \mathbf{W}^m||_2 > \epsilon\right)$ vanishes in the limit $m \to \infty$, for all $\epsilon > 0$. We can repeat this procedure for any $\epsilon' > 0$. This implies the desired result. □

C DETAILED DERIVATION OF OPTIMAL WEIGHTING SCHEMES

In general, we observe that

$$\begin{aligned}
\mathbf{Bias}(\widehat{\boldsymbol{\alpha}}^m_{\mathbf{W}}) &= \mathbf{W}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{W})(\boldsymbol{\alpha} + \boldsymbol{\Delta}) - \boldsymbol{\alpha} = (\mathbf{I} - \mathbf{W})\boldsymbol{\Delta}, \\
\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathbf{W}}) &= \mathbf{W}\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}})\mathbf{W}^\top + (\mathbf{I} - \mathbf{W})\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}})(\mathbf{I} - \mathbf{W})^\top.
\end{aligned}$$

C.1 OPTIMAL SCALAR WEIGHT

Here, we have

$$\begin{aligned}
&\frac{\partial}{\partial w}\mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{w\mathbf{I}_p}\right) \\
=\; &\frac{\partial}{\partial w}\left|\left|\mathbf{Bias}\left(\widehat{\boldsymbol{\alpha}}^m_{w\mathbf{I}_p}\right)\right|\right|_2^2 + \frac{\partial}{\partial w}\mathrm{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}^m_{w\mathbf{I}_p}\right)\right) \\
=\; &-2(1-w)||\boldsymbol{\Delta}||_2^2 + 2w\mathrm{Tr}\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}})\right) - 2(1-w)\mathrm{Tr}\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}})\right) \overset{!}{=} 0.
\end{aligned}$$

By rearranging, we get

$$w^m_* = \frac{\mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}})) + \|\boldsymbol{\Delta}\|_2^2}{\mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}})) + \mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}})) + \|\boldsymbol{\Delta}\|_2^2}.$$

C.2 OPTIMAL DIAGONAL WEIGHT MATRIX

Here, we see that the objective decouples into a sum over the individual dimensions

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{w\mathbf{I}_p}\right) = \sum_{k=1}^{p}\left(1 - w^{(k)}\right)^2\boldsymbol{\Delta}^{(k)\,2} + w^{(k)\,2}\mathbf{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}) + \left(1 - w^{(k)}\right)^2\mathbf{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}).$$

Thus, we optimize for each dimension k separately and obtain

$$w^{m(k)}_* = \frac{\mathrm{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}) + \Delta^{(k)\,2}}{\mathrm{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}) + \mathrm{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}) + \Delta^{(k)\,2}}.$$

C.3 OPTIMAL WEIGHT MATRIX

Using $\frac{\partial}{\partial \mathbf{W}} \mathrm{Tr}(\mathbf{W}\mathbf{A}\mathbf{W}^\top) = 2\mathbf{W}\mathbf{A}$, since $\mathbf{A}$ is symmetric, we observe that

$$\begin{aligned} & \frac{\partial}{\partial \mathbf{W}} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\mathbf{W}}\right) \\ = \;& 2\mathbf{W}\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}) + \mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}) + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right) - 2\left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top + \mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}})\right) \\ \overset{!}{=} \;& \mathbf{0}. \end{aligned}$$

We see that this minimum is attained for

$$\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}) + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right)\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^m_{\mathrm{I}}) + \mathbf{Cov}(\widehat{\boldsymbol{\alpha}}^n_{\mathrm{O}}) + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right)^{-1}.$$

D NON ZERO-MEAN EXOGENOUS VARIABLES

All results established here can readily be extended to settings, where any of the exogenous variables have non-zero mean, i.e., $\boldsymbol{\mu}_{\mathbf{N_X}}, \boldsymbol{\mu}_{\tilde{\mathbf{N}}_\mathbf{X}} := \mathbb{E}[\tilde{\mathbf{N}}_\mathbf{X}], \boldsymbol{\mu}_{\mathbf{N_Z}}, \mu_{N_Y}$ (see (1)–(3)) may be non-zero. In order to extend the practical estimators introduced here, one needs to consider the following two pre-processing steps:

First, we center both treatment distributions separately, without scaling:

$$\mathbf{x}'_i \;\leftarrow\; \mathbf{x}_i - n^{-1} \sum_{j \in 1,\dots,n} \mathbf{x}_j, \qquad \forall i \in 1, ..., n, \tag{J}$$

$$\mathbf{x}'_i \;\leftarrow\; \mathbf{x}_i - m^{-1} \sum_{j \in n+1,\dots,n+m} \mathbf{x}_j, \qquad \forall i \in n+1, ..., n+m. \tag{K}$$

In this manner, both treatment variables become zero-mean.

Furthermore, we add a dummy dimension with value one to all treatment vectors:

$$\mathbf{x}''_i \;\leftarrow\; (\mathbf{x}'_i,\; 1), \quad \forall i \in 1, ..., n+m.$$

This naturally adds one more dimension also to $\boldsymbol{\alpha}$, which corresponds to the intercept term. We then use the constructed $\mathbf{x}''_i$ to compute the weight matrices proposed in this work.

Finally, we see that the intercept term must be identical for both distributions, interventional and observational:

$$\mathbb{E}[Y \mid \mathbf{X}' = \mathbf{x}'] \;=\; \boldsymbol{\gamma}^\top \mathbb{E}[\mathbf{Z} \mid \mathbf{X}' = \mathbf{x}'] + \boldsymbol{\alpha}^\top \mathbf{x}' + \mu_{N_Y}.$$

We then have in the observational setting (data points $1, ..., n$) that

$$\begin{aligned} \boldsymbol{\gamma}^\top \mathbb{E}[\mathbf{Z} \mid \mathbf{X}' = \mathbf{x}'] &= \boldsymbol{\gamma}^\top \boldsymbol{\mu}_{\mathbf{N_Z}} + \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{\mathbf{N_Z}} \mathbf{B}^\top (\boldsymbol{\Sigma}_{\mathbf{N_X}} + \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{N_Z}}\mathbf{B}^\top)^{-1}(\mathbf{x}' - \mathbb{E}[\mathbf{X}']) \\ &= \boldsymbol{\gamma}^\top \boldsymbol{\mu}_{\mathbf{N_Z}} + \boldsymbol{\Delta}^\top \mathbf{x}', \end{aligned}$$

where $\mathbb{E}[\mathbf{X}'] = \mathbf{0}$ due to (J).

For the interventional data, we have independence between $\mathbf{X}'$ and $\mathbf{Z}$ by definition and so we trivially get

$$\boldsymbol{\gamma}^\top \mathbb{E}[\mathbf{Z} \mid \mathbf{X}' = \mathbf{x}'] = \boldsymbol{\gamma}^\top \boldsymbol{\mu}_{\mathbf{N_Z}}$$

here. Thus, the intercept is $\boldsymbol{\gamma}^\top \boldsymbol{\mu}_{\mathbf{N_Z}} + \mu_{N_Y}$ for both distributions and we fix $\hat{\Delta}^{(p+1)} = 0$.

E SAMPLE IMBALANCE

We see that the ground truth covariance matrices of $\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}$ adapt to changes in the sample sizes, keeping the distributions of all variables fixed. For instance, we see that

$$\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}) = (\mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}})^{-1}\sigma_{Y|\mathrm{do}(X)}^{2} = m^{-1}(m^{-1}\mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}})^{-1}\sigma_{Y|\mathrm{do}(X)}^{2}.$$

The term $(m^{-1}\mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}})^{-1}\sigma_{Y|\mathrm{do}(X)}^{2}$ is bounded in probability, for large enough m. Accordingly, this implies that $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}) \xrightarrow{\mathrm{P}} \mathbf{0}$. Thus, when keeping n fixed, we obtain $\mathbf{W}_{*}^{m} \xrightarrow{\mathrm{P}} \mathbf{I}_{p}$, for $m \to \infty$.

On the other hand, if we keep m fixed and consider the limit $n \to \infty$ instead, we observe that

$$\mathbf{W}_{*}^{m} \xrightarrow{\mathrm{P}} \boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}) + \boldsymbol{\Delta}\boldsymbol{\Delta}^{\top})^{-1}.$$

We note that we do not have $\mathbf{W}_{*}^{m} \xrightarrow{\mathrm{P}} \mathbf{0}$ here in general, because the bias in $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}$ remains, independent of the sample size n.