6.1 Generalizing the Simple Illustration

Consider the scenario where the subjects are grouped into \( m = n/2 \) pairs where the two subjects within each pair share the same \( x \). The \( x \) values between each pair \( i \) is evenly spaced controlled by \( \delta \) so that the covariate value in pair \( i \) is \( x = \delta(i - \frac{m+1}{2}) \). The subjects within each pair differ by a \( z \) value denoted \( a \) and \( -a \), identical for all pairs. This structure is the most adversarial if we optimize for observed imbalance.

The optimal restricted procedure will be to match \( m \) pairs: subjects \( \{1, 2\}, \{3, 4\}, \) etc. because this is the bipartite match configuration that provides the smallest imbalance, which will be zero in this example. Comparing this optimal restricted matching procedure with BCRD in terms of the MSE of the simple differences-in-means estimator, we find that the ratio of the standard deviation of the observed covariate to the standard deviation of the unobserved covariate is a critical quantity in this illustration, denoted \( \eta \). Using this equivalent parameterization, \( \delta = \frac{a \sqrt{\eta}}{\sqrt{m^2 - 1}} \eta \).

Tab. 1 shows for both the restricted and BCRD procedures (a) the mean squared observed imbalance as measured by the difference in averages of the two groups i.e. \( (\bar{x}_T - \bar{x}_C)^2 \), (b) the mean squared unobserved imbalance i.e. \( (\bar{z}_T - \bar{z}_C)^2 \) and (c) the MSE of the difference-in-means estimator.
Table 1: Metrics for a general adversarial example for two designs.

<table>
<thead>
<tr>
<th></th>
<th>mean squared observed imbalance</th>
<th>mean squared unobserved imbalance</th>
<th>MSE of the difference-in-means estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>random (BCRD)</td>
<td>( \frac{4a^2\eta^2}{2m-1} )</td>
<td>( \frac{2a^2m}{m} )</td>
<td>( \frac{a^2(\eta^2+1)}{2m-1} )</td>
</tr>
<tr>
<td>restricted (matching)</td>
<td>0</td>
<td>( \frac{2a^2m}{m} )</td>
<td>( \frac{a^2}{m} )</td>
</tr>
</tbody>
</table>

The problem is calibrated so that when \( \eta = 1 \), the observed and unobserved variables carry the same weight in determining the response. Note that the smaller the \( \eta \) (i.e., the less important the observed covariate), the more our estimator favors randomization over matching. In fact, randomization is preferred so long as \( \eta < \sqrt{\frac{m-1}{m}} \). For example, in the case given in Section 1.1, \( m = 4, a = 2 \) and \( \delta = 1 \) implying \( \eta \approx 0.559 < \sqrt{\frac{3}{4}} \approx 0.87 \) and randomization is preferred.

6.2 Derivation of the Conditional MSE

Since the estimator is unbiased the MSE equals the variance. We use the property that \( \text{Var}[\hat{\beta}_T | z, x] = \mathbb{E}[\hat{\beta}_T^2 | z, x] - \beta_T^2 \) where the expectation is taken over \( w \in \mathcal{W}_0 \). Recall that \( \hat{\beta}_T := w^\top \bar{y}/n \) so that

\[
\text{Var}[\hat{\beta}_T | z, x] = \mathbb{E}[(w^\top \bar{y}/n)^2 | z, x] - \beta_T^2 = \frac{1}{n^2} \mathbb{E}[(w^\top (\beta_T w + \beta_x x + z))^2 | z, x] - \beta_T^2.
\]

Note that \( w^\top w = \sum_{i=1}^n w_i^2 = n \), since \( w_i \in \{-1, +1\} \). After canceling out the constant \( \beta_T^2 \), we are left with:

\[
\text{Var}[\hat{\beta}_T | z, x] = \frac{1}{n^2} \mathbb{E}[(2n\beta_T w^\top (\beta_x x + z) + (w^\top (\beta_x x + z))^2 | z, x]].
\]

But \( \mathbb{E}_w[w^\top (\beta_x x + z) | z, x] = 0 \) by the mirror property. Since \( (w^\top (\beta_x x + z))^2 = (\beta_x x + z)^\top w w^\top (\beta_x x + z) \)

\[
\text{Var}[\hat{\beta}_T | z, x] = \frac{1}{n^2} \mathbb{E}[(\beta_x x + z)^\top w w^\top (\beta_x x + z) | z, x] = \frac{1}{n^2}(\beta_x x + z)^\top \Sigma_w(\beta_x x + z).
\]

6.3 Derivation of the Mean MSE

Using the result from Section 6.2,
Thus, it is a sum of scaled non-central \( \chi^2 \) where 
\[
\lambda \Sigma x
\]
and Pedersen (2012, equation 319) when assuming that the fourth moment of 
\( z \) with respect to \( V \) unobserved covariates realizations, i.e. 
\[
\text{where the expectation is taken over all randomizations and the variance is taken over all}
\]
6.5 Derivation of the Variance of the MSE

It is very difficult to optimize the inverse CDF over the space of legal matrices.

Baldessari (1967) proves that this quadratic form is distributed as
\[
(\beta_x + Z)^\top \Sigma_w (\beta_x + Z) \sim \sigma_z^2 \sum_{i=1}^{s} \lambda_i \chi_i^2 \left( \text{Mult} [\lambda_i], \frac{\sigma_z^2}{2} \beta_x^\top v_i v_i^\top x \right)
\]
where \( \lambda_1 \geq \ldots \geq \lambda_s \geq 0 \) and the \( v_1, \ldots, v_s \) are the unique eigenvalues and eigenvectors of \( \Sigma_w \) respectively. \( \text{Mult} [\lambda_i] \) is the multiplicity of the eigenvalue \( \lambda_i \) and \( \chi_i^2 (\nu, \lambda) \) denotes a non-central \( \chi^2 \) random variable with degrees of freedom \( \nu \) and non centrality parameter \( \lambda \). Thus, it is a sum of scaled non-central \( \chi^2 \) random variables.

Since the distribution is parameterized by the eigendecomposition of \( \Sigma_w \), it would be very difficult to optimize the inverse CDF over the space of legal matrices.

6.4 Distribution of the MSE Under Normality

If we assume \( Z \sim \mathcal{N}_n (0_n, \sigma_x^2 I_n) \), then \( \beta_x + Z \sim \mathcal{N}_n (x, \sigma_x^2 I_n) \). We now examine the distribution of the quadratic form, \( (\beta_x + Z)^\top \Sigma_w (\beta_x + Z) \) where \( \Sigma_w \) has properties outlined in the text. Baldessari (1967) proves that this quadratic form is distributed as
\[
(\beta_x + Z)^\top \Sigma_w (\beta_x + Z) \sim \sigma_z^2 \sum_{i=1}^{s} \lambda_i \chi_i^2 \left( \text{Mult} [\lambda_i], \frac{\sigma_z^2}{2} \beta_x^\top v_i v_i^\top x \right)
\]
where \( \lambda_1 \geq \ldots \geq \lambda_s \geq 0 \) and the \( v_1, \ldots, v_s \) are the unique eigenvalues and eigenvectors of \( \Sigma_w \) respectively. \( \text{Mult} [\lambda_i] \) is the multiplicity of the eigenvalue \( \lambda_i \) and \( \chi_i^2 (\nu, \lambda) \) denotes a non-central \( \chi^2 \) random variable with degrees of freedom \( \nu \) and non centrality parameter \( \lambda \). Thus, it is a sum of scaled non-central \( \chi^2 \) random variables.

Since the distribution is parameterized by the eigendecomposition of \( \Sigma_w \), it would be very difficult to optimize the inverse CDF over the space of legal matrices.

6.5 Derivation of the Variance of the MSE

We wish to derive an expression for the variance of the expected squared loss function where the expectation is taken over all randomizations and the variance is taken over all unobserved covariates realizations, i.e. \( \text{Var}_z \left[ \text{MSE}_w [\hat{\beta}_T | z, x] \right] \).

Since Section 6.2 proved that \( \text{MSE}_w [\hat{\beta}_T | z, x] = \frac{1}{n^2} (\beta_x + z)^\top \Sigma_w (\beta_x + z) \), the variance with respect to \( z \) is a variance of a quadratic form, which can be calculated via Petersen and Pedersen (2012, equation 319) when assuming that the fourth moment of \( z \) is finite (Assumption 2.5) and do not depend on \( x \) (Assumption 2.6). Thus we have,
\[
\text{Var}_z \left[ \text{MSE}_w [\hat{\beta}_T | z, x] \right] = \frac{1}{n^4} \left( n \kappa_z + 2 (\sigma_z^2)^2 ||\Sigma_w||_F^2 + 4 \sigma_z^2 \beta_x^\top \Sigma_w^2 x + \gamma_z \beta_x 1_n^\top \Sigma_w x \right),
\]
where \( \kappa_z := \mathbb{E}[z^4] - 3(\sigma^2)^2 \) and \( \gamma_z := \mathbb{E}[z^3] \). Since
\[
\sum_{i,j} \mathbb{E}_w [w_i w_j] x_j = \sum_{j=1}^n x_j \mathbb{E}_w [(w_1 + \ldots + w_n) w_j] = 0,
\]
as \( \sum_{i=1}^n w_i = 0 \), the \( 1_n^\top \Sigma_w x \) term in (1) is zero.

### 6.6 The Order of the Balance Term in Pairwise Matching

Since the distribution of \( x \) is Lipschitz continuous (Assumption 2.7), there exists a bounded increasing function \( G \) such that \( x_i = G(U_i) \) and \( U_1, \ldots, U_n \overset{iid}{\sim} U(0,1) \). Let \( U(1) < U(2) < \cdots < U(n) \) denote the order statistic. Since \( U(2) - U(1), U(4) - U(3), \ldots, U(n) - U(n-1) \overset{iid}{\sim} \exp 1/n \), then
\[
(U(1) - U(2))^2 + (U(3) - U(4))^2 + \cdots + (U(n-1) - U(n))^2 = O(1/n)
\]
almost surely. The Lipschitz condition implies that \( (x_{(i-1)} - x_{(i)})^2 \leq C(U_{(i-1)} - U_{(i)})^2 \) for some constant \( C \) and every \( i \); therefore, \( x^\top \Sigma x = (x_{(1)} - x_{(2)})^2 + (x_{(3)} - x_{(4)})^2 + \cdots + (x_{(n-1)} - x_{(n)})^2 = O(n^{-1}) \) almost surely and thus \( B_1 = O(n^{-3}) \).

### 6.7 Derivation of the Regression Estimator

For the model of equation (1) for arbitrary \( \beta_0 \) and \( \beta_x \), the least squares slope of \( w \) in a regression of \( y \) on \( w \) and \( x \) is
\[
\hat{\beta}_T = \frac{(x^\top x)(w^\top y) - (w^\top x)(x^\top y)}{(x^\top x)(w^\top w) - (w^\top x)^2}.
\]
Since \( w^\top w = n \) and \( (x^\top x) = n \) then,
\[
\hat{\beta}_T = \frac{w^\top y - B_x x^\top y}{n(1 - B_x^2)}
= \frac{(w^\top x + w^\top z + \beta_T w^\top w) - B_x (x^\top x + x^\top z + \beta_T x^\top w)}{n(1 - B_x^2)}
= \frac{(nB_x + nB_z + n\beta_T) - (nB_x + nrB_x + nB_x^2\beta_T)}{n(1 - B_x^2)}
= \frac{B_x + B_z + \beta_T - B_x - rB_x - B_x^2\beta_T}{1 - B_x^2}
= \beta_T + \frac{B_z - rB_x}{1 - B_x^2}
\]
where \( r := x^\top z/n \), a metric that measures dependence between \( x \) and \( z \) (but is not the correlation as it is not properly scaled).
6.8 Derivation of the MSE for the Regression Estimator

In equation (11) we consider the second order approximation to what we found in the previous section:

$$\hat{\beta}_T = \beta_T + (B_z - rB_x)(1 + B_x^2 + O(B_x^4)).$$

We now take the second order approximation of its squared error,

$$\left(\hat{\beta}_T - \beta_T\right)^2 = ((B_z - rB_x)(1 + B_x^2 + O(B_x^4)))^2$$

and simplify as:

$$= (B_z - rB_x)^2 (1 + B_x^4 + O(B_x^4)^2 + 2B_x^2 + 2O(B_x^4) + 2B_x^2O(B_x^4))$$

$$= (B_z - rB_x)^2 (1 + 2B_x^2 + O(B_x^4) + O(B_x^4))$$

$$= (B_z - rB_x)^2 (1 + 2B_x^2 + O(B_x^4))$$

$$= (B_z - rB_x)^2 + 2B_x^2 (B_z - rB_x)^2 + O(B_x^4) (B_z - rB_x)^2$$

$$= (B_z - rB_x)^2 + 2B_x^2 (B_z - rB_x)^2 + O (\max \{B_x^4B_z^2, rB_zB_x^5\}, B_x^6)$$

$$= (B_z - rB_x)^2 + 2B_x^2 (B_z - rB_x)^2 + O (\max \{B_x^4B_z^2, rB_zB_x^5\}, B_x^6)$$

$$= (B_z - rB_x)^2 + 2B_x^2B_z^2 + 4rB_x^3B_z + 2r^2B_x^4 + O (\max \{B_x^4B_z^2, rB_zB_x^5\}, B_x^6)$$

$$= (B_z - rB_x)^2 + 2B_x^2B_z^2 + O (\max \{B_x^4B_z^2, r^2B_x^4, rB_zB_x^5\}, B_x^6)$$

$$\approx (B_z - rB_x)^2 + 2B_x^2B_z^2$$

The $\text{MSE}_w[\hat{\beta}_T | z, x]$ is then the expectation of the above over $\mathbb{P}(w)$. The first term we can simplify as:

$$\mathbb{E}_w[(B_z - rB_x)^2 | x] = \mathbb{E}_w[r^2B_x^2 - 2rB_x + B_x^2 | x]$$

$$= \mathbb{E}_w[r^2(w^T x)/n^2 - 2rw^T x w^T z/n^2 + (w^T z)^2/n^2 | x]$$

$$= \frac{1}{n^2} (r^2x^T \Sigma_w x - 2rx^T \Sigma_w z + z^T \Sigma_w z)$$

$$= \frac{1}{n^2} (-r x + z)^T \Sigma_w (-r x + z)$$

$$= \frac{1}{n^2} (-x x^T z/n + z)^T \Sigma_w (-x x^T z/n + z)$$

$$= \frac{1}{n^2} (-P + I) z^T \Sigma_w (-P + I) z$$

$$= \frac{1}{n^2} z^T (I - P) \Sigma_w (I - P) z$$

$$= \frac{1}{n^2} z^T A z$$

where $P := xx^T/n$, the idempotent projection matrix onto the colsp$[x]$ and $A := (I - P)\Sigma_w(I - P)$. The second term we can simplify as:
\[
2 \mathbb{E}_w [B_x^2 B_z^2 | z, x] = \frac{2}{n^4} \mathbb{E}_w \left[ w^\top x x^\top w w^\top z z^\top w | z, x \right] \\
= \frac{2}{n^4} \mathbb{E}_w \left[ \text{tr} \left[ w^\top x x^\top w w^\top z z^\top w \right] | z, x \right] \\
= \frac{2}{n^4} \mathbb{E}_w \left[ \text{tr} \left[ z^\top w w^\top x x^\top w w^\top z \right] | z, x \right] \\
= \frac{2}{n^4} \mathbb{E}_w \left[ z^\top w w^\top x x^\top w w^\top z | z, x \right] \\
= \frac{2}{n^4} z^\top \mathbb{E}_w \left[ w w^\top x x^\top w w^\top | x \right] z \\
= \frac{2}{n^4} z^\top \mathbb{E}_w \left[ w w^\top P w w^\top | x \right] z
\]

Letting \( D := w w^\top P w w^\top \), a quartic form that cannot be simplified further,

\[
\text{MSE}_w \left[ \hat{\beta}_T | z, x \right] \approx \frac{1}{n^2} \left( \frac{2}{n} D + A \right) z,
\]

6.9 BCRD and CRD are Asymptotically Minimax for the Sup Criterion for the Regression Estimator

To show that CRD and BCRD under the balance property is asymptotically minimax, consider

\[
\gamma(\Sigma_W) = \left( \frac{2}{n} D + (I - P) \Sigma_w (I - P) \right).
\]

Note that \( \sup_{z, \|z\|^2 = 1} z^\top \gamma(\Sigma_W) z = \lambda_{\max}(\gamma(\Sigma_W)) \). We first show that for every design

\[
\lambda_{\max}(\gamma(\Sigma_W)) \geq 1,
\]

where \( \lambda_{\max}(\cdot) \) denotes the largest eigenvalue operator, and for BCRD and CRD

\[
\lambda_{\max}(\gamma(\Sigma_W)) \to 1,
\]

as \( n \to \infty \).

To show (3) we have

\[
\lambda_{\max}(\gamma(\Sigma_W)) \geq^{(1)} \lambda_{\max}((I - P) \Sigma_w (I - P)) \\
\geq^{(2)} \text{tr}[(I - P) \Sigma_w (I - P)] / n \\
=^{(3)} \text{tr}[(I - P) \Sigma_w] / n \\
\geq^{(4)} \text{tr}[\Sigma_w] / n \\
= 1.
\]
Here, inequality (1) is true because $D$ is positive-definite, inequality (2) uses the fact that the maximum is larger than the average, inequality (3) follows from trace properties and the equality $(I - P)^2 = (I - P)$ and finally inequality (4) holds since $P\Sigma_w$ is positive-definite.

We prove that (4) holds for BCRD. The proof that it holds for CRD is simpler and is of the same form. We show first that the term $\frac{2}{n}D$ in equation (2) can be ignored by showing that $z^\top Dz$ is bounded.

Note that $\sum_{i=1}^n x_i = 0$, $\sum_{i=1}^n x_i^2 = n$, and $nD_{ii} = \mathbb{E}_w \left[ \sum_{k,l} w_k w_l x_k x_l \right]$. Therefore, under BCRD, $nD_{ii} = \sum_{i=1}^n x_i^2 - \frac{1}{n-1}A$, where $A = \sum_{i \neq j} x_i x_j$. But $(x_1 + \cdots + x_n)^2 = \sum_{i=1}^n x_i^2 + A = 0$, hence $A = -n$. It follows straightforwardly that $D_{ii} = \frac{n}{n-1}$. Hence, when $\sum_{i=1}^n z_i^2 = 1$, $\sum_{i=1}^n z_i^2 D_{ii} = \frac{n}{n-1}$ which is bounded.

To compute the entries of $D$ for $i \neq j$, it is elementary to show that

$$E(w_i w_j w_k w_\ell) = \begin{cases} \frac{1}{(n-1)(n-3)} & k = i \text{ and } l = j; \text{ or } k = j \text{ and } l = i \\ \frac{1}{n-1} & i \neq j \neq k \neq \ell \\ & \text{otherwise}. \end{cases}$$

Therefore, after simple algebra,

$$D_{ij} = \mathbb{E}_w [w_i w_j P w_j w_j] = \mathbb{E}_w \left[ w_i w_j \sum_{k,\ell} w_k w_\ell x_k x_\ell / n \right]$$

$$= \frac{2x_i x_j}{n} + \frac{2}{n(n-1)}(x_i + x_j)^2 - \frac{1}{n-1} + \frac{3}{n(n-1)(n-3)}(2(x_i + x_j)^2 - n - 2x_i x_j).$$

(5)

Since $|x_i| \leq \sqrt{n}$ and $|x_j| \leq \sqrt{n}$, it is clear that all but the first term in (5) can be ignored in showing that $\sum_{i \neq j} D_{ij} z_i z_j$ is bounded. Finally,

$$\sum_{i \neq j} x_i x_j z_i z_j / n \leq \left( \sum_{i=1}^n x_i z_i / \sqrt{n} \right) \left( \sum_{j=1}^n x_j z_j / \sqrt{n} \right) \leq r_{xz}^2 \leq 1,$$

where $r_{xz}$ is the sample correlation between the $x$ and $z$ values. The second inequality follows since $\sum_{i=1}^n x_i = 0$, $\sum_{i=1}^n x_i^2 = n$, $\sum_{i=1}^n z_i^2 = 1$ and so $\sum_{i=1}^n (z_i - \bar{z})^2 \leq 1$. We conclude that $z^\top Dz$ is bounded and therefore the term $\frac{2}{n}D$ in equation (2) can be ignored.

Recall that for BCRD, $\Sigma_w = \frac{n}{n-1}I - \frac{1}{n-1} \mathbf{1} \mathbf{1}^\top$. Therefore the second term of (2) is

$$\sup_{z, \|z\|^2 = 1} z^\top (I - P) \Sigma_w (I - P) z$$

$$= \sup_{z, \|z\|^2 = 1} \left\{ \frac{n}{n-1} z^\top (I - P) z - \frac{1}{n-1} z^\top (I - P) \mathbf{1} \mathbf{1}^\top (I - P) z \right\}.$$

Because $P \mathbf{1} = \frac{1}{n} \mathbf{x} \mathbf{x}^\top \mathbf{1} = 0$ this term simplifies to
\[
\sup_{z, ||z||^2=1} \left( \frac{n}{n-1} ||z||^2 - \frac{n}{n-1} z^\top P z - \frac{1}{n-1} (z^\top 1)^2 \right)
\]
\[
= \frac{n}{n-1} - \inf_{z, ||z||^2=1} \left( \frac{n}{n-1} z^\top P z + \frac{1}{n-1} (z^\top 1)^2 \right)
\]
\[
= \frac{n}{n-1} - \inf_{z, ||z||^2=1} \left( \frac{1}{n-1} (z^\top x)^2 + \frac{1}{n-1} (z^\top 1)^2 \right)
\]
\[
= \frac{n}{n-1}.
\]

where the last equality is true because the infimum is obtained when \( z \) is orthogonal to both \( x \) and \( 1 \). Since \( \frac{n}{n-1} \to 1 \), equation (4) follows.

### 6.10 Derivation of the Mean Criterion for the Regression Estimator

We find the expectation over \( z \) assuming independence and homoskedasticity (Assumption 2.4) and using Petersen and Pedersen (2012, equation 318),

\[
\mathbb{E}_z \left[ \text{MSE}_{w} \left[ \hat{\beta}_T \mid z, x \right] \right] = \frac{1}{n^2} \mathbb{E}_z \left[ z^\top \left( \frac{2}{n} D + (I - P) \Sigma_w (I - P) \right) z \mid x \right]
\]
\[
= \frac{\sigma_z^2}{n^2} \left[ 2 \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right] \right]
\]
\[
= \frac{\sigma_z^2}{n^2} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]
\[
= \frac{\sigma_z^2}{n} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]
\[
= \frac{\sigma_z^2}{n} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]
\[
= \frac{\sigma_z^2}{n} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]
\[
= \frac{\sigma_z^2}{n} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]
\[
= \frac{\sigma_z^2}{n} \mathbb{E}_w \left[ \text{tr} \left[ \Sigma_w \right] \right]
\]

### 6.11 Derivation of the Standard Error of the MSE for the Regression Estimator

We evaluate the variance of the MSE derived in Appendix 6.8. Using Petersen and Pedersen (2012, equation 319), the variance is
\[ \text{Var}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T \mid z, x \right] \right] = \left( \sigma_z^2 \right)^2 \left\| A + \frac{2}{n} D \right\|_F^2 + \frac{\kappa_z}{n^4} ||a||^2, \]

where \( a \) is the diagonal of \( A + \frac{2}{n} D \) and \( \kappa_z := \mathbb{E}_z [Z^4 \mid x] - 3 (\sigma_z^2)^2 \). Since \( ||a||^2 \) is the sum of the squared entries in the diagonal, it must be less than or equal to the entire Frobenius norm squared and greater than or equal to the negative Frobenius norm squared. Thus,

\[ \text{Var}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T \mid x, z \right] \right] \leq \frac{1}{n^4} \left( \kappa_z - (\sigma_z^2)^2 \right) \left\| A + \frac{2}{n} D \right\|_F^2 \]

and

\[ \text{Var}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T \mid x, z \right] \right] \geq \frac{1}{n^4} \left( (\sigma_z^2)^2 - \kappa_z \right) \left\| A + \frac{2}{n} D \right\|_F^2 \]

thus

\[ \text{Var}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T \mid x, z \right] \right] = \frac{1}{n^4} O(1) \left\| A + \frac{2}{n} D \right\|_F^2. \]

Now we seek to bound the Frobenius norm squared term. We expand it as follows:

\[ \left\| A + \frac{2}{n} D \right\|_F^2 = \sum_{i,j} \left( A_{ij} + \frac{2}{n} D_{ij} \right)^2 = \sum_{i,j} A_{ij}^2 + \frac{4}{n} \sum_{i,j} A_{i,j} D_{i,j} + \frac{4}{n^2} \sum_{i,j} D_{i,j}^2 = \left\| A \right\|_F^2 + \frac{4}{n} \sum_{i,j} A_{i,j} D_{i,j} + \frac{4}{n^2} \left\| D \right\|_F^2. \] (6)

For the first term in equation (6),

\[ \left\| A \right\|_F^2 = \text{tr} \left[ (I - P) \Sigma_w (I - P) (I - P) \Sigma_w (I - P) \right] = \text{tr} \left[ (I - P) \Sigma_w (I - P) \right] = ||\Sigma_w||_F^2 - 2 \text{tr} \left[ P \Sigma_w^2 \right] + \text{tr} \left[ P \Sigma_w P \Sigma_w \right]. \]

Using the inequality \( \text{tr} [AB] \leq \sqrt{\text{tr} [A^2] \text{tr} [B^2]} \) for two positive-definite symmetric matrices \( A \) and \( B \), we find that

\[ \text{tr} \left[ P \Sigma_w^2 \right] \leq \sqrt{\text{tr} \left[ (P \Sigma_w)^2 \right] \text{tr} \left[ \Sigma_w^2 \right]} \leq \sqrt{\text{tr} \left[ P \Sigma_w \right]^2 \text{tr} \left[ \Sigma_w^2 \right]} = \mathcal{B}_x^2 ||\Sigma_w||_F \]

where the second inequality follows from the fact that \( \text{tr} [A^2] \leq \text{tr} [A]^2 \) if \( A \) is positive semi-definite and \( \mathcal{B}_x^2 := \text{tr} \left[ P \Sigma_w \right] = \frac{1}{n} \text{tr} \left[ \Sigma_w \right] = n \mathbb{E}_w [B^2 \mid x] = \mathbb{E}_w \left[ w^\top P w \mid x \right]. \)
Employing the trace inequality again we find,
\[
\text{tr} [P \Sigma_w P \Sigma_w] \leq \sqrt{\text{tr} \left[ (P \Sigma_w)^2 \right]} \text{tr} \left[ (P \Sigma_w)^2 \right] = \text{tr} \left[ (P \Sigma_w)^2 \right] \leq \text{tr} [P \Sigma_w]^2 = B_x^4.
\]

Putting these upper bounds together with analogous lower bounds of their negatives, we find,
\[
||A||_F^2 = ||\Sigma_w||_F^2 + ||\Sigma_w||_F O(B_x^2) + O(B_x^4) \quad \text{and} \quad ||A||_F = ||\Sigma_w||_F + O(B_x^2).
\]

The sum in the second term in equation (6) can be bounded by Cauchy-Schwartz:
\[
\left| \sum_{i,j} A_{i,j} D_{i,j} \right| \leq \sqrt{\left( \sum_{i,j} A_{i,j} \right)^2 \left( \sum_{i,j} D_{i,j} \right)^2} = \sqrt{||A||_F^2 ||D||_F^2} = ||A||_F ||D||_F
\]

To calculate the Frobenius norm of $D$, note that the $i$-th diagonal element of $D$ is equal to $E_w [w_i w_i^T P w | x] = E_w [w_i w | x] = B_x^2$ for all $i$ since $w_i \in \{ -1, +1 \}$. Since $D$ is positive semi-definite, then for each element $D_{ij}$ we have by Cauchy-Schwartz that $|D_{ij}| \leq \sqrt{D_{ii} D_{jj}} = B_x^2$ and thus $D_{ij}^2 \leq B_x^4$ meaning $||D||_F^2 \leq n^2 B_x^4$ and $||D||_F = nB_x^2$.

Putting this upper bounds together with its analogous lower bound of its negatives, we find,
\[
\frac{4}{n} \sum_{i,j} A_{i,j} D_{i,j} = ||A||_F O(B_x^2) = ||\Sigma_w||_F O(B_x^2) + O(B_x^4)
\]

The third term in equation (6) we have already resolved above to be
\[
\frac{4}{n^2} ||D||_F^2 = O(B_x^4)
\]

Aggregating all three terms together and combining like orders, we arrive at
\[
\text{Var}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T | z, x \right] \right] = \frac{1}{n^4} O(1) \left( ||\Sigma_w||_F^2 + ||\Sigma_w||_F O(B_x^2) + O(B_x^4) \right)
\]
and the standard deviation of
\[
\text{SD}_z \left[ \text{MSE}_w \left[ \hat{\beta}_T | z, x \right] \right] = \frac{1}{n^2} O(1) \left( ||\Sigma_w||_F + O(B_x^2) \right).
\]
Replication

All figures and tables can be reproduced by running the R code found at https://github.com/kapelner/harmonizing_designs/paper_duplication.

References
