Supplemental Materials for “Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness”

Timothy B. Armstrong∗
Yale University

Michal Kolesár†
Princeton University

July 20, 2020

D Proofs of auxiliary Lemmas and additional details

D.1 Proof of Lemma A.3

We will show that eq. (30) holds for (a) all $i,j$ with $d_i = d_j = 1 - d$, (b) all $i,j$ with $d_i = 1 - d_j = d$, and for part (ii) that it also holds (c) for all $i,j$ with $d_i = d_j = d$. Let $g_i$ denote the $i$th element of the vector $(g(x_1,d), \ldots, g(x_n,d))^\prime$.

For (a), if eq. (30) didn’t hold for some $i,j$ with $d_i = d_j = 1 - d$, then by the triangle inequality, for all $j'$ with $d_{j'} = d$,

$$g_j + C\|x_i - x_j\|_X < g_i \leq g_{j'} + C\|x_i - x_{j'}\|_X \leq g_j' + C\|x_i - x_j\|_X + C\|x_j - x_{j'}\|_X,$$

contradicting the assertion in both part (i) and part (ii) that eq. (30) holds with equality for at least one $j'$ with $d_{j'} = d$. Similarly, for (c), if it didn’t hold for some $i,j$, then for all $i'$ with $d_{i'} = 1 - d$, by the triangle inequality,

$$g_{i'} \leq g_j + C\|x_{i'} - x_j\|_X < g_i \leq g_{j'} + C\|x_{i'} - x_{j'}\|_X - C\|x_i - x_j\|_X \leq g_{i'} + C\|x_{i'} - x_i\|_X,$$

contradicting the assertion that eq. (30) holds with equality for at least one $i'$ with $d_{i'} = 1 - d$. Finally for (b), if eq. (30) didn’t hold for some $i',j'$ with $d_{i'} = 1 - d_{j'} = d$, then by the triangle inequality, denoting by $j^*(j')$ an element with $d_{j^*} = d$ such that eq. (30) holds with equality when $i = j'$ and $j = j^*$,

$$g_{i'} - g_{j^*(j')} = g_{i'} + C\|x_{j^*(j')} - x_{j'}\|_X - g_{j'} > C\|x_{j^*(j')} - x_{j'}\|_X + C\|x_{i'} - x_{j'}\|_X \geq C\|x_{j^*(j')} - x_{i'}\|_X,$$

which violates (c).

∗email: timothy.armstrong@yale.edu
†email: mkolesar@princeton.edu
D.2 Derivation of algorithm for solution path

Observe that \( \Lambda^0_{ij} = 0 \) unless for some \( k, i \in R_k^0 \) and \( j \in M_k^0 \), and similarly \( \Lambda^1_{ij} = 0 \) unless for some \( k, j \in R_k^1 \) and \( i \in M_k^1 \). Therefore, the first-order conditions for the Lagrangian can be written as

\[
m_j/\sigma^2(0) = \mu w(0) + \sum_{i \in R_k^0} \Lambda^0_{ij} \quad j \in M_k^0, \quad \mu w(1) = \sum_{j \in M_k^0} \Lambda^0_{ij} \quad i \in R_k^0, \tag{S1}
\]

\[
m_i/\sigma^2(1) = \mu w(1) + \sum_{j \in R_k^1} \Lambda^1_{ij} \quad i \in M_k^1, \quad \mu w(0) = \sum_{i \in M_k^1} \Lambda^1_{ij} \quad j \in R_k^1. \tag{S2}
\]

Summing up these conditions then yields

\[
\sum_{j \in M_k^0} m_j/\sigma^2(0) = \mu w(0) \cdot \#M_k^0 + \sum_{j \in M_k^0} \sum_{i \in R_k^0} \Lambda^0_{ij} = \#M_k^0 \cdot \mu w(0) + \#R_k^0 \cdot \mu w(1),
\]

\[
\sum_{i \in M_k^1} m_i/\sigma^2(1) = \mu w(1) \cdot \#M_k^1 + \sum_{i \in M_k^1} \sum_{j \in R_k^1} \Lambda^1_{ij} = \#M_k^1 \cdot \mu w(1) + \#R_k^1 \cdot \mu w(0).
\]

Following the argument in Osborne et al. (2000, Section 4), by continuity of the solution path, for a small enough perturbation \( s \), \( N^d(\mu + s) = N^d(\mu) \), so long as the elements of \( \Lambda^d(\mu) \) associated with the active constraints are strictly positive. In other words, the set of active constraints doesn’t change for small enough changes in \( \mu \). Hence, the partition \( M_k^d \) remains the same for small enough changes in \( \mu \) and the solution path is differentiable. Differentiating the preceding display yields

\[
\frac{1}{\sigma^2(0)} \sum_{j \in M_k^0} \frac{\partial m_j(\mu)}{\partial \mu} = \#M_k^0 \cdot w(0) + \#R_k^0 \cdot w(1),
\]

\[
\frac{1}{\sigma^2(1)} \sum_{i \in M_k^1} \frac{\partial m_i(\mu)}{\partial \mu} = \#M_k^1 \cdot w(1) + \#R_k^1 \cdot w(0).
\]

If \( j \in M_k^0 \), then there exists a \( j' \) and \( i \) such that the constraints associated with \( \Lambda^0_{ij} \) and \( \Lambda^0_{ij'} \) are both active, so that \( m_j + \|x_i - x_j\|_{\mathcal{X}} = r_i = m_{j'} + \|x_i - x_j\|_{\mathcal{X}} \), which implies that \( \partial m_j(\mu)/\partial \mu = \partial m_{j'}(\mu)/\partial \mu \). Since all elements in \( M_k^0 \) are connected, it follows that the derivative \( \partial m_j(\mu)/\partial \mu \) is the same for all \( j \) in \( M_k^0 \). Similarly, \( \partial m_j(\mu)/\partial \mu \) is the same for all \( j \) in \( M_k^1 \). Combining these observations with the preceding display implies

\[
\frac{1}{\sigma^2(0)} \frac{\partial m_j(\mu)}{\partial \mu} = w(0) + \frac{\#R_k^0(j)}{\#M_k^0(j)} w(1), \quad \frac{1}{\sigma^2(1)} \frac{\partial m_i(\mu)}{\partial \mu} = w(1) + \frac{\#R_k^1(i)}{\#M_k^1(i)} w(0),
\]

where \( k(i) \) and \( k(j) \) are the partitions that \( i \) and \( j \) belong to. Differentiating the first-order conditions (S1) and (S2) and combining them with the restriction that \( \partial \Lambda^d_{ij}(\mu)/\partial \mu = 0 \) if \( N^d_{ij}(\mu) = 0 \)
then yields the following set of linear equations for $\partial \Lambda^d(\mu)/\partial \mu$:

\[
\frac{\# \mathcal{R}_k^0}{\# \mathcal{M}_k^0} w(1) = \sum_{i \in \mathcal{R}_k^0} \frac{\partial \Lambda^d_{ij}(\mu)}{\partial \mu}, \quad w(1) = \sum_{j \in \mathcal{M}_k^0} \frac{\partial \Lambda^d_{ij}(\mu)}{\partial \mu},
\]

\[
\frac{\# \mathcal{R}_k^1}{\# \mathcal{M}_k^1} w(0) = \sum_{j \in \mathcal{R}_k^1} \frac{\partial \Lambda^d_{ij}(\mu)}{\partial \mu}, \quad w(0) = \sum_{i \in \mathcal{M}_k^1} \frac{\partial \Lambda^d_{ij}(\mu)}{\partial \mu}, \quad \frac{\partial \Lambda^d_{ij}(\mu)}{\partial \mu} = 0 \text{ if } N^d_{ij}(\mu) = 0.
\]

Therefore, $m(\mu)$, $\Lambda^0(\mu)$, and $\Lambda^1(\mu)$ are all piecewise linear in $\mu$. Furthermore, since for $i \in \mathcal{R}_k^0$, $r_i(\mu) = m_j(\mu) + \|x_i - x_j\|_\chi$ where $j \in \mathcal{M}_k^0$, it follows that

\[
\frac{\partial r_i(\mu)}{\partial \mu} = \frac{\partial m_j(\mu)}{\partial \mu} = \sigma^2(0) \left[ w(0) + \frac{\# \mathcal{R}_k^0}{\# \mathcal{M}_k^0} w(1) \right].
\]

Similarly, since for $j \in \mathcal{R}_k^1$, and $i \in \mathcal{M}_k^1$ $r_j(\mu) = m_i(\mu) + \|x_i - x_j\|_\chi$, where $j \in \mathcal{M}_k^1$, we have

\[
\frac{\partial r_j(\mu)}{\partial \mu} = \frac{\partial m_i(\mu)}{\partial \mu} = \sigma^2(1) \left[ w(1) + \frac{\# \mathcal{R}_k^1}{\# \mathcal{M}_k^1} w(0) \right].
\]

Thus, $r(\mu)$ is also piecewise linear in $\mu$.

Differentiability of $m$ and $\Lambda^d$ is violated if the condition that the elements of $\Lambda^d(\mu)$ associated with the active constraints are all strictly positive is violated. This happens if one of the non-zero elements of $\Lambda^d(\mu)$ decreases to zero, or else if a non-active constraint becomes active, so that for some $i$ and $j$ with $N^0_{ij}(\mu) = 0$, $r_i(\mu) = m_j(\mu) + \|x_i - x_j\|_\chi$, or for some $i$ and $j$ with $N^1_{ij}(\mu) = 0$, $r_j(\mu) = m_i(\mu) + \|x_i - x_j\|_\chi$. This determines the step size $s$ in the algorithm.

### D.3 Proof of Lemma B.2

For ease of notation, let $f_i = f(x_i, d_i)$, $\sigma^2_i = \sigma^2(x_i, d_i)$, and let $\bar{f}_i = J^{-1}\sum_{j=1}^{J} f_{ij}(i)$ and $\bar{u}_i = J^{-1}\sum_{j=1}^{J} u_{\ell_j(i)}$. Then we can decompose

\[
\frac{J + 1}{J} (u_i^2 - u_i^2) = [f_i - \bar{f}_i + u_i - \bar{u}_i]^2 - \frac{J + 1}{J} u_i^2
\]

\[
= [(f_i - \bar{f}_i)^2 + 2(u_i - \bar{u}_i)(f_i - \bar{f}_i)] - 2\bar{u}_i u_i + \frac{2}{J^2} \sum_{j=1}^{J} \sum_{k=1}^{J} u_{\ell_j(i)} u_{\ell_k(i)} + \frac{1}{J^2} \sum_{j=1}^{J} (u_{\ell_j(i)}^2 - u_i^2)
\]

\[
= T_{1i} + 2T_{2i} + 2T_{3i} + T_{4i} + T_{5i} + \frac{1}{J^2} \sum_{j=1}^{J} (\sigma^2_{\ell_j(i)} - \sigma^2_i),
\]
where

\[ T_{1i} = [(f_i - \bar{f}_i)^2 + 2(u_i - \bar{u}_i)(f_i - \bar{f}_i)], \quad T_{2i} = \bar{u}_i u_i \]

\[ T_{3i} = \frac{1}{J^2} \sum_{j=1}^{J} \sum_{k=1}^{J-1} u_{\ell_j(i)} u_{k(i)}, \quad T_{4i} = \frac{1}{J^2} \sum_{j=1}^{J} (u_{\ell_j(i)}^2 - \sigma_{\ell_j(i)}^2), \quad T_{5i} = \sigma_i^2 - u_i^2. \]

Since \( \max_i \|x_{\ell_j(i)} - x_i\| \to 0 \) and since \( \sigma^2(\cdot, d) \) is uniformly continuous, it follows that

\[ \max_i \max_{1 \leq j \leq J} |\sigma_{\ell_j(i)}^2 - \sigma_i^2| \to 0, \]

and hence that \( |\sum_{i=1}^{n} a_{ni} J^{-1} \sum_{j=1}^{J} (\sigma_{\ell_j(i)}^2 - \sigma_i^2)| \leq \max_i \max_{j=1}^J (\sigma_{\ell_j(i)}^2 - \sigma_i^2) \sum_{i=1}^{n} a_{ni} \to 0. \) To prove the lemma, it therefore suffices to show that the sums \( \sum_{i=1}^{n} a_{ni} T_{qi} \) all converge to zero.

To that end,

\[ E|\sum_{i} a_{ni} T_{1i}| \leq \max_{i} (f_i - \bar{f}_i)^2 \sum_{i} a_{ni} + 2 \max_{i} |f_i - \bar{f}_i| \sum_{i} a_{ni} E|u_i - \bar{u}_i|, \]

which converges to zero since \( \max_{i} |f_i - \bar{f}_i| \leq \max_{i} \max_{j=1}^J (f_i - f_{\ell_j(i)}) \leq C_n \max_{i} \|x_i - x_{\ell_j(i)}\| x \to 0. \) Next, by the von Bahr-Esseen inequality,

\[ E|\sum_{i=1}^{n} a_{ni} T_{5i}|^{1+1/2K} \leq 2 \sum_{i=1}^{n} a_{ni}^{1+1/2K} E|T_{5i}|^{1+1/2K} \leq 2 \max_{i} a_{ni}^{1/2K} \max_{j} E|T_{5j}|^{1+1/2K} \sum_{k=1}^{n} a_{nk} \to 0. \]

Let \( \mathcal{I}_j \) denote the set of observations for which an observation \( j \) is used as a match. To show that the remaining terms converge to zero, let use the fact \( \#\mathcal{I}_j \) is bounded by \( JL \), where \( L \) is the kissing number, defined as the maximum number of non-overlapping unit balls that can be arranged such that they each touch a common unit ball (Miller et al., 1997, Lemma 3.2.1; see also Abadie and Imbens, 2008). \( L \) is a finite constant that depends only on the dimension of the covariates (for example, \( L = 2 \) if \( \text{dim}(x_i) = 1 \)). Now,

\[ \sum_{i} a_{ni} T_{4i} = \frac{1}{J^2} \sum_{j=1}^{n} (u_j - \sigma_j^2) \sum_{i \in \mathcal{I}_j} a_{ni}, \]

and so by the von Bahr-Esseen inequality,

\[ E|\sum_{i} a_{ni} T_{4i}|^{1+1/2K} \leq \frac{2}{J^{2+1/K}} \sum_{j=1}^{n} E|u_j - \sigma_j^2|^{1+1/2K} \left( \sum_{i \in \mathcal{I}_j} a_{ni} \right)^{1+1/2K} \]

\[ \leq \frac{(JL)^{1/2K}}{J^{2+1/K}} \max_{k} E|u_k - \sigma_k^2|^{1+1/2K} \max_{i} a_{ni}^{1+1/2K} \sum_{j \in \mathcal{I}_j} \sum_{i} a_{ni}, \]

which is bounded by a constant times \( \max_{i} a_{ni}^{1+1/2K} \sum_{j=1}^{n} \sum_{i \in \mathcal{I}_j} a_{ni} = \max_{i} a_{ni}^{1+1/2K} J \sum_{i} a_{ni} \to 0. \)
Next, since \( E[u_i u'_i u_{\ell_j(i)} u_{\ell_k(i')}] \) is non-zero only if either \( i = i' \) and \( \ell_j(i) = \ell_k(i') \), or else if \( i = \ell_k(i') \) and \( i' = \ell_j(i) \), we have \( \sum_{i'=1}^n a_{nn'} E[u_i u_{\ell_k(i')} u_{\ell_j(i)} u_{\ell_k(i')}] \leq \max_{i'} a_{nn'} \left( \sigma^2_{\ell_j(i)} + \sigma^2_{\ell_k(i')} \right) \), so that

\[
\text{var} \left( \sum_i a_{nn} T_{2i} \right) = \frac{1}{J^2} \sum_{i,j,k,i'} a_{nn'} \text{var} \left( u_i u_{\ell_k(i')} u_{\ell_j(i)} u_{\ell_k(i')} \right) \leq 2K^2 \max_{i'} a_{nn'} \sum_i a_{ni} \to 0.
\]

Similarly for \( j \neq k \) and \( j' \neq k \), \( \sum_{i'=1}^n a_{nn'} E[u_{\ell_j(i)} u_{\ell_k(i)} u_{\ell_j(i')} u_{\ell_k(i')} \leq \max_{i'} 2\sigma^2_{\ell_j(i)} \sigma^2_{\ell_k(i)} \), so that

\[
\text{var} \left( \sum_i a_{nn} T_{3i} \right) = \frac{1}{J^4} \sum_{i,j,j',k,k'=1} a_{nn} a_{nn'} a_{nn'} \text{var} \left( u_i u_{\ell_k(i)} u_{\ell_j(i')} u_{\ell_k(i')} \right) \leq 2K^2 \max_{i'} a_{nn'} \sum_i a_{ni} \to 0.
\]

D.4 Standard errors for PATE

We now consider construction of the standard error \( \text{se}_r(\hat{L}_k) \). For matching estimators with a fixed number of matches, standard errors for the PATE are available, for example, in Abadie and Imbens (2006). For completeness, we provide a generic formulation and consistency result that applies to arbitrary estimators \( \hat{L}_k \) in our setting.

In Theorems 4.2 and 4.3, we gave conditions under which the conditional standard error \( \text{se}(\hat{L}_k) \) is consistent in the sense that \( \text{se}(\hat{L}_k) / \sum_{i=1}^n k(X_i, D_i) \sigma^2_P(X_i, D_i) \) converges in probability to one conditional on \( \{X_i, D_i\}_{i=1}^n \), along with conditions on the marginal distribution of \( (X_i, D_i) \) such that this holds for \( \{X_i, D_i\}_{i=1}^\infty \) in a probability one set. This implies that \( \text{se}(\hat{L}_k) / \sum_{i=1}^n k(X_i, D_i) \sigma^2_P(X_i, D_i) \) converges in probability to one unconditionally under these conditions. Thus, if Assumption B.1 holds as well, \( \text{se}(\hat{L}_k)^2 / V_{1,n}(P) \) will converge in probability to one.

Thus, it suffices to estimate \( nV_{2,n}(P) = E_P([f_P(X_i, 1) - f(X_i, 0) - \tau(P)]^2) \). Abadie and Imbens (2006, Theorem 7) give consistency conditions for the matching estimator described in the text. We therefore focus on the estimator \( n\hat{V}_2 = \frac{1}{n} \sum_{i=1}^n (\hat{f}(X_i, 1) - \hat{f}(X_i, 0))^2 - \hat{L}_k^2 \).

**Theorem D.1.** Suppose that \( \max_{1 \leq i \leq n, d \in \{0,1\}} |\hat{f}(X_i, d) - f_P(X_i, d)| \overset{P}{\to} 0 \) and \( \hat{L}_k \overset{P}{\to} \tau(P) \) uniformly over \( P \in \mathcal{P} \), and that Assumption B.1 holds, with \( n[V_{1,n}(P) + V_{2,n}(P)] \) bounded away from zero uniformly over \( P \in \mathcal{P} \). Let \( \hat{V}_{2,n} \) be given above. Then \( [\hat{V}_{2,n} - V_{2,n}(P)]/[V_{1,n}(P) + V_{2,n}(P)] \) converges in probability to zero uniformly over \( P \in \mathcal{P} \). Furthermore, if \( \text{se}_r(\hat{L}_k)^2 = \text{se}(\hat{L}_k)^2 + \hat{V}_{2,n} \) where \( \text{se}(\hat{L}_k)^2 / V_{1,n}(P) \) converges in probability to one uniformly over \( P \in \mathcal{P} \), then \( [V_{1,n}(P) + V_{2,n}(P)] / \text{se}_r(\hat{L}_k)^2 \overset{P}{\to} 1 \) uniformly over \( P \in \mathcal{P} \).
Proof. We have

\[
\left| \frac{\hat{V}_{2,n}}{n} - \frac{V_{2,n}(P)}{n} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \left[ \hat{f}(X_i, 1) - \hat{f}(X_i, 0) \right]^2 - \left[ f_P(X_i, 1) - f_P(X_i, 0) \right]^2 \right\} + \tau(P)^2 - \hat{L}_k^2 \right|
\]

\leq 2 \max_{1 \leq i \leq n, d \in \{0,1\}} \left| \hat{f}(X_i, d) - f_P(X_i, d) \right|^2 + \left| \hat{L}_k^2 - \tau(P)^2 \right|

which converges in probability to zero uniformly over \( P \in \mathcal{P} \). By the \( O(1/n) \) lower bound on \( V_{1,n}(P) + V_{2,n}(P) \), it then follows that \( [\hat{V}_{2,n} - V_{2,n}(P)]/[V_{1,n}(P) + V_{2,n}(P)] \) converges in probability to zero uniformly over \( P \in \mathcal{P} \).

References


