Supplement to "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness"

Timothy B. Armstrong* Yale University

Michal Kolesár† Princeton University

January 18, 2021

These supplemental materials are organized as follows. Supplemental Appendix C gives additional empirical results. Supplemental Appendix D proves Lemma A.3, gives the derivation of the solution path in the proof of Theorem 2.2, completes the proof of Theorem 2.3, proves Lemma B.1 and Lemma B.2, gives conditions for asymptotic efficiency of the matching estimator with a single match, and finally verifies Assumption B.1 for the matching estimator.

C Additional empirical results: Other choices of distance

A disadvantage of the distance based on $A = A_{\text{main}}$ is that it requires prior knowledge of the relative importance of different pretreatment variables in explaining the outcome variable. An alternative is to specify the distance using moments of the pretreatment variables in a way that ensures invariance to scale transformations. For example, Abadie and Imbens (2011) form matching estimators using the weighted Euclidean norm (so $q = 2$) with $A = A_{\text{ne}} \equiv \text{diag}(1/\text{std}(x_1), \ldots, 1/\text{std}(x_p))$, where std denotes sample standard deviation. Table S1 shows the diagonal elements of A_{ne}. It can be seen that this distance is most likely not the best way of encoding a researcher's prior beliefs about Lipschitz constraints. For example, the bound on the difference in average earnings between blacks and non-black non-Hispanics is substantially smaller than the bound on the difference in average earnings between Hispanics and non-black non-Hispanics.

If the constant C is to be chosen conservatively, the derivative of $f(x, d)$ with respect to each of these variables must be bounded by C times the corresponding element in this table. If one allows for somewhat persistent earnings, then C should be chosen in the range of 10 or above: to allow previous years' earnings to have a one-to-one effect, we would need to take $C = 1/\sqrt{.07^2 + .07^2} = 10.1$. For this C, when δ is chosen to optimize confidence interval (CI) length, the resulting CI is given by 1.72 ± 7.63, which is much wider than the CIs reported in Table 2.

*email: timothy.armstrong@yale.edu
†email: mkolesar@princeton.edu

Table S1: Diagonal elements of the weight matrix A in definition of the norm in eq. (24) for the main specification, A_{main}, and alternative specification, A_{ne}.

						Earnings		Employed	
	Age	Educ.	Black	Hispanic	Married	1974	1975	1974	1975
A_{main}	0.15	0.60	2.50	2.50	2.50	0.50	0.50	0.10	0.10
A_{ne}	0.10	0.33	2.20	5.49	2.60	0.07	0.07	2.98	2.93

In Theorem 2.3, we showed that the matching estimator with a single match is optimal for C large enough. For this result, it is important that the norm used to construct the matches is the same as the norm defining the Lipschitz class. To illustrate this point, consider a matching estimator considered in Abadie and Imbens (2011), that uses $q = 2$ and $A = A_{\text{ne}}$. The root mean squared error (RMSE) efficiency of this estimator under our main specification (A_{main}, $q = 1$ and $C = 1$) is 77.5%; for CI length, its efficiency is 74.6%. This is considerably lower than the efficiencies of the matching estimator that matched on the norm defining the Lipschitz class reported in Section 5.2. Furthermore, the efficiency is never higher than 80.1%, even for large values of C.

D Proofs of auxiliary Lemmas and additional details

D.1 Proof of Lemma A.3

We will show that eq. (29) 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 ith element of the vector $(g(x_1, d), \ldots, g(x_n, d))'$.

For (a), if eq. (29) didn't hold for some i, j with $d_i = d_j = 1 - d$, then by the triangle inequality, for all j' with $d_{j'} = d$,

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

contradicting the assertion in both part (i) and part (ii) that eq. (29) 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,

$$g_{i'} \leq g_j + C\|x_{i'} - x_j\|_{\mathcal{X}} < g_i + C\|x_{i'} - x_j\|_{\mathcal{X}} - C\|x_i - x_j\|_{\mathcal{X}} \leq g_i + C\|x_{i'} - x_i\|_{\mathcal{X}},$$

contradicting the assertion that eq. (29) holds with equality for at least one i' with $d_{i'} = 1 - d$. Finally, for (b), if eq. (29) 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. (29) holds with equality when $i = j'$ and $j = j^*$,

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

which violates (c).

D.2 Derivation of algorithm for solution path

Observe that $\Lambda^0_{ij} = 0$ unless for some k, $i \in \mathcal{R}^0_k$ and $j \in \mathcal{M}^0_k$, and similarly $\Lambda^1_{ij} = 0$ unless for some k, $j \in \mathcal{R}^1_k$ and $i \in \mathcal{M}^1_k$. Therefore, the first-order conditions for the Lagrangian can be written as

$$m_j/\sigma^2(0) = \mu w(0) + \sum_{i \in \mathcal{R}^0_k} \Lambda^0_{ij} \qquad j \in \mathcal{M}^0_k, \qquad \mu w(1) = \sum_{j \in \mathcal{M}^0_k} \Lambda^0_{ij} \qquad i \in \mathcal{R}^0_k, \tag{S1}$$

$$m_i/\sigma^2(1) = \mu w(1) + \sum_{j \in \mathcal{R}^1_k} \Lambda^1_{ij} \qquad i \in \mathcal{M}^1_k, \qquad \mu w(0) = \sum_{i \in \mathcal{M}^1_k} \Lambda^1_{ij} \qquad j \in \mathcal{R}^1_k. \tag{S2}$$

Summing up these conditions then yields

$$\sum_{j \in \mathcal{M}^0_k} m_j/\sigma^2(0) = \mu w(0) \cdot \#\mathcal{M}^0_k + \sum_{j \in \mathcal{M}^0_k} \sum_{i \in \mathcal{R}^0_k} \Lambda^0_{ij} = \#\mathcal{M}^0_k \cdot \mu w(0) + \#\mathcal{R}^0_k \cdot \mu w(1),$$
$$\sum_{i \in \mathcal{M}^1_k} m_i/\sigma^2(1) = \mu w(1) \cdot \#\mathcal{M}^1_k + \sum_{i \in \mathcal{M}^1_k} \sum_{j \in \mathcal{R}^1_k} \Lambda^1_{ij} = \#\mathcal{M}^1_k \cdot \mu w(1) + \#\mathcal{R}^1_k \cdot \mu w(0).$$

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 μ. Hence, the partition $\mathcal{M}^d_k$ remains the same for small enough changes in μ and the solution path is differentiable. Differentiating the preceding display yields

$$\frac{1}{\sigma^2(0)} \sum_{j \in \mathcal{M}^0_k} \frac{\partial m_j(\mu)}{\partial \mu} = \#\mathcal{M}^0_k \cdot w(0) + \#\mathcal{R}^0_k \cdot w(1),$$
$$\frac{1}{\sigma^2(1)} \sum_{i \in \mathcal{M}^1_k} \frac{\partial m_i(\mu)}{\partial \mu} = \#\mathcal{M}^1_k \cdot w(1) + \#\mathcal{R}^1_k \cdot w(0).$$

If $j \in \mathcal{M}^0_k$, then there exists a j' and i such that the constraints associated with Λ^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 $\mathcal{M}^0_k$ are connected, it follows that the derivative $\partial m_j(\mu)/\partial \mu$ is the same for all j in $\mathcal{M}^0_k$. Similarly, $\partial m_j(\mu)/\partial \mu$ is the same for all j in $\mathcal{M}^1_k$. Combining these observations with the preceding display implies

$$\frac{1}{\sigma^2(0)} \frac{\partial m_j(\mu)}{\partial \mu} = w(0) + \frac{\#\mathcal{R}^0_{k(j)}}{\#\mathcal{M}^0_{k(j)}} w(1), \qquad \frac{1}{\sigma^2(1)} \frac{\partial m_i(\mu)}{\partial \mu} = w(1) + \frac{\#\mathcal{R}^1_{k(i)}}{\#\mathcal{M}^1_{k(i)}} w(0),$$

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 $\partial\Lambda^d(\mu)/\partial\mu$:

$$\frac{\#\mathcal{R}_k^0}{\#\mathcal{M}_k^0}w(1) = \sum_{i\in\mathcal{R}_k^0}\frac{\partial\Lambda_{ij}^0(\mu)}{\partial\mu}, \qquad w(1) = \sum_{j\in\mathcal{M}_k^0}\frac{\partial\Lambda_{ij}^0(\mu)}{\partial\mu},$$
$$\frac{\#\mathcal{R}_k^1}{\#\mathcal{M}_k^1}w(0) = \sum_{j\in\mathcal{R}_k^1}\frac{\partial\Lambda_{ij}^1(\mu)}{\partial\mu}, \qquad w(0) = \sum_{i\in\mathcal{M}_k^1}\frac{\partial\Lambda_{ij}^1(\mu)}{\partial\mu}, \qquad \frac{\partial\Lambda_{ij}^d(\mu)}{\partial\mu} = 0 \qquad \text{if } N_{ij}^d(\mu) = 0.$$

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

Differentiability of m and Λ^d is violated if the condition that the elements of Λ^d 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 $N_{ij}^0(\mu) = 0$, $r_i(\mu) = m_j(\mu) + \|x_i - x_j\|_{\mathcal{X}}$, or for some i and j with $N_{ij}^1(\mu) = 0$, $r_j(\mu) = m_i(\mu) + \|x_i - x_j\|_{\mathcal{X}}$. This determines the step size s in the algorithm.

D.3 Bounds on optimal δ for Theorem 2.3

Theorem 2.3 follows from Theorem A.5 so long as the optimal δ for the fixed-length confidence interval (FLCI) and RMSE criteria do not increase without bound as C increases. This section shows that this is indeed the case.

Let $S(\delta, C) = \text{sd}(\hat{L}_\delta)$ and let $B(\delta, C) = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta)$ denote standard deviation and worst-case bias when $\mathcal{F}$ is given by the Lipschitz class with constant C, and $\hat{L}_\delta$ is computed with this class. Let $\mathcal{A}(C)$ denote the feasible set of worst-case bias and standard deviation pairs for this problem. Note that the set $\mathcal{A}(C)$ is convex. In particular, given estimators $\hat{L}_1$ and $\hat{L}_2$ with worst-case bias B_1, B_2 and standard deviation S_1, S_2, the estimator $\lambda\hat{L}_1 + (1-\lambda)\hat{L}_2$ has worst-case bias bounded by $\lambda B_1 + (1-\lambda)B_2$ and standard deviation bounded by $\lambda S_1 + (1-\lambda)S_2$, which then allows for the construction of an affine estimator with worst-case bias and standard deviation exactly equal to these quantities by adding a nonrandom constant and a multiple of a $\mathcal{N}(0,1)$ variable independent of the observed data (adding a $\mathcal{N}(0,1)$ variable to the sample will not change the calculations for the optimal estimator for RMSE or FLCI length).

Let $R(B, S)$ be the RMSE criterion ($R(B, S) = \sqrt{B^2 + S^2}$) or the FLCI length criterion

$(R(B,S) = \text{cv}_\alpha(B/S)S)$. Let $\delta^* = \delta^*(C)$ minimize $R(B(\delta,C), S(\delta,C))$. Then $B(\delta^*,C), S(\delta^*,C)$ optimizes $R(B,S)$ over the feasible set $\mathcal{A}(C)$. Let $\delta \neq \delta^*$ be given. By convexity of the feasible set $\mathcal{A}(C)$, we have, for all $t \in [0,1]$,

$$R((B(\delta,C) - B(\delta^*,C))t + B(\delta^*,C), (S(\delta,C) - S(\delta^*,C))t + S(\delta^*,C)) - R(B(\delta^*,C), S(\delta^*,C)) \geq 0.$$

Dividing both sides by t and taking the limit as $t \to 0$, we obtain

$$R_1^*(C)[B(\delta,C) - B(\delta^*,C)] + R_2^*(C)[S(\delta,C) - S(\delta^*,C)] \geq 0,$$

where $(R_1^*(C), R_2^*(C))$ is the derivative of $R(B,S)$ at $(B(\delta^*,C), S(\delta^*,C))$. It now follows that δ^* minimizes

$$2B(\delta) + [2R_2^*(C)/R_1^*(C)]S(\delta)$$

over $\delta > 0$. Note, however, that this is simply the worst-case β quantile of excess length of a one-sided $1-\alpha$ CI when $z_{1-\alpha} + z_\beta = 2R_2^*(C)/R_1^*(C)$, so this means that $\delta^*(C)$ is also optimal for this criterion. By Theorem A.1, the estimator $\hat{L}_{\tilde{\delta}}$ where $\tilde{\delta} = 2R_2^*(C)/R_1^*(C)$ is also optimal for this criterion. Furthermore, the estimator that optimizes this criterion is unique in this setting, so it follows that the estimator that optimizes the criterion $R(B,S)$ is equal to the estimator $\hat{L}_{\tilde{\delta}}$.

To show that this estimator is equal to the matching estimator with a single match once C is large enough, it now suffices to show that $R_2^*(C)/(2R_1^*(C))$ is bounded as $C \to \infty$ so that $C > KR_2^*(C)/(2R_1^*(C))$ once C is large enough. This can be checked by noting that, for the FLCI length and RMSE criteria, $R_1^*(C)$ is bounded from below and $R_2^*(C)$ is bounded from above, over the set $(B(\delta,C), S(\delta,C))$ with $C > 0$, using the fact that $S(\delta,C)$ is bounded from above and below away from zero over this set.

D.4 Proof of Lemma B.1

Let $A_n = \{x \in [a,b]^p$: there exists j such that $D_j = 0$ and $\|x - X_j\| \leq 2h\}$. Then $\#\mathcal{I}_n(h) = \sum_{i \in \mathcal{N}_{1,n}}[\mathbb{I}\{X_i \in [a,b]^p\} - \mathbb{I}\{X_i \in A_n\}]$. Note that, conditional on $\mathcal{E}$, the random variables $\mathbb{I}\{X_i \in A_n\}$ with $i \in \mathcal{N}_{1,n}$ are i.i.d. Bernoulli(ν_n) with $\nu_n = P(X_i \in A_n \mid \mathcal{E}) = \int \mathbb{I}\{x \in A_n\} f_{X|D}(x \mid 1)\, dx \leq K\lambda(A_n)$ where $f_{X|D}(x \mid 1)$ is the conditional density of X_i given $D_i = 1$, λ is the Lebesgue measure and K is an upper bound on this density. Under the assumption that $\limsup_n h_n n^{1/p} \leq \eta$, we have $\lambda(A_n) \leq (4h_n)^p n \leq 8^p \eta^p$ where the last inequality holds for large enough n. Thus, letting $\overline{\nu} = 8^p \eta^p K$, we can construct random variables Z_i for each $i \in \mathcal{N}_{1,n}$ that are i.i.d. Bernoulli$(\overline{\nu})$ conditional on $\mathcal{E}$ such that $\mathbb{I}\{X_i \in A_n\} \leq Z_i$. Applying the strong law of large numbers, it follows that

$$\begin{aligned}\liminf_n \#\mathcal{I}_n(h)/n &\geq \liminf_n \frac{\#\mathcal{N}_{1,n}}{n} \frac{1}{\#\mathcal{N}_{1,n}} \sum_{i \in \mathcal{N}_{1,n}} (\mathbb{I}\{X_i \in [a,b]^p\} - Z_i) \\ &\geq P(D_i = 1)(P(X_i \in [a,b]^p \mid D_i = 1) - 8^p \eta^p K)\end{aligned}$$

almost surely. This will be greater than η for η small enough.

D.5 Proof of Lemma B.2

The result follows from verifying the conditions of Theorem F.1 in Armstrong and Kolesár (2018). In particular, we need to show that the weights k are such that $\sum_{i=1}^n k(x_i,d_i)u_i/\operatorname{sd}_k$ converges in distribution to $N(0,1)$ (condition (S13) in Armstrong and Kolesár, 2018) and $\sum_i \hat{u}_i^2 k(x_i,d_i)^2/\operatorname{sd}_k^2$ converges in probability to 1, uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$ (S14), where $\operatorname{sd}_k^2 = \sum_{i=1}^n \sigma^2(x_i,d_i)k(x_i,d_i)^2$.

Under the moment bounds on u_i, eq. (22) directly implies the Lindeberg condition that is needed for condition (S13) to hold. To show that it also implies (S14), note that (S14) is equivalent to the requirement that $\sum_{i=1}^n \hat{u}_i^2 a_{ni} - \sum_{i=1}^n \sigma^2(x_i,n_i)a_{ni}$ converges to zero uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$, where

$$a_{ni} = k(x_i,d_i)^2 / \sum_{j=1}^n [\sigma^2(x_j,d_j)k(x_j,d_j)^2].$$

By an inequality of von Bahr and Esseen (1965),

$$\begin{aligned} E\left|\sum_{i=1}^n (u_i^2 - \sigma^2(x_i,d_i))a_{ni}\right|^{1+1/(2K)} &\le 2\sum_{i=1}^n a_{ni}^{1+1/(2K)} E|u_i^2 - \sigma^2(x_i,d_i)|^{1+1/(2K)} \\ &\le \max_{1\le i\le n} a_{ni}^{1/(2K)} E|u_i^2 - \sigma^2(x_i,d_i)|^{1+1/(2K)} \cdot \sum_{i=1}^n a_{ni}. \end{aligned}$$

Note that, by boundedness of $\sigma(x,d)$ away from zero and infinity, $\sum_{i=1}^n a_{ni}$ is uniformly bounded. Furthermore, it follows from eq. (22) that $\max_{1\le i\le n} a_{ni} \to 0$. From this and the moment bounds on u_i, it follows that the above display converges to zero. It therefore suffices to show that $\sum_{i=1}^n (\hat{u}_i^2 - u_i^2)a_{ni}$ converges to zero. This follows from the following result.

Lemma D.1. *Consider the model in eq. (1). Suppose that $1/K \le Eu_i^2 \le K$ and $E|u_i|^{2+1/K} \le K$ for some constant K, and that $\sigma^2(x,d)$ is uniformly continuous in x for $d \in \{0,1\}$. Let $\ell_j(i)$ be the jth closest unit to i, with respect to some norm $\|\cdot\|$, among units with the same value of the treatment. Let $\hat{u}_i^2 = \frac{J}{J+1}(Y_i - \sum_{j=1}^J Y_{\ell_j(i)}/J)^2$, and let $a_{ni} \ge 0$ be a non-random sequence such that $\max_i a_{ni} \to 0$, and that $\sum_{i=1}^n a_{ni}$ is uniformly bounded. If $\max_i C_n\|x_{\ell_J(i)} - x_i\| \to 0$, then $\sum_i a_{ni}(\hat{u}_i^2 - u_i^2)$ converges in probability to zero, uniformly over $\mathcal{F}_{\text{Lip}}(C_n)$.*

Proof. The proof is based on the arguments in Abadie and Imbens (2008). For ease of notation, let $f_i = f(x_i,d_i)$, $\sigma_i^2 = \sigma^2(x_i,d_i)$, and let $\overline{f}_i = J^{-1}\sum_{j=1}^J f_{\ell_j(i)}$ and $\overline{u}_i = J^{-1}\sum_{j=1}^J u_{\ell_j(i)}$. Then we can decompose

$$\begin{aligned} \frac{J+1}{J}(\hat{u}_i^2 - u_i^2) &= [f_i - \overline{f}_i + u_i - \overline{u}_i]^2 - \frac{J+1}{J}u_i^2 \\ &= [(f_i - \overline{f}_i)^2 + 2(u_i - \overline{u}_i)(f_i - \overline{f}_i)] - 2\overline{u}_i u_i + \frac{2}{J^2}\sum_{j=1}^J \sum_{k=1}^{j-1} u_{\ell_j(i)}u_{\ell_k(i)} + \frac{1}{J^2}\sum_{j=1}^J (u_{\ell_j(i)}^2 - u_i^2) \end{aligned}$$

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

where

$$T_{1i} = [(f_i - \overline{f}_i)^2 + 2(u_i - \overline{u}_i)(f_i - \overline{f}_i)], \qquad T_{2i} = \overline{u}_i u_i$$
$$T_{3i} = \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{j-1} u_{\ell_j(i)}u_{\ell_k(i)}, \qquad T_{4i} = \frac{1}{J^2}\sum_{j=1}^{J}(u^2_{\ell_j(i)} - \sigma^2_{\ell_j(i)}), \qquad T_{5i} = \sigma^2_i - u^2_i.$$

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

$$\max_i \max_{1\le j\le J} |\sigma^2_{\ell_j(i)} - \sigma^2_i| \to 0,$$

and hence that $|\sum_{i=1}^n a_{ni} J^{-1} \sum_{j=1}^J (\sigma^2_{\ell_j(i)} - \sigma^2_i)| \le \max_i \max_{j=1,\dots,J}(\sigma^2_{\ell_j(i)} - \sigma^2_i) \sum_{i=1}^n a_{ni} \to 0$. 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,

$$E|\sum_i a_{ni} T_{1i}| \le \max_i (f_i - \overline{f}_i)^2 \sum_i a_{ni} + 2 \max_i |f_i - \overline{f}_i| \sum_i a_{ni} E|u_i - \overline{u}_i|,$$

which converges to zero since $\max_i |f_i - \overline{f}_i| \le \max_i \max_{j=1,\dots,J}(f_i - f_{\ell_j(i)}) \le C_n \max_i \|x_i - x_{\ell_J(i)}\|_{\mathcal{X}} \to 0$. Next, by the von Bahr-Esseen inequality,

$$E|\sum_{i=1}^n a_{ni} T_{5i}|^{1+1/2K} \le 2\sum_{i=1}^n a_{ni}^{1+1/2K} E|T_{5i}|^{1+1/2K} \le 2 \max_i a_{ni}^{1/2K} \max_j E|T_{5j}|^{1+1/2K} \sum_{k=1}^n a_{nk} \to 0.$$

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 we use the fact $\#\mathcal{I}_j$ is bounded by $J\overline{L}$, where $\overline{L}$ 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). $\overline{L}$ is a finite constant that depends only on the dimension of the covariates (for example, $\overline{L} = 2$ if $\dim(x_i) = 1$). Now,

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

and so by the von Bahr-Esseen inequality,

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

$$\leq \frac{(J\overline{L})^{1/2K}}{J^{2+1/K}} \max_k E|u_k - \sigma_k^2|^{1+1/2K} \max_i a_{ni}^{1+1/2K} \sum_{j=1}^n \sum_{i \in \mathcal{I}_j} a_{ni},$$

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 $\sum_{i'=1}^n a_{ni'} E[u_i u_{i'} u_{\ell_j(i)} u_{\ell_k(i')}] \leq \max_{i'} a_{ni'} \left(\sigma_i^2 \sigma_{\ell_j(i)}^2 + \sigma_{\ell_j(i)}^2 \sigma_i^2\right)$, so that

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

Similarly for $j \neq k$ and $j' \neq k$, $\sum_{i'=1}^n a_{ni'} E[u_{\ell_j(i)} u_{\ell_k(i)} u_{\ell_{j'}(i')} u_{\ell_{k'}(i')}] \leq \max_{i'} 2\sigma_{\ell_j(i)}^2 \sigma_{\ell_k(i)}^2$, so that

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

□

D.6 Asymptotic efficiency of the matching estimator

By Theorem 2.2, the matching estimator with $M = 1$ is efficient in finite samples if the Lipschitz constant C is large enough. We now give conditions for its asymptotic optimality.

Theorem D.1. *Suppose that the assumptions of Theorem 4.1 hold, and that $\sigma^2(x, d)$ is bounded away from zero and infinity. Suppose that, for some functions $\overline{G}\colon \mathbb{R}^+ \to \mathbb{R}^+$ and $\underline{G}\colon \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{t \to 0} \overline{G}(\underline{G}^{-1}(t))^2 / [t / \log t^{-1}]^{2/p+1} = 0$,*

$$\underline{G}(a) \leq P(\|X_i - x\|_{\mathcal{X}} \leq a,\, D_i = d) \leq \overline{G}(a).$$

Let $R^*_{n,match,RMSE}$ denote the worst-case RMSE of the matching estimator with $M = 1$, and let $R^*_{n,opt,RMSE}$ denote the minimax RMSE among linear estimators, conditional on $\{X_i, D_i\}_{i=1}^n$, for the class $\mathcal{F}_{\text{Lip}}(C)$. Then $R^*_{n,match,RMSE} / R^*_{n,opt,RMSE} \to 1$ almost surely. The same holds with "RMSE" replaced by "CI length" or "β quantile of excess length of a one-sided CI."

If X_i has sufficiently regular support and the conditional density of X_i given D_i is bounded away from zero on the support of X_i for both $D_i = 0$ and $D_i = 1$, then the conditions of Theorem D.1 hold with $\underline{G}(a)$ and $\overline{G}(a)$ both given by constants times a^p, so that $\overline{G}(\underline{G}(a))$ decreases like a as $a \to 0$. Thus, the conditions of Theorem D.1 hold so long as $p > 2$ and there is sufficient overlap.

Proof. Let $\text{sd}_{\delta_{\text{RMSE}},n}$ and $\overline{\text{bias}}_{\delta_{\text{RMSE}},n}$ denote the standard deviation and worst-case bias of the minimax linear estimator and let $\text{sd}_{\text{match},1}$ and $\overline{\text{bias}}_{\text{match},1}$ denote the standard deviation and worst-case bias of the estimator with a single match (conditional on $\{(X_i, D_i)_{i=1}^n\}$). Since worst-case bias

is increasing in δ and variance is decreasing in δ, and since the matching estimator with $M = 1$ solves the modulus problem for small enough δ by Theorem 2.3, we have $\overline{\text{bias}}_{\delta_{\text{RMSE}},n} \geq \overline{\text{bias}}_{\text{match},1}$. Thus,

$$1 \leq \frac{\overline{\text{bias}}^2_{\text{match},1} + \text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n} + \text{sd}^2_{\delta\text{RMSE},n}} \leq \frac{\overline{\text{bias}}^2_{\delta\text{RMSE},n} + \text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n} + \text{sd}^2_{\delta\text{RMSE},n}} \leq 1 + \frac{\text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n} + \text{sd}^2_{\delta\text{RMSE},n}}.$$

By the arguments in the proof of Theorem 4.1, there exists $\varepsilon > 0$ such that $\overline{\text{bias}}_{\delta\text{RMSE},n} \geq \varepsilon n^{-2/p}$ almost surely. In addition, by Theorem 37 in Chapter 2 of Pollard (1984), the conditions of Theorem 4.3 hold almost surely (with $\underline{G}(a)$ and $\overline{G}(a)$ multiplied by some positive constants). Arguing as in the proof of Theorem 4.3 then gives the bound $\text{sd}^2_{\text{match},1} \leq [2 \max_{1 \leq i \leq n} K_1(i)]^2/n \leq [2n\overline{G}(a_n)]^2/n$ for any sequence $a_n = \underline{G}^{-1}(c_n(\log n)/n)$ with $c_n = n\overline{G}(a_n)/\log n \to \infty$. Plugging these bounds into the above display gives a bound proportional to

$$\overline{G}(\underline{G}^{-1}(c_n(\log n)/n))^2 n^{2/p+1} = b(c_n(\log n)/n) \left[\frac{c_n(\log n)/n}{\log n - \log c_n - \log\log n}\right]^{2/p+1} n^{2/p+1},$$

where $b(t) = \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}]^{2/p+1}$. If $\lim_{t \to 0} b(t) = 0$, then this can be made to converge to zero by choosing c_n to increase slowly enough. Similar arguments apply to the other performance criteria. □

D.7 Verification of the conditions in Theorem B.1 for the matching estimator

For matching estimators with a fixed number of matches we use results from Abadie and Imbens (2006) and Abadie and Imbens (2016) to verify Assumption B.1. Since such results appear to be available only for the case where X_i is scalar, we restrict ourselves to this case, and we leave the question of verifying Assumption B.1 when X_i is multivariate for future research. Since these results are stated for a single underlying distribution, we restrict attention to the case where the distribution of (X_i, D_i) is fixed over $P \in \mathcal{P}$ (but where the conditional expectation function f_P is allowed to vary over the given class $\mathcal{F}$).

Theorem D.2. *Suppose that the class $\mathcal{P}$ is such that the marginal distribution of (X_i, D_i) and the conditional variance function $\sigma^2_P(x, d)$ is the same for all $P \in \mathcal{P}$, and such that the following conditions hold: (i) X_i is scalar, and is supported on a compact interval $[a, b]$ with continuous density (ii) $\sigma^2_P(x, d)$ is continuous and uniformly bounded away from zero and infinity (iii) $0 < P(D_i = 1) < 1$ and letting $g(x \mid d)$ denote the density of X_i given D_i, $g(x \mid 1)/g(x \mid 0)$ is uniformly bounded from above and below away from zero on $[a, b]$. Suppose, in addition, that, for some η, $E_P(u_i^{2+\eta} \mid X_i = x, D_i = d) \leq 1/\eta$ for $d \in \{0, 1\}$, all x and all $P \in \mathcal{P}$. Then Assumption B.1 holds for the weights $k(X_i, D_i) = \frac{1}{n}(2D_i - 1)\left(1 + \frac{K_M(i)}{M}\right)$ for the matching estimator with M matches.*

Proof. Part (i) of Assumption B.1 follows from Lemma S.11 in Abadie and Imbens (2016). The formula for $V_{1,n}(P)$ follows from this lemma as well, and is given by a constant times $1/n$ (where,

under our assumptions, the constant is strictly positive and does not depend on P). Thus, to verify part (ii) of Assumption B.1, it suffices to show this condition with $V_{1,n}(P)$ replaced by $1/n$. To this end, note that replacing $V_{1,n}(P)$ with $1/n$ in this condition gives

$$n^2 E_P[k(X_i,D_i)^2 u_i^2 \mathbb{I}\{k(X_i,D_i)^2 u_i^2 > \varepsilon/n\}] = E_P[(1+K_M(i)/M)^2 u_i^2 \mathbb{I}\{(1+K_M(i))^2 u_i^2 > \varepsilon \cdot n\}].$$

This will converge to zero by the standard arguments showing that the Lyapunov condition implies the Lindeberg condition, so long as $E_P[(1+K_M(i)/M)^{2+\eta} u_i^{2+\eta}]$ is uniformly bounded. Indeed, the bound on the conditional $2+\eta$ moment of u_i implies that this is bounded by a constant times $E_P[(1+K_M(i)/M)^{2+\eta}]$, which is bounded uniformly in i and n by Lemma S.8 in Abadie and Imbens (2016). □

We now consider construction of the standard error $\text{se}_\tau(\hat{L}_k)$. 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 $\text{se}(\hat{L}_k)^2 / \sum_{i=1}^n k(X_i,D_i)^2 \sigma_P^2(X_i,D_i)$ 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 $\text{se}(\hat{L}_k)^2 / \sum_{i=1}^n k(X_i,D_i)^2 \sigma_P^2(X_i,D_i)$ 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.3. *Suppose that* $\max_{1\le i\le n, d\in\{0,1\}} |\hat{f}(X_i,d) - f_P(X_i,d)| \xrightarrow{p} 0$ *and* $\hat{L}_k \xrightarrow{p} \tau(P)$ *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* $\text{se}_\tau(\hat{L}_k)^2 = \text{se}(\hat{L}_k)^2 + \hat{V}_{2,n}$ *where* $\text{se}(\hat{L}_k)^2 / V_{1,n}(P)$ *converges in probability to one uniformly over* $P \in \mathcal{P}$*, then* $[V_{1,n}(P) + V_{2,n}(P)] / \text{se}_\tau(\hat{L}_k)^2 \xrightarrow{p} 1$ *uniformly over* $P \in \mathcal{P}$*.*

Proof. We have

$$\begin{aligned}
&|\hat{V}_{2,n}/n - V_{2,n}(P)/n| \\
&\qquad = \left| \frac{1}{n}\sum_{i=1}^n \{[\hat{f}(X_i,1) - \hat{f}(X_i,0)]^2 - [f_P(X_i,1) - f_P(X_i,0)]^2\} + \tau(P)^2 - \hat{L}_k^2 \right| \\
&\qquad\qquad \le 2 \max_{1\le i\le n, d\in\{0,1\}} |\hat{f}(X_i,d) - f_P(X_i,d)|^2 + |\hat{L}_k^2 - \tau(P)^2|,
\end{aligned}$$

which converges in probability to zero uniformly over $P \in \mathcal{P}$. By the $\mathcal{O}(1/n)$ 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

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, (91/92):175–187.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Miller, G. L., Teng, S.-H., Thurston, W., and Vavasis, S. A. (1997). Separators for sphere-packings and nearest neighbor graphs. *Journal of the ACM*, 44(1):1–29.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–404.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York, NY.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.