Supplement to "Optimality of Matched-Pair Designs in Randomized Controlled Trials"

Yuehao Bai
Department of Economics
University of Michigan
yuehaob@umich.edu

November 23, 2020

Abstract

This document provides proofs for all results for the author's paper "Optimality of Matched-Pair Designs in Randomized Controlled Trials," as well as some additional results.

KEYWORDS: Matched-pair design, stratified randomization, randomized controlled trial, ex-post bias, treatment effect, stratification, pilot experiment, matched pairs

JEL CLASSIFICATION CODES: C12, C13, C14, C90

S.1 Proof of main results

For the rest of the appendix we introduce the following definition for the convex combination of matched-pair designs.

Definition S.1.1. For $\lambda, \lambda' \in \Lambda_n^{\text{pair}}$ and $\delta \in [0,1]$, define $\delta\lambda \bigoplus (1-\delta)\lambda'$ as the randomization between λ and λ' such that λ is implemented with probability δ. Define the convex hull formed by all convex combinations of any matched-pair designs as

$$\text{co}(\Lambda_n^{\text{pair}}) = \left\{ \bigoplus_{1\le j\le J} \delta_j \lambda^j : \lambda^j \in \Lambda_n^{\text{pair}} \text{ and } \delta_j \ge 0 \text{ for } 1 \le j \le J, \sum_{1\le j\le J} \delta_j = 1, 1 \le J < \infty \right\} . \quad \text{(S.1)}$$

S.1.1 Proof of Theorem 3.1

Define $V(\lambda)$ as the objective in (18) multiplied by n^2. We have

$$\begin{aligned} V(\lambda) &= n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n | X^{(n)}, D^{(n)}] | X^{(n)}] \\ &= \operatorname{Var}_\lambda \left[\sum_{1\le i\le 2n} [D_i E[Y_i(1)|X_i] - (1-D_i)E[Y_i(0)|X_i]] \middle| X^{(n)} \right] \\ &= \operatorname{Var}_\lambda \left[\sum_{1\le i\le 2n} D_i (E[Y_i(0)|X_i] + E[Y_i(1)|X_i]) \middle| X^{(n)} \right] \\ &= (g^{(n)})' \operatorname{Var}_\lambda[D^{(n)}] g^{(n)} . \end{aligned}$$

Recall from Section 2 that Λ_n^{pair} is the set of all matched-pair designs. For any $\lambda = \{\{\pi(1), \pi(2)\}, \ldots, \{\pi(2n-1), \pi(2n)\}\} \in \Lambda_n^{\text{pair}}$,

$$V(\lambda) = \frac{1}{4} \sum_{1\le s\le n} (g_{\pi(2s-1)} - g_{\pi(2s)})^2 . \quad \text{(S.2)}$$

By Lemma S.3.2, we have $V(\lambda^g(X^{(n)})) \le V(\lambda)$.

Recall the definition of convex combinations of matched-pair designs from Definition S.1.1. To conclude the proof, note that by Lemma S.3.1, for any $\lambda \in \Lambda$ we have

$$\lambda = \bigoplus_{1\le j\le J} \delta_j \lambda^j ,$$

where $\lambda^j \in \Lambda_n^{\text{pair}}$ and $\delta_j \ge 0$ for $1 \le j \le J$, $\sum_{1\le j\le J} \delta_j = 1$, and $1 \le J < \infty$. Then,

$$\text{MSE}(\lambda | X^{(n)}) = \sum_{1\le j\le J} \delta_j \text{MSE}(\lambda^j | X^{(n)}) \ge \min_{1\le j\le J} \text{MSE}(\lambda^j | X^{(n)}) \ge \text{MSE}(\lambda^g(X^{(n)}) | X^{(n)}) ,$$

where the last inequality follows because $\lambda^g(X^{(n)})$ minimizes $\mathrm{MSE}(\lambda|X^{(n)})$ over $\Lambda_n^{\mathrm{pair}}$. The theorem therefore follows. ■

S.1.2 Proof of Theorem 5.3

First, note that the assumptions in Lemma S.3.4 hold because of Lemma S.3.7 and Assumption 5.4, and that $\hat{g}_m$ is a fixed function conditional on $\tilde{W}^{(m)}$. Hence, by Lemma S.3.4 with $\tau = \frac{1}{2}$, with probability one for $\tilde{W}^{(m)}$, as $n \to \infty$,

$$\sup_{t \in \mathbf{R}} \left| Q\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \leq t | \tilde{W}^{(m)}\} - \Phi(z/\varsigma_{\hat{g}_m}) \right| \to 0 \,, \tag{S.3}$$

where

$$\varsigma^2_{\hat{g}_m} = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2} E[(E[Y_i(1) + Y_i(0) | \hat{g}_m(X_i), \tilde{W}^{(m)}] - E[Y_i(1) + Y_i(0)])^2] \,. \tag{S.4}$$

On the other hand, note that the assumptions in Lemma S.3.5 hold because of Lemma S.3.7 and Assumption 5.4, and that $\hat{g}_m$ is a fixed function conditional on $\tilde{W}^{(m)}$. Hence, by Lemma S.3.5 with $\tau = \frac{1}{2}$, with probability one for $\tilde{W}^{(m)}$, for all $\epsilon > 0$, as $n \to \infty$,

$$Q\{|(\hat{\varsigma}_n^{\hat{g}_m})^2 - \varsigma^2_{\hat{g}_m}| > \epsilon | \tilde{W}^{(m)}\} \to 0 \,. \tag{S.5}$$

The conditional convergence in (40) follows immediately from (S.3) and (S.5). Since the conditional convergence holds with probability one for $\tilde{W}^{(m)}$, and $\phi_n^{\hat{g}_m}(W^{(n)}) \in [0, 1]$, the unconditional convergence follows from the dominated convergence theorem. ■

S.1.3 Proof of Theorem 5.1

The first assertion follows from Lemma S.3.4 with $h = g$ and $\tau = \frac{1}{2}$. We now show that under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (23), $\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_g^2)$ for ς_g^2 defined in (34) as $m, n \to \infty$. By repeating arguments in the proof of Lemma S.3.4, we write

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) = A_n - B_n + C_n - D_n \,,$$

where

$$A_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (Y_i(1)D_i - E[Y_i(1)D_i | g^{(n)}, D^{(n)}])$$
$$B_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (Y_i(0)(1-D_i) - E[Y_i(0)(1-D_i) | g^{(n)}, D^{(n)}])$$
$$C_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (E[(Y_i(1)+Y_i(0))D_i | g^{(n)}, D^{(n)}] - D_i E[Y_i(1)+Y_i(0)])$$
$$D_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (E[Y_i(0) | g^{(n)}, D^{(n)}] - E[Y_i(0)]) \,.$$

Note that unlike in Lemma S.3.4, the quantities above are conditioned on $g^{(n)}$ for g defined in (19), instead of $\hat{g}_m^{(n)}$. Note that by Assumptions 5.2(c), 5.3, and Lemma S.3.8,

$$\frac{1}{n} \sum_{1 \le s \le n} (g_{\pi^{\hat{g}_m}(2s-1)} - g_{\pi^{\hat{g}_m}(2s)})^2 \xrightarrow{P} 0 \,. \tag{S.6}$$

Since Assumption 5.2(a)–(b) and (S.6) hold, by repeating arguments in the proof of Lemma S.3.4 with $\tau = \frac{1}{2}$, it is straightforward to establish that as $m, n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_g^2) \,, \tag{S.7}$$

Note that (S.6) is enough to derive the asymptotic representation for C_n so that we need not impose Lipschitz conditions on $E[Y_i(d)|g(X_i)]$. ■

S.1.4 Proof of Theorem 5.2

In light of Theorem 5.1, we only need to show that $(\hat{\varsigma}_n^{\hat{g}_m})^2 \xrightarrow{P} \varsigma_g^2$ as $m, n \to \infty$. Similar arguments as those used in Lemma S.3.5 go through if (S.6) holds and

$$\frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} |g_{\pi^{\hat{g}_m}(4j-k)} - g_{\pi^{\hat{g}_m}(4j-l)}|^2 \xrightarrow{P} 0 \tag{S.8}$$

for $k \in \{2, 3\}$ and $l \in \{0, 1\}$. Since (S.8) follows from Assumptions 5.3 by Lemma S.3.8, the proof is concluded.

S.1.5 Proof of Theorem 5.5

To begin with, note that we need only establish that as $m, n \to \infty$,

$$\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) \xrightarrow{d} N(0, \nu \varsigma_{\text{pilot}}^2 + (1-\nu) 2\varsigma^2) \,, \tag{S.9}$$

and the rest follows from Slutsky's lemma. We prove (S.9) by contradiction. Suppose (S.9) does not hold. Then, there exists a subsequence still denoted by $\{m, n\}$ for notational simplicity, along which as $m, n \to \infty$,

$$\sup_{t \in \mathbf{R}} \left| \sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) - \Phi(z/\sqrt{\nu \varsigma_{\text{pilot}}^2 + (1-\nu) 2\varsigma^2}) \right| \to c \,, \tag{S.10}$$

where $c > 0$, and

$$\frac{m}{m+2n} \to \nu \in [0,1] \,.$$

Now consider this subsequence. Since the two convergences in the Lemma S.3.8 hold in probability, there exists a further subsequence along which they hold with probability one. By repeating the proof of Theorem 5.2, we could see that along this subsequence, as $m, n \to \infty$, with probability one for $\tilde{W}^{(m)}$,

$$\sup_{t \in \mathbf{R}} \left| Q\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \le t | \tilde{W}^{(m)}\} - \Phi(z/\varsigma_g) \right| \to 0 \,. \tag{S.11}$$

Along the subsequence we construct, since $\frac{m}{m+2n} \to \nu$, by (S.11), Slutsky's lemma and Lemma S.3.3,

$$\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) \xrightarrow{d} N(0, \nu \varsigma_{\text{pilot}}^2 + (1-\nu) 2\varsigma^2) \,,$$

which is a contradiction to (S.10). The theorem therefore holds. ■

S.1.6 Proof of Theorem 5.4

Follows from Theorem 4.2 in Bai et al. (2019) and by repeating arguments in the proof of Lemma S.3.4. ■

S.2 Supplementary results

The next theorem shows that the infeasible optimal stratification has a similar structure to (20) when $\tau \neq \frac{1}{2}$.

Theorem S.2.1. *Suppose the sample size is kn for $k \in \mathbb{Z}$ and the treatment assignment scheme satisfies $\tau_s \equiv \tau = \frac{l}{k}$, where $l \in \mathbb{Z}$, $0 < l < k$, and k and l are relatively prime. Then, (5) is solved by $\lambda^{\tau,g}$ defined in (22), where $g^\tau_{\pi^{\tau,g^\tau}(1)} \le \ldots \le g^\tau_{\pi^{\tau,g^\tau}(kn)}$ for g^τ defined in (21).*

PROOF OF THEOREM S.2.1. First, note that

$$\hat{\theta}_n = \frac{1}{kn} \sum_{1 \le i \le kn} \left(\frac{1}{\tau} Y_i(1) D_i - \frac{1}{1-\tau} Y_i(0)(1 - D_i) \right) \,.$$

Next,

$$\text{MSE}(\lambda | X^{(n)}) = (E_\lambda[\hat{\theta}_n | X^{(n)}] - \theta(Q))^2 + \text{Var}_\lambda[\hat{\theta}_n | X^{(n)}] \,.$$

By repeating arguments in the proof of Lemma 3.1,

$$E_\lambda[\hat{\theta}_n|X^{(n)}] - \theta(Q) = \frac{1}{kn}\sum_{1\le i\le kn}(E[Y_i(1)|X_i] - E[Y_i(0)|X_i]) - \theta(Q)\,,$$

identical across all $\lambda \in \Lambda_n$, so that we need only consider conditional variances of $\hat{\theta}$ given $X^{(n)}$ which could be decomposed as in (11). By repeating arguments in the proof of Lemma 3.1, for any $\lambda \in \Lambda_n$, the first term of the right-hand side of (11) equals

$$\frac{1}{k^2n^2}\sum_{1\le i\le kn}\left(\frac{\mathrm{Var}[Y_i(1)|X_i]}{\tau} + \frac{\mathrm{Var}[Y_i(0)|X_i]}{1-\tau}\right)\,,$$

again identical across all $\lambda \in \Lambda_n$. Therefore, we need only consider

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}]\,.$$

By repeating arguments in the proof of Lemma S.3.1, a stratum of size kl where $l > 1$ is a convex combination of stratifications with strata only of size k. We could therefore focus on the case where each stratum is of size k. For any stratification of the form $\lambda = \{\{\pi((s-1)k+1, \dots \pi(sk)\} : 1 \le s \le n\}$,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk}(g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2\,,$$

where g^τ_i is defined in (21) and

$$\bar{g}^\tau_s = \frac{1}{k}\sum_{(s-1)k+1\le j\le sk} g^\tau_{\pi(j)}\,.$$

To see this, first note that units are independent across strata, so that by repeating arguments in the proof of Lemma 3.1,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1\le s\le n}\mathrm{Var}_\lambda\left[\sum_{(s-1)k+1\le j\le sk} g^\tau_{\pi(j)}D_{\pi(j)}\right]\,.$$

Next,

$$\begin{aligned}
&\mathrm{Var}_\lambda\left[\sum_{(s-1)k+1\le j\le sk} g^\tau_{\pi(j)}D_{\pi(j)}\right] \\
&= \frac{1}{\binom{k}{l}}\sum_{(s-1)k+1\le j_1<\dots<j_l\le sk}\left(\sum_{1\le \iota\le l} g^\tau_{\pi(j_\iota)} - l\bar{g}^\tau_s\right)^2 \\
&= \frac{l}{k}\sum_{(s-1)k+1\le j\le sk}(g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 + \frac{1}{\binom{k}{l}}\sum_{(s-1)k+1\le j_1<\dots<j_l\le sk}\sum_{1\le \iota_1\ne\iota_2\le l}(g_{\pi(j_{\iota_1})} - \bar{g}^\tau_s)(g_{\pi(j_{\iota_2})} - \bar{g}^\tau_s) \\
&= \frac{l}{k}\sum_{(s-1)k+1\le j\le sk}(g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 + \frac{\binom{k-2}{l-2}}{\binom{k}{l}}\left[\left(\sum_{(s-1)k+1\le j\le sk} g^\tau_{\pi(j)} - k\bar{g}^\tau_s\right)^2 - \sum_{(s-1)k+1\le j\le sk}(g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2\right]
\end{aligned}$$

$$\propto \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 \,,$$

where the first equality holds by definition, the second holds by expanding the square, the third holds by accounting for cross product terms, and the fourth holds because the first term inside the square bracket on the fourth line is 0. Therefore, the problem is reduced to optimal univariate clustering of kn units on the real line where each cluster is of size k, and the conclusion follows by arguing similarly to in the proof of Lemma S.3.2. ■

For a measurable function $h : \mathbf{R}^p \to \mathbf{R}$, let π^h be a permutation of $\{1, \ldots, kn\}$ such that $h_{\pi^{\tau,h}(1)} \le \ldots \le h_{\pi^{\tau,h}(kn)}$. Define

$$\lambda^{\tau,h}(X^{(n)}) = \{\{\pi^{\tau,h}((s-1)k+1), \ldots \pi^{\tau,h}(sk)\} : 1 \le s \le n\} \,. \tag{S.12}$$

Further define $\bar{h}^\tau_s = \frac{1}{k} \sum_{(s-1)k+1 \le j \le sk} h_{\pi^{\tau,h}(j)}$.

Assumption S.2.1. h satisfies

(a) $0 < E[\mathrm{Var}[Y_i(d)|h(X_i)]]$ for $d \in \{0, 1\}$.

(b) $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz for $r = 1, 2$ and $d = 0, 1$.

(c) $\frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}^\tau_s|^2 \xrightarrow{P} 0$.

The next theorem is the limiting counterpart to Theorems 3.1 and S.2.1. It shows that across all stratifications defined by (S.12) for h satisfying Assumption S.2.1, the asymptotic variance of $\hat{\theta}_n$ is minimized by choosing $h = g^\tau$ defined in (21).

Theorem S.2.2. *Suppose $h : \mathbf{R}^p \to \mathbf{R}$ be a measurable function that satisfies Assumption S.2.1. Then,*

$$\varsigma^2_{\tau,g^\tau} \le \varsigma^2_{\tau,h} \,,$$

for $\varsigma^2_{\tau,g^\tau}$ and $\varsigma^2_{\tau,h}$ defined in (S.17) and g^τ defined in (21). Moreover, the inequality is strict unless $E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau} \middle| h(X_i)\right] = g^\tau(X_i)$ with probability one under Q.

PROOF OF THEOREM S.2.2. By the definition of $\varsigma^2_{\tau,h}$ in (S.17), minimizing $\varsigma^2_{\tau,h}$ with respect to h is equivalent to maximizing

$$E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] .$$

Next, note that

$$E\left[\left(g^\tau(X_i) - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right]$$

$$= E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] + E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right]$$

$$= E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\right)^2\right] + E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right], \tag{S.13}$$

$$\geq E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right]. \tag{S.14}$$

where the last inequality is strict except unless $E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] = g^\tau(X_i)$ with probability one under Q. To show (S.13), note that

$$E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\right)\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)\right]$$
$$= E\left[E\left[g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\middle| h(X_i)\right]\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)\right]$$
$$= 0\,,$$

where the second equality holds because

$$E[g^\tau(X_i)|h(X_i)] = E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]$$

by the law of iterated expectation. The lemma is thus proved. ■

If τ_s's are allowed to differ across s, then $\hat{\theta}_n$ is generally inconsistent for θ. In such settings researchers often use the estimator from the fully saturated regression in Bugni et al. (2019). For $1 \leq s \leq S$ and $d \in \{0, 1\}$, define

$$\hat{\mu}_{n,s}(1) = \frac{1}{n_s \tau_s} \sum_{i \in \lambda_s : D_i = 1} Y_i$$

and

$$\hat{\mu}_{n,s}(0) = \frac{1}{n_s(1-\tau_s)} \sum_{i \in \lambda_s : D_i = 0} Y_i\,.$$

The estimator is

$$\hat{\theta}_n^{\text{sat}} = \sum_{1 \leq s \leq S} \frac{n_s}{n}(\hat{\mu}_{n,s}(1) - \hat{\mu}_{n,s}(0))\,. \tag{S.15}$$

Note that $\hat{\theta}_n^{\text{sat}}$ and $\hat{\theta}_n$ coincide whenever $\tau_s \equiv \tau \in (0, 1)$. See Bugni et al. (2018), Tabord-Meehan (2020), and Bugni et al. (2019) for more details. By repeating arguments used in the proof of Theorem 3.1 and Theorem S.2.1, we could find the stratification that minimizes $\text{MSE}(\hat{\theta}_n^{\text{sat}}|X^{(n)})$, which is defined as in (4) with $\hat{\theta}_n$ replaced by θ_n^{sat}. The solution is as follows: we first calculate the stratification defined in (22) with τ, g, and $X^{(n)}$ defined separately for each subpopulation, and then take the union of those stratifications. Moreover, the next theorem enables us to derive feasible procedures similar to (23) when treated fractions are allowed to vary across subpopulations. In particular, it reveals any

plug-in estimator that satisfies the regularity conditions in Assumption S.2.1 leads to a procedure under which the asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ is no greater than and typically strictly less than that under procedures with each subpopulation as a stratum.

Theorem S.2.3. *Suppose the sample size is n. Define a function $f : \mathbf{R}^p \to \{1, \ldots, R\}$ where $R \geq 1$ is an integer. Define $N_r = \{i : f(X_i) = r\}$, $X^{N_r} = (X_i : i \in N_r)$, $n_r = |N_r|$, and $p(r) = Q\{f(X_i) = r\}$. Define $\lambda^{\text{large}} = \bigcup_{1 \leq r \leq R} N_r$. For $1 \leq r \leq R$, let τ_r be the treated fraction in N_r. Define functions $h^r : \mathbf{R}^p \to \mathbf{R}$ for $1 \leq r \leq R$. Define $\lambda^{\text{small}} = \bigcup_{1 \leq r \leq R} \lambda^{\tau_r, h^r}(X^{N_r})$, where $\lambda^{\tau_r, h^r}(X^{N_r})$ is defined in (S.12). Suppose Q satisfies Assumption 5.1. Then, under λ^{large}, for $\hat{\theta}_n^{\text{sat}}$ defined in (S.15), as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{large}}) \,,$$

where

$$\varsigma^2_{\text{large}} = E\left[\frac{\text{Var}[Y_i(1)]}{\tau_{f_i}} + \frac{\text{Var}[Y_i(0)]}{1-\tau_{f_i}} - \tau_{f_i}(1-\tau_{f_i})E\left[\left(E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| f(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau_{f_i}} + \frac{E[Y_i(0)]}{1-\tau_{f_i}}\right)\right)^2\right]\right] .$$

Suppose in addition that $h^r, 1 \leq r \leq R$ satisfy Assumption S.2.1, under Q restricted to $\{x \in \mathbf{R}^p : f(x) = r\}$. Then, under λ^{small}, for $\hat{\theta}_n^{\text{sat}}$ defined in (S.15), as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{small}}) \,,$$

where

$$\varsigma^2_{\text{small}} = E\left[\frac{\text{Var}[Y_i(1)]}{\tau_{f_i}} + \frac{\text{Var}[Y_i(0)]}{1-\tau_{f_i}} - \tau_{f_i}(1-\tau_{f_i})E\left[\left(E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| h^{f_i}(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau_{f_i}} + \frac{E[Y_i(0)]}{1-\tau_{f_i}}\right)\right)^2\right]\right] .$$

In addition, $\varsigma^2_{\text{small}} \leq \varsigma^2_{\text{large}}$, where the inequality is strict unless

$$E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| h^{f_i}(X_i)\right] = E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| f(X_i)\right]$$

with probability one under Q. Moreover, among all choices of $(h^r : 1 \leq r \leq R)$, $\varsigma^2_{\text{small}}$ is minimized by setting $h^r = g^{\tau_r}$, where g^{τ_r} is defined in (21).

Remark S.2.1. Tabord-Meehan (2020) considers stratification trees, which leads to a small number of large strata, with different treated fractions in each stratum. Using results from Theorem S.2.3, it is straightforward to combine his procedure with procedures in this paper. The asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ under the combined procedure is no greater than and typically strictly less than that under his procedure alone. The combined procedure is as follows: First, perform the procedure in Tabord-Meehan (2020), which produces a finite number of strata with a target treated fraction for each stratum. Second, we view each stratum as a subpopulation and calculate the stratification in (S.12) either with a fixed function h or some plug-in estimate, with τ equal the target treated fraction. Finally, we take the union of these stratifications. The desired property now follows from Theorem S.2.3. ■

PROOF OF THEOREM S.2.3. The first convergence holds by Theorem 3.1 of Bugni et al. (2019). For the second convergence, note that

$$\begin{pmatrix} \sqrt{n_1}(\hat{\mu}_{n,1}(1) - \hat{\mu}_{n,1}(0)) \\ \vdots \\ \sqrt{n_R}(\hat{\mu}_{n,R}(1) - \hat{\mu}_{n,R}(0)) \end{pmatrix} \xrightarrow{d} N\left(0, \operatorname{diag}(\varsigma^2_{\tau_r,h^r} : 1 \le r \le R)\right) .$$

Meanwhile, note that $\frac{n_r}{n} \xrightarrow{P} p(r)$ for $1 \le r \le R$. The convergence then follows by the Slutsky's lemma. The last two results could be shown similarly to Theorem S.2.2. ■

S.3 Auxiliary Lemmas

In the rest of the appendix, we use $a \lesssim b$ to denote that there exists $c \ge 0$ such that $a \le cb$.

Lemma S.3.1. *If the treatment assignment scheme satisfies Assumption 2.1, then* $\Lambda_n \subseteq \operatorname{co}(\Lambda_n^{\text{pair}})$.

PROOF OF LEMMA S.3.1. We first prove that $\lambda_0 = \{\{X_1, \ldots, X_{2n}\}\}$ is a convex combination of matched-pair designs. Indeed,

$$\lambda_0 = \frac{1}{|\Lambda_n^{\text{pair}}|} \bigoplus_{\lambda \in \Lambda_n^{\text{pair}}} \lambda ,$$

where

$$|\Lambda_n^{\text{pair}}| = \frac{\binom{2n}{n} n!}{2^n} .$$

Next, consider $\lambda = \{\lambda_1, \ldots, \lambda_S\}$. Let $\Lambda_n^{\text{pair}}(\lambda_s)$ denote the set of all matched-pair designs of units in λ_s. Then,

$$\lambda = \frac{1}{\prod_{1 \le s \le S} |\Lambda_n^{\text{pair}}(\lambda_s)|} \bigoplus_{\xi^s \in \Lambda_n^{\text{pair}}(\lambda_s) : 1 \le s \le S} \bigcup_{1 \le s \le S} \xi^s ,$$

and the conclusion follows. ■

Example S.3.1. Let $n = 4$ and define

$$\begin{aligned} \lambda^0 &= \{\{1,2,3,4\}\} \\ \lambda^1 &= \{\{1,2\},\{3,4\}\} \\ \lambda^2 &= \{\{1,3\},\{2,4\}\} \\ \lambda^3 &= \{\{1,4\},\{2,3\}\} . \end{aligned}$$

We have $\lambda^0 = \frac{1}{3}\lambda^1 \bigoplus \frac{1}{3}\lambda^2 \bigoplus \frac{1}{3}\lambda^3$. ■

Lemma S.3.2. *Suppose* $m \ge 2$, *and* $x_1, \ldots, x_{2m}$ *are real number such that* $x_1 \le \ldots \le x_{2m}$. *Then, for*

any $\pi \in \Pi_n$,

$$\sum_{k=1}^{m} x_{\pi(2k-1)} x_{\pi(2k)} \leq \sum_{k=1}^{m} x_{2k-1} x_{2k} \,. \tag{S.16}$$

PROOF OF LEMMA S.3.2. We need only consider the case where there exists $k_1 < k_2 < k_3 < k_4$ such that at least one of $\pi(k_1), \pi(k_2)$ is greater than at least one of $\pi(k_3), \pi(k_4)$ because the lemma trivially holds otherwise. Suppose without loss of generality that $\pi(k_2) < \pi(k_3) < \pi(k_4) < \pi(k_1)$, then it is easy to verify that

$$x_{\pi(k_1)} x_{\pi(k_2)} + x_{\pi(k_3)} x_{\pi(k_4)} \leq x_{\pi(k_2)} x_{\pi(k_3)} + x_{\pi(k_1)} x_{\pi(k_4)}$$

so that by interchanging two indices we decrease the sum weakly. A finite number of those interchanges maps π back to the identity operator, and hence (S.16) holds. ■

Lemma S.3.3. *Let X_n, Y_n, Z_n be random variables. Suppose $Y_n = g(Z_n) \xrightarrow{d} Y$ as $n \to \infty$, where $g : \mathbf{R} \to \mathbf{R}$ is measurable and $X_n \xrightarrow{d} X$ conditional on Z_n, with probability one for Z_n. Furthermore, suppose the distributions of both X and Y are continuous everywhere. Then*

$$(X_n, Y_n) \xrightarrow{d} (X, Y) \,,$$

where $X \perp\!\!\!\perp Y$.

PROOF OF LEMMA S.3.3. Since X and Y both have continuous distribution function, we need only show for any $x, y \in \mathbf{R}$,

$$P\{X_n \leq x, Y_n \leq y\} \to P\{X \leq x\} P\{Y \leq y\} \,.$$

To this end, note that

$$\begin{aligned}
&P\{X_n \leq x, Y_n \leq y\} - P\{X \leq x\} P\{Y \leq y\} \\
&= E[E[I\{X_n \leq x\} I\{Y_n \leq y\} | Z_n]] - P\{X \leq x\} P\{Y \leq y\} \\
&= E[E[I\{X_n \leq x\} | Z_n] I\{Y_n \leq y\}] - P\{X \leq x\} P\{Y \leq y\} \\
&= E[(E[I\{X_n \leq x\} | Z_n] - P\{X \leq x\}) I\{Y_n \leq y\}] + E[P\{X \leq x\}(I\{Y_n \leq y\} - P\{Y \leq y\})] \\
&= E[(P\{X_n \leq x | Z_n\} - P\{X \leq x\}) I\{Y_n \leq y\}] + (P\{Y_n \leq y\} - P\{Y \leq y\}) P\{X \leq x\}
\end{aligned}$$

For the first term on the right-hand side, note that

$$P\{X_n \leq x | Z_n\} - P\{X \leq x\} \to 0$$

with probability one for Z_n, and hence the product inside the expectation converges to 0 with probability one as well, which in turn implies the expectation converges to 0 by the dominated convergence theorem since probabilities are bounded. The second term converges to 0 because of the definition of convergence in distribution and the fact that the distribution of Y has no discontinuity. ■

Lemma S.3.4. *Suppose the sample size is kn for $k \in \mathbb{Z}$ and the treatment assignment scheme satisfies $\tau_s \equiv \tau = \frac{l}{k}$, where $l \in \mathbb{Z}$, $0 < l < k$, and they are relatively prime. Suppose Q satisfies Assumption 5.1 and h satisfies Assumption S.2.1. Then, under $\lambda^{\tau,h}(X^{(n)})$ defined in (S.12), as $n \to \infty$,*

$$\sqrt{kn}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\tau,h}) \,,$$

where

$$\varsigma^2_{\tau,h} = \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \tau(1-\tau)E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] . \tag{S.17}$$

PROOF OF LEMMA S.3.4. To begin with, note that

$$\sqrt{kn}(\hat{\theta}_n - \theta(Q)) = A_n - B_n + C_n - D_n \,,$$

where

$$\begin{aligned}
A_n &= \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} \left(\frac{Y_i(1)D_i}{\tau} - E\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right]\right) \\
B_n &= \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} \left(\frac{Y_i(0)(1-D_i)}{1-\tau} - E\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right]\right) \\
C_n &= \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} \left(E\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right] - E[Y_i(1)]\right) \\
D_n &= \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} \left(E\left[\frac{Y_i(0)(1-D_i)}{1-\tau}\middle| h^{(n)}, D^{(n)}\right] - E[Y_i(0)]\right) .
\end{aligned}$$

Note that, conditional on $h^{(n)}$ and $D^{(n)}$, A_n and B_n are independent and C_n and D_n are constant.

We first study the limiting behavior of A_n. Conditional on $h^{(n)}$ and $D^{(n)}$, the terms in the sum are independent but not identically distributed. Therefore, we proceed to verify that the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$. To that end, define

$$s_n^2 = s_n^2(h^{(n)}, D^{(n)}) = \sum_{1 \le i \le kn} \mathrm{Var}\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right]$$

and note that

$$\begin{aligned}
s_n^2 &= \sum_{1 \le i \le kn} \mathrm{Var}\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right] \\
&= \frac{1}{\tau^2} \sum_{1 \le i \le kn} D_i \,\mathrm{Var}[Y_i(1)|h^{(n)}] \\
&= \frac{1}{\tau^2} \sum_{1 \le i \le kn : D_i = 1} \mathrm{Var}[Y_i(1)|h(X_i)] \,,
\end{aligned}$$

where the second equality follows from (2) and the third follows from the fact that units are i.i.d. It follows that

$$\tau \frac{s_n^2}{kn} = \frac{1}{kn} \sum_{1 \le i \le kn} \mathrm{Var}[Y_i(1)|h(X_i)] + \left(\frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} \mathrm{Var}[Y_i(1)|h(X_i)] - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} \mathrm{Var}[Y_i(1)|h(X_i)] \right) . \tag{S.18}$$

By Assumption 5.1,

$$\frac{1}{kn} \sum_{1 \le i \le kn} \mathrm{Var}[Y_i(1)|h(X_i)] \xrightarrow{P} E[\mathrm{Var}[Y_i(1)|h(X_i)]] < E[Y_i(1)] < \infty \, . \tag{S.19}$$

Meanwhile,

$$\begin{aligned}
&\left| \frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} \mathrm{Var}[Y_i(1)|h(X_i)] - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} \mathrm{Var}[Y_i(1)|h(X_i)] \right| \\
&\lesssim \left| \frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} h_i - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} h_i \right| \\
&= \frac{1}{\tau kn} \left| \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk : D_{\pi^{\tau,h}(j)} = 1} (h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau) \right| \\
&\le \frac{1}{\tau kn} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk : D_{\pi^{\tau,h}(j)} = 1} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\
&\lesssim \frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\
&\le \left(\frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau|^2 \right)^{1/2} \xrightarrow{P} 0 \, ,
\end{aligned} \tag{S.20}$$

where the first inequality holds by Assumption S.2.1(b), the second holds by using Assumption S.2.1(c), the third holds by inspection, the last holds by the Cauchy-Schwarz inequality, and the equality holds by inspection. Combining (S.18), (S.19), and (S.20), we have

$$\frac{s_n^2}{kn} \xrightarrow{P} \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} > 0 \, , \tag{S.21}$$

where the inequality holds by Assumption S.2.1(a).

We now argue that the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$, i.e., for any $\epsilon > 0$,

$$E_n = \frac{1}{s_n^2 \tau^2} \sum_{1 \le i \le kn} E[|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]|^2 I\{|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\} | h^{(n)}, D^{(n)}] \xrightarrow{P} 0 \, .$$

To this end, first note that for any $M > 0$,

$$P\{\epsilon\tau s_n > M\} \to 1 \tag{S.22}$$

because of (S.21). Next, note that

$$E[Y_i(1)D_i|h^{(n)}, D^{(n)}] = E[Y_i(1)|h(X_i)]D_i$$

because of (2). As a result, for any $M > 0$

$$\begin{aligned}
E_n &= \frac{1}{s_n^2\tau^2} \sum_{1\le i\le kn: D_i=1} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 I\{|Y_i(1) - E[Y_i(1|h^{(n)}, D^{(n)}]| > \epsilon\tau s_n\}|h^{(n)}, D^{(n)}] \\
&\le \frac{1}{s_n^2\tau^2} \sum_{1\le i\le kn} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 I\{|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]| > \epsilon\tau s_n\}|h^{(n)}, D^{(n)}] \\
&\le \frac{1}{s_n^2\tau^2} \sum_{1\le i\le kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)] > M\}|h^{(n)}, D^{(n)}] + o_p(1) \\
&= \frac{kn}{s_n^2\tau^2}\frac{1}{kn} \sum_{1\le i\le kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}|h^{(n)}, D^{(n)}] + o_p(1)
\end{aligned} \tag{S.23}$$

$$\xrightarrow{P} (E[\mathrm{Var}[Y_i(1)|h(X_i)]])^{-1} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] , \tag{S.24}$$

where the first inequality holds by inspection, the second holds because of (S.22) and the equality follows because (2) and $Q_n = Q^{kn}$, and the convergence in probability follows from (S.21) and the fact that Assumption S.2.1(a) implies

$$\begin{aligned}
&E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] \\
&\qquad \le E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2] = E[\mathrm{Var}[Y_i(1)|h(X_i)]] \le E[Y_i^2(1)] < \infty .
\end{aligned}$$

In addition, by the dominated convergence theorem,

$$\lim_{M\to\infty} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] = 0 .$$

To show that $E_n \xrightarrow{P} 0$, fix any subsequence, which we still call $\{n\}$ with some abuse of notation, and we argue that there is a further subsequence along which E_n converges to 0 almost surely. Indeed, for the subsequence $\{n\}$, for any fixed M, the preceding display is bounded by (S.23), which we define as $U_n(M)$. We know from above that $U_n(M) \xrightarrow{P} U(M)$, where $U(M)$ is defined as (S.24). Hence, there exists a further subsequence $\{n\}$ along which $U_n(M) \to U(M)$ almost surely. We then choose a sequence $\{M_n\}_{n\ge1}$ such that $M_n \to \infty$. By the dominated convergence theorem, $\lim_{n\to\infty} U(M_n) = 0$. By a diagonalizing argument, we could construct a further subsequence $\{n\}$ along which $U_n(M_n) \to 0$. Along this subsequence, since $E_n \le U_n(M_n)$ for each n, the almost sure limit of E_n must be zero

because it is non-negative.

We now argue that

$$\sup_{t\in\mathbf{R}}\left|P\{A_n \le t|h^{(n)}, D^{(n)}\} - \Phi\left(t/\sqrt{E[\mathrm{Var}[Y_i(1)|h(X_i)]]/\tau}\right)\right| \xrightarrow{P} 0\,.$$

Fix any subsequence. Since $E_n \xrightarrow{P} 0$, there exists a further subsequence along which $E_n \to 0$ with probability one for $h^{(n)}, D^{(n)}$. Because of the Lindeberg condition and (S.21), it follows that with probability one for $h^{(n)}, D^{(n)}$, $A_n \xrightarrow{d} N(0, E[\mathrm{Var}[Y_i(1)|h(X_i)]]/\tau)$ conditional on $h^{(n)}, D^{(n)}$. But then the left-hand side of the preceding display must converge almost surely to 0 by Pólya's theorem. Since for any subsequence there exists a further subsequence along which it converges to 0 almost surely, it must converge to 0 in probability.

A similar argument establishes that

$$\sup_{t\in\mathbf{R}}\left|P\{B_n \le t|h^{(n)}, D^{(n)}\} - \Phi\left(t/\sqrt{E[\mathrm{Var}[Y_i(0)|h(X_i)]]/(1-\tau)}\right)\right| \xrightarrow{P} 0\,.$$

Since A_n and B_n are independent conditional on $h^{(n)}$ and $D^{(n)}$, it follows by a similar subsequencing argument as above that

$$\sup_{t\in\mathbf{R}}\left|P\{A_n - B_n \le t|h^{(n)}, D^{(n)}\} - \Phi\left(t/\sqrt{E[\mathrm{Var}[Y_i(1)|h(X_i)]/\tau + E[\mathrm{Var}[Y_i(0)|h(X_i)]]/(1-\tau)}\right)\right| \xrightarrow{P} 0\,. \tag{S.25}$$

To study C_n, note that by (2),

$$C_n = \frac{1}{\sqrt{kn}}\sum_{1\le i\le kn}\left(E\left[\frac{Y_i(1)}{\tau}\middle|h(X_i)\right]D_i - E[Y_i(1)]\right)\,.$$

So we have

$$E[C_n|h^{(n)}] = \frac{1}{\sqrt{kn}}\sum_{1\le i\le kn}\left(E[Y_i(1)|h(X_i)] - E[Y_i(1)]\right)\,.$$

Furthermore, by Assumptions S.2.1(b)-(c),

$$\mathrm{Var}[C_n|h^{(n)}] \propto \frac{1}{kn}\sum_{1\le s\le n}(h_{\pi^{\tau,h}(i)} - \bar{h}^s_\tau)^2 \xrightarrow{P} 0\,,$$

where the first relation could be established by repeating the arguments used in the last step of establishing Theorem S.2.1. It therefore follows by Markov's inequality that for any $\epsilon > 0$,

$$P\{|C_n - E[C_n|h^{(n)}]| > \epsilon|h^{(n)}\} \xrightarrow{P} 0\,,$$

and since probabilities are bounded and hence uniformly integrable,

$$P\{|C_n - E[C_n|h^{(n)}]| > \epsilon\} \xrightarrow{P} 0\,,$$

and hence

$$C_n = \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} (E[Y_i(1)|h(X_i)] - E[Y_i(1)]) + o_p(1) \,.$$

A similar proof shows that

$$D_n = \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} (E[Y_i(0)|h(X_i)] - E[Y_i(0)]) + o_p(1) \,.$$

and therefore

$$\begin{aligned} C_n - D_n &= \frac{1}{\sqrt{kn}} \sum_{1 \le i \le kn} (E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)])) + o_p(1) \\ &\xrightarrow{d} N\left(0, E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\right]\right) \,. \end{aligned}$$

We now show by contradiction that

$$\sup_{t \in \mathbf{R}} |P\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \le t\} - \Phi(t/\varsigma_h)| \to 0 \,.$$

Suppose not, then there must exist a subsequence along which the left-hand side of the above display converges to some $\delta > 0$. Along this subsequence, we could find a further subsequence along which the left-hand side of (S.25) converges to 0 with probability one for $h^{(n)}$ and $D^{(n)}$, i.e.,

$$A_n - B_n \xrightarrow{d} N\left(0, \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau}\right)$$

with probability one for $h^{(n)}$ and $D^{(n)}$. Since $C_n - D_n$ is constant for each $h^{(n)}$ and $D^{(n)}$, Lemma S.3.3 establishes that

$$\begin{aligned} A_n - B_n + C_n - D_n \xrightarrow{d} N\Bigg(0, \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau} + \\ E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\right]\Bigg) \,, \end{aligned}$$

which, by Pólya's Theorem, implies a contradiction.

Finally, note that

$$\begin{aligned} &\frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau} + E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\right] \\ &= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \frac{\mathrm{Var}[E[Y_i(1)|h(X_i)]]}{\tau} - \frac{\mathrm{Var}[E[Y_i(0)|h(X_i)]]}{1-\tau} + \\ &\qquad E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\right] \\ &= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \frac{1-\tau}{\tau} E[(E[Y_i(1)|h(X_i)] - E[Y_i(1)])^2] - \frac{\tau}{1-\tau} E[(E[Y_i(0)|h(X_i)] - E[Y_i(0)])^2] \\ &\qquad - 2E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)])(E[Y_i(0)|h(X_i)] - E[Y_i(0)])\right] \end{aligned}$$

$$= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \tau(1-\tau)E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right],$$

and the result follows. ■

Assumption S.3.1. h satisfies

$$\frac{1}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} |h_{\pi^h(4j-k)} - h_{\pi^h(4j-l)}|^2 \xrightarrow{P} 0$$

for $k \in \{2,3\}$ and $l \in \{0,1\}$.

Lemma S.3.5. *Define*

$$\hat{\rho}_n = \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} (Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})$$

and

$$(\hat{\varsigma}_n^h)^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 .$$

Suppose the treatment assignment scheme satisfies Assumption 2.1, Q satisfies Assumption 5.1, and h satisfies Assumptions S.2.1 and S.3.1. Then, under $\lambda^{\frac{1}{2},h}$ defined in (S.12),

$$(\hat{\varsigma}_n^h)^2 \xrightarrow{P} \varsigma^2_{\frac{1}{2},h} .$$

PROOF OF LEMMA S.3.5. To begin with, note that $\hat{\mu}_n(d) \xrightarrow{P} E[Y_i(d)]$ and $\hat{\sigma}_n^2(d) \xrightarrow{P} \mathrm{Var}[Y_i(d)]$ for $d \in \{0,1\}$, by Lemma 6.5 in Bai et al. (2019). Next, we show that

$$E[\hat{\rho}_n | h^{(n)}] \xrightarrow{P} \rho^2 . \tag{S.26}$$

For notational simplicity, we define $\mu_d(h_i) = E[Y_i(d)|h(X_i) = h_i]$ for $d \in \{0,1\}$. To see this, note that

$$\begin{aligned}
&E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}] \\
&= \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&= \frac{1}{4}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j-2)}))(g_h(h_{\pi^h(4j-1)}) + g_h(h_{\pi^h(4j)})) \\
&= \frac{1}{4}\sum_{k\in\{2,3\},l\in\{0,1\}} g_h^2(h_{\pi^h(4j-k)}) + g_h^2(h_{\pi^h(4j-l)}) - (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 .
\end{aligned}$$

As a result,

$$E[\hat{\rho}_n|h^{(n)}] = \sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}]$$
$$= \frac{1}{2n}\sum_{1\le i\le 2n} g_h^2(h(X_i)) - \frac{1}{4n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor}\sum_{k\in\{2,3\},l\in\{0,1\}} (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 .$$

(S.26) then follows from Assumption S.2.1(b), S.3.1, the fact that

$$E[g_h^2(h(X_i))] \lesssim E[E[Y_i(1)|h(X_i)]^2] + E[E[Y_i(0)|h(X_i)]^2]$$
$$\le E[E[Y_i^2(1)|h(X_i)]] + E[E[Y_i^2(0)|h(X_i)]] = E[Y_i^2(1) + Y_i^2(0)] < \infty$$

because of Assumption 5.1, and an application of the WLLN.

It remains to show $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$. We will prove

$$\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} (Y_{\pi^h(4j-2)}Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]) \xrightarrow{P} 0 ,$$

and the others follow similarly. We will repeatedly use the following elementary inequalities for any $a, b \in \mathbf{R}$ and $\lambda > 0$:

$$|a+b|I\{|a+b| > \lambda\} \le 2|a|I\{|a| > \lambda/2\} + 2|b|I\{|b| > \lambda/2\}$$
$$|ab|I\{|ab| > \lambda\} \le |a|^2 I\{|a| > \sqrt{\lambda}\} + |b|^2 I\{|b| > \sqrt{\lambda}\} .$$

To begin with,

$$E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}] = \frac{1}{2}\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) + \frac{1}{2}\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})$$

Next, note that

$$\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]|I\{|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]| > \lambda\}|h^{(n)}]$$
$$\le \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|I\{|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}]$$
$$+ E[|E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]|I\{|E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]| > \sqrt{\lambda/2}\}|h^{(n)}]$$
$$\le \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[Y_{\pi^h(4j-2)}^2 I\{|Y_{\pi^h(4j-2)}| > \sqrt{\lambda/2}\}|h^{(n)}] + E[Y_{\pi^h(4j)}^2 I\{|Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}]$$
$$+ |\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)})|I\{\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) > \lambda/2\}$$
$$+ |\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})|I\{\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)}) > \lambda/2\}$$
$$\le \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[Y_{\pi^h(4j-2)}^2(1) I\{|Y_{\pi^h(4j-2)}(1)| > \sqrt{\lambda/2}\}|h^{(n)}] + E[Y_{\pi^h(4j-2)}^2(0) I\{|Y_{\pi^h(4j-2)}(0)| > \sqrt{\lambda/2}\}|h^{(n)}]$$

$$
\begin{aligned}
&+ E[Y^2_{\pi^h(4j)}(1)I\{|Y_{\pi^h(4j)}(1)| > \sqrt{\lambda/2}\}|h^{(n)}] + E[Y^2_{\pi^h(4j)}(0)I\{|Y_{\pi^h(4j)}(0)| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&+ \mu_1^2(h_{\pi^h(4j-2)})I\{|\mu_1(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j)})I\{|\mu_0(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} \\
&+ \mu_1^2(h_{\pi^h(4j)})I\{|\mu_1(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j-2)})I\{|\mu_0(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} \\
\lesssim \frac{1}{2n} &\sum_{1\le i\le 2n} E[Y_i^2(1)I\{|Y_i(1) > \sqrt{\lambda/2}|\}|h(X_i)] + E[Y_i^2(0)I\{|Y_i(1) > \sqrt{\lambda/2}|\}|h(X_i)] \\
&+ E[Y_i^2(1)|h(X_i)]I\{E[Y_i^2(1)|h(X_i)] > \sqrt{\lambda/2}\} + E[Y_i^2(0)|h(X_i)]I\{E[Y_i^2(0)|h(X_i)] > \sqrt{\lambda/2}\} \\
\xrightarrow{P} &E[Y_i^2(1)I\{|Y_i(1) > \sqrt{\lambda/2}|\}] + E[Y_i^2(0)I\{|Y_i(1) > \sqrt{\lambda/2}|\}] + E[E[Y_i^2(1)|h(X_i)]I\{E[Y_i^2(1)|h(X_i)] > \sqrt{\lambda/2}\}] \\
&+ E[E[Y_i^2(0)|h(X_i)]I\{E[Y_i^2(0)|h(X_i)] > \sqrt{\lambda/2}\}]\,, \qquad \text{(S.27)}
\end{aligned}
$$

where the last line follows from WLLN and the law of iterated expectation. Since by Assumption 5.1 we have $E[Y_i^2(d)] < \infty$ and hence $E[E[Y_i(d)|h(X_i)]^2] < E[Y_i^2(d)]$ by Jensen's inequality, the limit as $\lambda \to \infty$ of the last line is 0, by the dominated convergence theorem. We finish the proof by arguing by contradiction. Suppose

$$\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]$$

does not converge in probability to 0. There must then exist $\epsilon > 0$ and $\delta > 0$ and a subsequence, which for simplicity we again denote by $\{n\}$, such that

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon\} \to \delta \qquad \text{(S.28)}$$

along this subsequence. But because of (S.27), there exists a further subsequence along which the condition in Lemma 6.3 of Bai et al. (2019) holds with probability one for $h^{(n)}$, but then along this subsequence $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$ conditional on $h^{(n)}$ with probability one for $h^{(n)}$, i.e., for any $\epsilon > 0$, with probability one for $h^{(n)}$,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon|h^{(n)}\} \to 0\,.$$

Since probabilities are bounded and hence uniformly integrable,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon\} \to 0$$

along the chosen subsequence, which implies a contradiction to (S.28). ■

Lemma S.3.6. *Suppose U_i, $1 \le i \le n$ are i.i.d. random variables where $E|U_i|^r < \infty$. Then*

$$n^{-1/r} \max_{1\le i\le n} |U_i| \xrightarrow{P} 0\,.$$

PROOF OF LEMMA S.3.6. Note that for all $\epsilon > 0$,

$$P\left\{n^{-1/r} \max_{1\le i\le n} |U_i| > \epsilon\right\} = P\left\{\max_{1\le i\le n} |U_i|^r > n\epsilon^r\right\}$$

$$\leq nP\{|U_i|^r > n\epsilon^r\} \leq \frac{n}{n\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] = \frac{1}{\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] \to 0\,,$$

where the convergence follows because of the dominated convergence theorem and that $E|U_i|^r < \infty$. ■

Lemma S.3.7. *Suppose $E[h^2(X_i)] < \infty$. Then Assumptions S.2.1(c) and S.3.1 hold.*

PROOF OF LEMMA S.3.7. We prove the case where $\tau = \frac{1}{2}$ and the results follow similarly for any $\tau \in (0,1)$. Note that

$$\sum_{1\leq s\leq n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \leq |h_{\pi^h(2n)} - h_{\pi^h(1)}|^2 \leq 4 \max_{1\leq i\leq 2n} h^2(X_i)\,,$$

where the first inequality follows from the definition of π^h and the second inequality follows by inspection, and therefore it follows from Lemma S.3.6 that

$$\frac{1}{n}\sum_{1\leq s\leq n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \leq \frac{4}{n} \max_{1\leq i\leq 2n} h^2(X_i) \xrightarrow{P} 0\,.$$

Assumption S.2.1(c) thus holds. To see Assumption S.3.1 holds, note that

$$\frac{1}{n}\sum_{1\leq j\leq \lfloor \frac{n}{2}\rfloor} |h_{\pi^h(4j-k)} - h_{\pi^h(4j-l)}|^2 \lesssim \frac{1}{n}|h_{\pi^h(2n)} - h_{\pi^h(1)}|^2\,,$$

and the result follows similarly as above. ■

Lemma S.3.8. *Suppose g satisfies Assumption 5.2(c) and $\hat{g}_m$ satisfies Assumption 5.3. Then, as $m, n \to \infty$,*

$$\frac{1}{n}\sum_{1\leq s\leq n} |g_{\pi^{\hat{g}_m}(2s-1)} - g_{\pi^{\hat{g}_m}(2s)}|^2 \xrightarrow{P} 0\,,$$

and

$$\frac{1}{n}\sum_{1\leq j\leq \lfloor \frac{n}{2}\rfloor} |g_{\pi^{\hat{g}_m}(4j-k)} - g_{\pi^{\hat{g}_m}(4j-l)}|^2 \xrightarrow{P} 0$$

for $k \in \{2,3\}$ and $l \in \{0,1\}$.

PROOF OF LEMMA S.3.8. We only prove the first conclusion as the second could be shown by similar arguments. We first show that Assumption 5.3 implies

$$\frac{1}{n}\sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \xrightarrow{P} 0\,. \tag{S.29}$$

Suppose Assumption 5.3 holds. For any $\epsilon > 0$, $\delta > 0$, there exists $M > 0$ such that for $m > M$,

$$P\left\{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2\, Q_X(dx) > \frac{\epsilon\delta}{2}\right\} \leq \frac{\delta}{2}\,. \tag{S.30}$$

By Markov's inequality again, if

$$\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) \leq \frac{\epsilon\delta}{2} \,,$$

then by the independence of $\tilde{W}^{(m)}$ and $W^{(n)}$,

$$P\left\{ \frac{1}{2n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon \middle| \tilde{W}^{(m)} \right\} \leq \frac{E\left[\frac{1}{2n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \middle| \tilde{W}^{(m)} \right]}{\epsilon} = \frac{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx)}{\epsilon} \leq \frac{\delta}{2} \,. \tag{S.31}$$

Then,

$$\begin{aligned} P\left\{ \frac{1}{2n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon \right\} &\leq P\left\{ \frac{1}{2n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon \middle| \tilde{W}^{(m)} \right\} P\left\{ \int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) \leq \frac{\epsilon\delta}{2} \right\} \\ &\quad + P\left\{ \int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) > \frac{\epsilon\delta}{2} \right\} \\ &\leq \frac{\delta}{2}\left(1 - \frac{\delta}{2}\right) + \frac{\delta}{2} \leq \delta \,, \end{aligned}$$

where the first inequality follows by definition, and the second inequality follows from (S.30) and (S.31).

Next, note that since $|a+b|^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbf{R}$,

$$\begin{aligned} &\frac{1}{n} \sum_{1\leq s\leq n} |g_{\pi^{\hat{g}_m}(2s-1)} - g_{\pi^{\hat{g}_m}(2s)}|^2 \\ &\lesssim \frac{1}{n} \sum_{1\leq s\leq n} |\hat{g}_{\pi^{\hat{g}_m}(2s-1)} - \hat{g}_{\pi^{\hat{g}_m}(2s)}|^2 + \frac{1}{n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \,. \end{aligned} \tag{S.32}$$

Next, note that

$$\begin{aligned} &\frac{1}{n} \sum_{1\leq s\leq n} |\hat{g}_{\pi^{\hat{g}_m}(2s-1)} - \hat{g}_{\pi^{\hat{g}_m}(2s)}|^2 \\ &\leq \frac{1}{n} \max_{1\leq i\leq 2n} |\hat{g}_i|^2 \\ &\lesssim \frac{1}{n} \max_{1\leq i\leq 2n} |g_i|^2 + \frac{1}{n} \max_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \\ &\lesssim \frac{1}{n} \max_{1\leq i\leq 2n} |g_i|^2 + \frac{1}{n} \sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \,. \end{aligned} \tag{S.33}$$

The conclusion then follows from (S.29), (S.32), (S.33), Assumption 5.2(c) and an application of Lemma S.3.6. ■

S.3.1 Sufficient conditions for Lipschitz continuity

Let f denote the density function of X. Recall that $C^{(r)}$ is the class of functions which are rth continuously differentiable. We impose the following assumption on h in Assumption S.2.1 and f.

Assumption S.3.2. The function h and density function f satisfy the following conditions.

(a) $h \in C^{(2)}$.

(b) $\frac{\partial h(x)}{\partial x_p} \neq 0$ Lebesgue a.e.

(c) $f \in C^{(2)}$.

Lemma S.3.9 (Theorem 24.4 of Munkres (1997)). *Let O be open in $\mathbf{R}^p$ and $f : O \to \mathbf{R}$ be of class $C^{(r)}$ for $r \geq 1$. Let M be the set of points x for which $f(x) = 0$ and N be the set of points x for which $f(x) \geq 0$. Suppose M is non-empty and $Df(x)$ has rank 1 at each point of M. Then N is a p-manifold in $\mathbf{R}^p$ and $\partial N = M$.*

Lemma S.3.10. *Suppose Assumption S.3.2(a)–(b) hold. Then $M = \{x : h(x) = z\}$ is a $(p-1)$-manifold in $\mathbf{R}^p$.*

PROOF OF LEMMA S.3.10. For each $x \in M$, we aim at providing a coordinate patch on M about x. Indeed, by Assumption S.3.2(a)–(b) and Theorem 9.2 (implicit function theorem) of Munkres (1997), there exists an open set U containing $u = (x_1, \ldots, x_{p-1})$, an open ball $B(z)$ containing z and an open set O in $\mathbf{R}$ containing x_p, and a function $k : U \times B(z) \to \mathbf{R}^p$ of class $C^{(2)}$ such that $h(u, k(u, z')) = z'$ for all $u \in U$, $z' \in B(z)$ and $x \in O$. Moreover, $k(U \times B(z)) = O$. Define the coordinate patch $\alpha(u; z) = (u, k(u, z))$. The conclusion follows by Theorem 5-2 of Spivak (1965). ■

Note that $M = \{x : h(x) = z\}$ is a $(p-1)$-manifold by Lemmas S.3.9 and S.3.10. In what follows, we will need the definition of the integral of a function g over the manifold M. In order to do so, note that there exists a coordinate patch as $\{\alpha_j : U_j \subseteq \mathbf{R}^{p-1} \to V_j \subseteq M, j \in \mathcal{J}\}$, where $\alpha_j(u) = \alpha_j(u, z)$, and each $\alpha_j(u) = (u, k_j(u))$ for some function $k_j : U \to \mathbf{R}$ which is of class C^2, as shown in the proof of Lemma S.3.10, and $\alpha_j(U_j) = V_j$. Next, there exists a partition of unity $\{\phi_i : i \in \mathcal{I}\}$ dominated by the $\{V_j : j \in \mathcal{J}\}$. Moreover, both $\mathcal{I}$ and $\mathcal{J}$ could be chosen to be countable, according to Section 25 of Munkres (1997). The integral of a scalar function g over the manifold is written as

$$\int_M g \, \mathrm{d}V = \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}} \int_{U_j} [(g\phi_i) \circ \alpha_j] V(D\alpha_j) \,,$$

where $V(A) = \sqrt{\det(A'A)}$ is the volume. We have

$$D\alpha_j = \begin{bmatrix} I_{p-1} & \dfrac{\partial k_j(u, z)}{\partial u} \end{bmatrix} \,,$$

so that

$$V(D\alpha_j) = \sqrt{1 + \frac{\partial k_j(u,z)}{\partial u'}\frac{\partial k_j(u,z)}{\partial u}} = \frac{\|\nabla h(u, k_j(u,z))\|}{|D_p h(u, k_j(u,z))|},$$

where $D_p = \frac{\partial}{\partial x_p}$, by the implicit function theorem and matrix determinant lemma. Note that on one hand, for each $j \in \mathcal{J}$, only a finite number of ϕ_i is positive, and on the other hand, $\{\phi_i : i \in \mathcal{I}\}$ is dominated by the coordinate patch, which means that each ϕ_i is supported on a compact set inside a single V_j. As a result, the order of the above double sum could be interchanged.

By p.345 of Bogachev (2007), the conditional expectation of a function g on the manifold M is defined as

$$E[g(X)|M] = \lim_{t\to 0} \frac{E[g(X)I\{z \le h(X) \le z+t\}]}{P\{z \le h(X) \le z+t\}}.$$

Lemma S.3.11. *Suppose Assumption S.3.2(a)-(c) hold. Then*

$$E[g(X)|M] = \frac{\int_M \frac{fg}{\|\nabla h\|}\,\mathrm{d}V}{\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V}. \tag{S.34}$$

For a continuously differentiable function $h : \mathbf{R}^p \to \mathbf{R}$, $x \in \mathbf{R}^p$ is a critical point of h if $\nabla h(x) = 0$, where $\nabla h(x)$ is the gradient of h at x; otherwise x is a regular point of h. A value z is a critical value of h if the set $\{x : h(x) = z\}$ contains at least one critical point; otherwise z is a regular value of h.

PROOF OF LEMMA S.3.11. By L'Hospital's rule,

$$E[g(X)|M] = \frac{\lim_{t\to 0} \frac{E[g(X)I\{z \le h(X) \le z+t\}]}{t}}{\lim_{t\to 0} \frac{P\{z \le h(X) \le z+t\}}{t}},$$

and the lemma follows from Lemma A.1 of Chernozhukov et al. (2018). In particular, the denominator equals the one in (S.34) directly by that lemma, while for the numerator we merely need to redefine the 'density' function as fg and the same proof goes through. ■

Lemma S.3.12. *Suppose Assumption S.3.2(a)-(b) hold. Let $M = \{x : h(x) = z\}$, where z is a regular value of h on $\mathbf{R}^p$. Then for any $g \in C^{(2)}$,*

$$\frac{\partial}{\partial z}\int_M g\,\mathrm{d}V = \int_M \frac{D_p g}{D_p h}\,\mathrm{d}V + \int_M g\frac{1}{\|\nabla h\|^2}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M g\frac{D_{pp}h}{D_p^2 h}\,\mathrm{d}V. \tag{S.35}$$

PROOF OF LEMMA S.3.12. To begin with, note that

$$\begin{aligned}&\frac{\partial}{\partial z}\int_{U_j}[(g\phi_i)\circ\alpha_j]V(D\alpha_j)\\ &\quad= \int_{U_j} D_p(g\phi_i)\frac{\partial k_j(u,z)}{\partial z}\frac{\|\nabla h\|}{|D_p h|}\end{aligned}$$

$$+\int_{U_j} g\phi_i \frac{|D_p h|}{\|\nabla h\|} \frac{\partial k_j(u,z)}{\partial z} \frac{1}{D_p^4 h} \left(D_p^2 h \sum_{1\le i\le p} D_i h D_{ip} h - D_p h D_{pp} h \sum_{1\le i\le p} D_i^2 h \right) , \quad \text{(S.36)}$$

where $D_{ij}h = \partial_i \partial_j h$ for any function $h \in C^{(2)}$. we have suppressed the arguments of h, being $(u, k_j(u,z))$. Note that it is legitimate to pass diffentiation inside the integral by the dominated convergence theorem. By the Implicit Function Theorem again,

$$\frac{\partial k_j(u,z)}{\partial z} = \frac{1}{D_p h(u, k_j(u,z))} . \quad \text{(S.37)}$$

By Theorem 7.17 of Rudin (1976), we know that $\frac{\partial}{\partial z}\int_M g(x)\,\mathrm{d}V$ is the sum over $i \in \mathcal{I}, j \in \mathcal{J}$ of the two terms in (S.36). Using (S.37), the sum of the first term is

$$\begin{aligned}
&\sum_{j\in\mathcal{J}} \sum_{i\in\mathcal{I}} \int_{U_j} (\phi_i D_p g + g D_p \phi_i) \frac{1}{D_p h} \frac{\|\nabla h\|}{|D_p h|} \\
&= \sum_j \int_{U_j} \frac{D_p g}{D_p h} V(D\alpha_j) \\
&= \int_M \frac{D_p g}{D_p h}\,\mathrm{d}V , \quad \text{(S.38)}
\end{aligned}$$

because $\sum_{i\in\mathcal{I}} \phi_i = 1$ and hence $\sum_{i\in\mathcal{I}} D_p\phi_i = D_p \sum_{i\in\mathcal{I}} \phi_i = 0$. Again, the interchange of differentiation and sum is allowed because the sum is actually over a finite number of terms, by definition of a partition of unity. The sum of the second term is

$$\begin{aligned}
&\sum_{j\in\mathcal{J}} \sum_{i\in\mathcal{I}} \int_{U_j} g\phi_i \frac{|D_p h|}{\|\nabla h\|} \frac{1}{D_p^4 h} \sum_{1\le i\le p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h) \\
&= \sum_{j\in\mathcal{J}} \int_{U_j} g \frac{D_p^2 h}{\|\nabla h\|^2} \frac{1}{D_p^4 h} \sum_{1\le i<p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h) V(D\alpha) \\
&= \int_M g \frac{1}{\|\nabla h\|^2 D_p^2 h} \sum_{1\le i\le p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h)\,\mathrm{d}V \\
&= \int_M g \frac{1}{\|\nabla h\|^2} \sum_{1\le i\le p} \frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M g \frac{D_{pp} h}{D_p^2 h}\,\mathrm{d}V . \quad \text{(S.39)}
\end{aligned}$$

(S.35) now follows from (S.38) and (S.39). ■

Theorem S.3.1. *Suppose Assumption S.3.2 holds. If z is a regular value of h, then*

$$\frac{\partial}{\partial z} E[g(X)|M] = \frac{\int_M \frac{D_p(fg/D_p h)}{\|\nabla h\|}\,\mathrm{d}V \int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V - \int_M \frac{D_p(f/D_p h)}{\|\nabla h\|}\,\mathrm{d}V \int_M \frac{fg}{\|\nabla h\|}\,\mathrm{d}V}{\left[\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V\right]^2} . \quad \text{(S.40)}$$

PROOF OF THEOREM S.3.1. To begin with, replace g in Lemma S.3.12 with $\frac{f}{\|\nabla h\|}$. We then have

$$\begin{aligned}
&\frac{\partial}{\partial z}\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V \\
&= \int_M \frac{\|\nabla h\| D_p f - \dfrac{f\sum_{1\le i\le p} D_i h D_{ip} h}{\|\nabla h\|}}{\|\nabla h\|^2 D_p h}\,\mathrm{d}V \\
&\quad + \int_M \frac{f}{\|\nabla h\|^3}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M \frac{f D_{pp} h}{\|\nabla h\| D_p^2 h}\,\mathrm{d}V \\
&= \int_M \frac{D_p f D_p h - f D_{pp} h}{\|\nabla h\| D_p^2 h}\,\mathrm{d}V \\
&= \int_M \frac{D_p(f/D_p h)}{\|\nabla h\|}\,\mathrm{d}V\,. \qquad \text{(S.41)}
\end{aligned}$$

By the same arguments,

$$\frac{\partial}{\partial z}\int_M \frac{fg}{\|\nabla h\|}\,\mathrm{d}V = \int_M \frac{D_p(fg/D_p h)}{\|\nabla h\|}\,\mathrm{d}V\,. \qquad \text{(S.42)}$$

(S.40) now follows from (S.41) and (S.42) together with the quotient rule. ■

In general, by the Law of Iterated Expectation

$$E[Y_i^r(d)|h(X)=z] = E[E[Y_i^r(d)|X]|h(X)=z]\,.$$

Suppose h and the density function of X, $f(X)$ satisfy the smoothness conditions in Assumption S.3.2, the derivative

$$\frac{\partial}{\partial z}E[g(X)|h(X)=z]$$

is given in Theorem S.3.1, where $g(x) = E[Y_i^r(d)|X=x]$ for $r=1,2$ and $d=0,1$. In particular, it is equal to

$$\begin{aligned}
&E\left[\frac{D_p g}{D_p h} + \frac{g D_p f}{f D_p h} - \frac{g D_{pp} h}{D_p^2 h}\middle| h(X)=z\right] - E\left[\frac{D_p f}{f D_p h} - \frac{D_{pp} h}{D_p^2 h}\middle| h(X)=z\right] E\left[g\middle| h(X)=z\right] \\
&= E\left[\frac{D_p g}{D_p h}\middle| h(X)=z\right] + \mathrm{Cov}\left[\frac{D_p f}{f D_p h} - \frac{D_{pp} h}{D_p^2 h}, g\middle| h(X)=z\right]\,. \qquad \text{(S.43)}
\end{aligned}$$

Lemma S.3.13. *Each of the following conditions imply the boundedness of* (S.43).

1. *h is linear, $\|D_p g\|_\infty < \infty$, $\|g\|_\infty < \infty$ and $\|D_p(\ln f)\|_\infty < \infty$.*
2. *h is linear, $\sup_{z\in\mathbf{R}}|E[D_p g|h(X)=z]| < \infty$, $\sup_{z\in\mathbf{R}}|E[g^2|h(X)=z]| < \infty$ and $\sup_{z\in\mathbf{R}}|E[D_p^2(\ln f)|h(X)=z]| < \infty$.*
3. *h includes linear and interaction terms, $\left\|\frac{D_p g}{D_p h}\right\|_\infty < \infty$, $\|g\|_\infty < \infty$ and $\left\|\frac{D_p(\ln f)}{D_p h}\right\|_\infty < \infty$.*

PROOF OF LEMMA S.3.13. Follows from inspection. ■

S.4 Details of penalized matching

In this section, we consider the solution to the Bayesian problem in (33) a particular example that motivates the penalized matching procedure defined by (31). For simplicity, we focus on the special case under which and $Y_i(d) \sim N(X_i'\beta(d), \sigma^2)$ for $d \in \{0,1\}$. Note that the potential outcomes are homoskedastic conditional on the covariates. Define $\beta = \beta(1) + \beta(0)$, and we have $g(x) = x'\beta$. As before, we suppose $\tilde{W}^{(m)} = ((\tilde{Y}_j, \tilde{X}_j', \tilde{D}_j)' : 1 \le j \le m)$ is available from a pilot experiment. Suppose the prior on $\beta(d)$ is $G_d \overset{d}{=} N(\eta(d), \Sigma(d))$ for $d \in \{0,1\}$, being independent across $d \in \{0,1\}$. The prior distribution of β is then $G(d\beta) \overset{d}{=} N(\eta(1)+\eta(0), \Sigma(1)+\Sigma(0))$. We could show that the posterior distribution of $\beta(d)$ conditional on $\tilde{W}^{(m)}$ is

$$\bar{G}_d(d\beta|\tilde{W}^{(m)}) \overset{d}{=} N(\bar{\eta}, \bar{\Sigma}) \,,$$

where for $d \in \{0,1\}$,

$$\bar{\eta}(d) = \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' + \Sigma^{-1}(d) \right)^{-1} \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{Y}_j + \Sigma^{-1}(d)\eta(d) \right)$$
$$\bar{\Sigma}(d) = \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' + \Sigma^{-1}(d) \right)^{-1} .$$

Define $\bar{\eta} = \bar{\eta}(1) + \bar{\eta}(0)$ and $\bar{\Sigma} = \bar{\Sigma}(1) + \bar{\Sigma}(0)$. The posterior distribution for β is

$$\bar{G}(d\beta|\tilde{W}^{(m)}) \overset{d}{=} (\bar{\eta}, \bar{\Sigma}) \,,$$

since $G_d(d\beta)$'s are independent across $d \in \{0,1\}$.

The next lemma provides the solution to the Bayesian problem in (33), where the choice set is over all measurable functions $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$.

Lemma S.4.1. *The solution to* (33) *maps each* $(\tilde{w}^{(m)}, x^{(n)})$ *to* $\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \le s \le n/2\}$, *where* π *solves*

$$\min_{\pi \in \Pi_n} \sum_{1 \le s \le n} \bar{d}\left(x_{\pi(2s-1)}, x_{\pi(2s)}\right) \,,$$

where

$$\bar{d}(x_1, x_2) = (x_1'\bar{\eta} - x_2'\bar{\eta})^2 + (x_1 - x_2)'\bar{\Sigma}(x_1 - x_2) \,. \tag{S.44}$$

PROOF. First note that by (9) and (12), (33) is equivalent to

$$\min_u \iiint L(u(\tilde{w}^{(m)}, x^{(n)})|\beta, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) G(d\beta) \,. \tag{S.45}$$

Next, note that we could solve the problem pointwise for $\tilde{w}^{(m)}$ and $x^{(n)}$ since (S.45) is equivalent to

$$\min_u \bar{R}(u|\tilde{W}^{(m)}) \,, \tag{S.46}$$

where

$$\bar{R}(u|\tilde{W}^{(m)}) = \int L(u(\tilde{W}^{(m)}, x^{(n)})|\beta, x^{(n)}) \bar{G}(d\beta|\tilde{W}^{(m)}) \,.$$

To solve (S.46), first note that since $\bar{R}(u|\tilde{W}^{(m)})$ is linear in u, by Lemma S.3.1, it is solved by a matched-pair design. Next,

$$\bar{R}(u|\tilde{W}^{(m)}) = \sum_{1 \le s \le n} ((x'_{\pi(2s-1)}\bar{\eta} - x'_{\pi(2s)}\bar{\eta})^2 + (x_{\pi(2s-1)} - x_{\pi(2s)})'\bar{\Sigma}(x_{\pi(2s-1)} - x_{\pi(2s)})) \,.$$

As a result, minimizing it is equivalent to minimizing the sum of the distances defined in (S.44). ■

Finally, we want the prior to be irrelevant. For the purpose, suppose that $\Sigma = cI$ where I is an identity matrix. We let the constant $c \to \infty$, so that the prior diverges to a diffuse (uninformative) one. Then, $\bar{\eta}(d)$ converges to $\hat{\beta}_m(d)$ in (24) and $\bar{\Sigma}(d)$ converges to $\hat{\Sigma}_m(d)$ defined in (25). Therefore, we define $\hat{\beta}_m$ as in (26) and $\hat{\Sigma}_m$ as in (27). The metric (S.44) converges to the metric defined in (32).

S.5 Minimax matching

This section describes the minimax procedure in detail. First note that $L(\lambda|h, X^{(n)})$ depends on h only through $h^{(n)}$, and hence (49) is equivalent to

$$\min_{\lambda \in \Lambda} \max_{h^{(n)} \in G} L(\lambda|h^{(n)}) \,, \tag{S.47}$$

where

$$L(\lambda|h^{(n)}) = L(\lambda|h, X^{(n)})$$

and

$$G = \{h^{(n)} : h \in \mathcal{G}, h_1 = 0\} \,.$$

The restriction $h_1 = 0$ is a location normalization, since $L(\lambda|h^{(n)})$ only depends on $h^{(n)}$ through pairwise differences and is therefore shift-invariant. In order to solve (S.47) computationally, we impose the following requirement on G:

Assumption S.5.1. G is a bounded polyhedron in $\mathbf{R}^n$.

We now provide examples of G that satisfy Assumption S.5.1.

Example S.5.1. Consider the class of Lipschitz functions:

$$G = \{h^{(n)} : |h_i - h_j| \leq M\|X_i - X_j\| \text{ for } i \neq j,\ h_1 = 0\}\ . \tag{S.48}$$

G satisfies Assumption S.5.1. ■

Example S.5.2. When $p > 2$, i.e., X_i is multivariate, consider the class of functions which are Lipschitz along each dimension:

$$G = \left\{ h^{(n)} : |h_i - h_j| \leq \sum_{1 \leq l \leq p} M_l |X_{il} - X_{jl}| \text{ for } i \neq j,\ h_1 = 0 \right\}\ .$$

G satisfies Assumption S.5.1. ■

Example S.5.3. Consider the class of functions Lipschitz in a known index. For a known function w, define

$$G = \left\{ h^{(n)} : |h_i - h_j| \leq M|\nu(X_i) - \nu(X_j)| \text{ for } i \neq j,\ h_1 = 0 \right\}\ . \tag{S.49}$$

G satisfies Assumption S.5.1. ■

Example S.5.4. Consider the class of linear functions with coefficients in a bounded polyhedron. For a bounded polyhedron $\mathcal{B}$ in $\mathbf{R}^p$, define

$$G = \{X^{(n)}\beta - X_1'\beta \mathbf{1}_n : \beta \in \mathcal{B}\}\ .$$

G satisfies Assumption S.5.1. ■

Example S.5.5. Consider the class of monotonically increasing functions. Without loss of generality assume that $X_1 \leq \ldots \leq X_n$. For $M > 0$, define

$$G = \{h^{(n)} : h_i \leq h_j \text{ for } i < j, h_n \leq M, h_1 = 0\}\ .$$

Since G is bounded and defined by linear inequalities, it satisfies Assumption S.5.1. ■

Example S.5.6. Consider the class of convex functions. Without loss of generality assume that $X_1 \leq \ldots \leq X_n$. For $M > 0$, define

$$G = \left\{ h^{(n)} : h_i \leq \frac{X_{i+1} - X_i}{X_{i+1} - X_{i-1}} h_{i-1} + \frac{X_i - X_{i-1}}{X_{i+1} - X_{i-1}} h_{i+1}, 2 \leq i \leq 2n-1, |h_n| \leq M, h_1 = 0 \right\}\ .$$

Since G is bounded and defined by linear inequalities, it satisfies Assumption S.5.1. ■

Consider the minimax problem (S.47) with G defined in (S.49). The following theorem shows that without any information of how the covariate affects potential outcomes beyond the index, the best we could do is to match on the index itself.

Theorem S.5.1. *The solution to* (S.47) *with* G *defined in* (S.49) *is* $\lambda^\nu = \{\{\pi^\nu(2s-1), \pi^\nu(2s)\} : 1 \le s \le n\}$ *where* $\nu_{\pi^\nu(1)} \le \ldots \le \nu_{\pi^\nu(2n)}$.

PROOF OF THEOREM S.5.1. Without loss of generality, consider $p = 1$ and $\nu(x) = x$. The general case is proved in exactly the same way. We use another expression of (48). Define $\Delta_i = g_{\pi(i+1)} - g_{\pi(i)}$ for $i = 1, \ldots, 2n-1$. For $\lambda^0 = \{\{1, \ldots, 2n\}\}$,

$$\begin{aligned}
&L(\lambda^0 | g, X^{(n)}) \\
&= \frac{1}{2n(2n-1)} \sum_{1 \le i \le 2n} \left[(2n-1) g_i - \sum_{j \ne i} g_j \right]^2 \\
&= \frac{1}{2n(2n-1)} \sum_{1 \le i \le 2n} \left[- \sum_{1 \le j \le i-1} j \Delta_j + \sum_{i \le j \le 2n-1} (2n-j) \Delta_j \right]^2 \\
&= \frac{1}{2n(2n-1)} \left[\sum_{1 \le i \le 2n-1} 2n(2n-i) i \Delta_i^2 + 2 \sum_{k < l \le 2n-1} 2n(2n-l) k \Delta_k \Delta_l \right] \\
&= \frac{1}{2n-1} \left[\sum_{1 \le i \le 2n-1} (2n-i) i \Delta_i^2 + 2 \sum_{k < l \le 2n-1} (2n-l) k \Delta_k \Delta_l \right] .
\end{aligned}$$

As a result, for a general stratification λ, the loss function (48) equals

$$L(\lambda | g, X^{(n)}) = \sum_{1 \le s \le S} \frac{1}{n_s - 1} \left[\sum_{1 \le i \le n_s - 1} (n_s - i) i \Delta_{i,s}^2 + 2 \sum_{k < l \le n_s - 1} (n_s - l) k \Delta_{k,s} \Delta_{l,s} \right] . \tag{S.50}$$

Note that $g^{\mathrm{mm}}(x) = Mx$ simultaneously maximizes (S.50) for every λ. But we know the stratification that solves

$$\min_{\lambda \in \Lambda} L(\lambda | g^{\mathrm{mm}}, X^{(n)})$$

is the "optimal non-bipartite matching" of X on $\mathbf{R}$, i.e. λ^x. ■

For a prespecified $\theta_0 \in \mathbf{R}$, consider the problem of testing (35) at level $\alpha \in (0,1)$. We use the test in (38) by setting $\hat{g}_m = \nu$.

Corollary S.5.1. *Suppose the treatment assignment scheme satisfies Assumption 2.1 and* Q *satisfies Assumption 5.1 and* $h = \nu$ *satisfies Assumption S.2.1 with* $\tau = \frac{1}{2}$. *Then, for the problem of testing* (35) *at level* $\alpha \in (0,1)$, ϕ_n^ν *satisfies*

$$\lim_{n \to \infty} E[\phi_n^\nu(W^{(n)})] = \alpha ,$$

whenever Q *additionally satisfies the null hypothesis, i.e.,* $\theta(Q) = \theta_0$.

For other specifications of G in (S.47), there does not exist a clean result as Theorem S.5.1, as illustrated by the following example.

Example S.5.7. Let $n = 4$ and $X_1 = (0,0)', X_2 = (1,0)', X_3 = (0,1)', X_4 = (1,1)'$. Let $n = 4$ and define

$$\begin{aligned}\lambda^0 &= \{\{1,2,3,4\}\}\\ \lambda^1 &= \{\{1,2\},\{3,4\}\}\\ \lambda^2 &= \{\{1,3\},\{2,4\}\}\\ \lambda^3 &= \{\{1,4\},\{2,3\}\}\,.\end{aligned}$$

Let G be as defined in (S.48) with $M = 1$. Then λ^0 solves (S.47). Indeed, for $\lambda = \lambda^1$, the worst case occurs at $h^{(n)} = (0, \sqrt{2}-1, \sqrt{2}-1, \sqrt{2})$, with the loss equal to 2. For $\lambda = \lambda^2$ or λ^3, the worst case occurs at $h^{(n)} = (0,1,1,0)$, with the loss equal to 2. In contrast, the worst case for $\lambda = \lambda^0$ occurs at $h^{(n)} = (0, \sqrt{2}-1, \sqrt{2}-1, \sqrt{2})$, and the loss is $(10 - 4\sqrt{2})/3 < 2$. ■

The key reason why (S.47) is hard to solve when $p > 1$ is that the choice set Λ is not convex. In principle, we could convexify the problem by considering the $\mathrm{co}(\Lambda)$, the convex hull of Λ. That amounts to allowing for mixing over (potentially a large number of) matched-pair designs, which is hard to interpret and is almost never used in practice. Although Λ is not convex, we can still provide computational strategies to solve (S.47). Note that $L(\lambda|h^{(n)})$ is convex in $h^{(n)}$, which combined with Assumption S.5.1 implies that the inner maximum in (S.47) is attained on the vertices of G, which we denote by V. Then, the minimax problem is equivalent to

$$\min_{\lambda\in\Lambda}\max_{h^{(n)}\in V} L(\lambda|h^{(n)})\,. \tag{S.51}$$

We now apply results from graph theory to reformulate (S.51) into Mixed Integer Linear Programs (MILPs). We first recall some definitions from the graph theory and connect them to the optimal stratification problem. For more details, see Bertsimas and Tsitsiklis (1997).

An undirected graph $\Gamma = (N, E)$ consists of a set of nodes N and a set of edges E. Each element of E is an unordered pair $\{i,j\}$ where $i \in N$ and $j \in N$. Define $q_e = 1$ if $e \in E$ and define $\mathbf{q} = (q_e)_{e\in E}$. Define $q_{ij} = q_{i,j}$. The degree of i is defined as $d_i = \sum_j q_{ij}$. The graph Γ is complete if $q_{ij} = 1$ for all $i \neq j$. A subset U of N is a clique in Γ if $\{i,j\} \in E$ for all $i,j \in U$. The set of induced edges by U is $E(U) = \{\{i,j\} \in E : i,j \in U, i \neq j\}$. A clique partition of Γ is $\Gamma^C = (N, E(U_1,\ldots,U_S))$ for $E(U_1,\ldots,U_S) = \cup_{s=1}^S E(U_s)$ where each U_s is a clique in Γ^C (and Γ), and $\{U_s\}_{s=1}^S$ is a partition of N, i.e., $N = \cup_{s=1}^S U_s$ and $U_s \bigcap U_t = \varnothing$ for $s \neq t$.

In terms of stratification, a unit is a node and an edge $\{i,j\} \in E$ if units i and j are in the same stratum. A stratum is a clique. A stratification $\lambda = \{\lambda_s\}_{s=1}^S$ of $N = \{1,\ldots,n\}$ induces a clique partition $\Gamma^\lambda = (N, E(\lambda_1,\ldots,\lambda_S))$ of $\Gamma = (N,E)$ for $E = \{\{i,j\} : i,j \in N, i \neq j\}$ where the size of each clique λ_s is even, or equivalently the degree of each node in Γ^λ is odd.

Define $c_e = (h_i - h_j)^2$ as the cost of edge $e = \{i,j\} \in E$, $\mathbf{c} = (c_e)_{e\in E}$ and $C = \{\mathbf{c} : h^{(n)} \in V\}$. By

(48),

$$L(\lambda|h^{(n)}) = L(\lambda|h, X^{(n)}) = \sum_{1\le s\le S} \frac{1}{n_s - 1} \sum_{i,j\in\lambda_s, i<j} (h_i - h_j)^2 \,.$$

If $n_s \equiv 2$, then it equals

$$\sum_{e\in E} c_e q_e \,.$$

If $n_s > 2$ for some s, then we need to introduce other binary variables to indicate n_s. The minimax problem (S.47) is equivalent to the following MILP which solves the cost minimization problem over size-bounded stratifications within Λ, i.e., λ with $n_s \le 2K$ for all s.

$$\min_{\mathbf{q}} \quad z \tag{S.52}$$

$$\begin{aligned}
\text{subject to} \quad & \sum_{e\in E} c_e \left(\sum_{1\le k\le K} \frac{u_{ik}}{2k-1} \right) I\{i \in e\} \le z, \text{ for all } \mathbf{c} \in C \,, \\
& \sum_{l\in N} q_{il} = \sum_{1\le k\le K} (2k-1) u_{ik}, \text{ for all } i \in N \,, \\
& u_{ik} \in \{0,1\}, \text{ for all } i \in N,\ 1 \le k \le K \,, \\
& q_{e_1} + q_{e_2} - q_{e_3} \le 1, \text{ for all } e_1, e_2, e_3 \in E \,, \qquad \text{(S.53)} \\
& q_e \in \{0,1\}, \text{ for all } e \in E \,.
\end{aligned}$$

We impose an upper bound on the size of each stratum, $2k$. $u_{ik}, k = 1, \ldots, K-1$ are binary indicators of whether the stratum of unit i has size $2k$. The first set of constraints express the loss function (48). The second set of constraints say the degree of each node is $2k-1$, the stratum size minus one. The third set of constraints restrict u_{ik} to be binary. The fourth and the most important set of constraints, (S.53), are called triangle inequalities in the clique partition literature. See Grötschel and Wakabayashi (1990). They ensure that the solution to (S.52) is indeed a clique partition, i.e., a stratification. However, our problem differs from the standard clique partition problem in two ways: we only allow an even number of units within each clique; and the final weights on each edge in the total cost depends on the degrees of either of its nodes, rather than being a constant.

The program (S.52) is computationally intensive even when $k = 2$ and becomes prohibitive quickly as n increases. Therefore, we consider two relaxations of it. The first relaxation is to optimize over Λ^p instead of Λ. For a matched-pair design $\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \le s \le n\}$,

$$L(\lambda|h, X^{(n)}) = \sum_{1\le s\le n} (h_{\pi(2s-1)} - h_{\pi(2s)})^2 \,.$$

As a result, we introduce the program as

$$\min_{\mathbf{q}} \quad z \tag{S.54}$$

$$\text{subject to} \quad \sum_{e\in E} c_e q_e \le z, \text{ for all } \mathbf{c} \in C \,,$$

$$\sum_{j\in N} q_{ij} = 1, \text{ for all } i \in N\,,$$
$$q_e \in \{0,1\}, \text{ for all } e \in E\,.$$

The solution to (S.54) is $\lambda^{\text{mm}} = \{e \in E : q_e = 1\}$. We define the permutation π^{mm} such that $\lambda^{\text{mm}} = \{\{\pi^{\text{mm}}(2s-1), \pi^{\text{mm}}(2s)\} : 1 \le s \le n\}$. (S.54) is feasible even when n is large and requires substantially less computational budget than (S.52). Moreover, as simulation evidence in Section in Table S.1 shows, the solution to (S.54) is frequently the same with (S.52) for a small K and (S.55).

The second relaxation is the following hierarchical procedure.

Algorithm S.5.1.

1. Solve (S.54). Denote the solution by $\mathbf{q}^0$ and denote $\Lambda^0 = \{e \in E : q_e = 1\}$.
2. For $k \ge 0$, repeat steps (a) and (b) below.
   (a) For $\mathbf{q}^k = (q^k_{AB})_{A,B\in\Lambda^k, A\neq B}$, solve

$$\begin{aligned} \min_{\mathbf{q}^k} \quad & z \\ \text{subject to} \quad & \sum_{A,B\in\Lambda^k} q_{AB}c_{AB} + \sum_{A\in\Lambda^k} c_A \le z, \text{ for all } \mathbf{c} \in C\,, \\ & \sum_{B\in\Lambda^k} q^k_{AB} \le 1, \quad \text{for all } A \in \Lambda^k\,, \\ & q^k_{AB} \in \{0,1\}, \text{ for all } A, B \in \Lambda^k\,, \end{aligned} \tag{S.55}$$

   where $c_A = L(\lambda|g, X_A)$, for $X_A = \{X_i : i \in A\}$ and $c_{AB} = c_{A\bigcup B} - c_A - c_B$.
   (b) Update

$$\Lambda^{k+1} = \{A\bigcup B : q^k_{AB} = 1\}\bigcup\{A : \sum_{B\in\Lambda^k} q^k_{AB} = 0\}$$

   until $\Lambda^{k^*} = \Lambda^{k^*+1}$. Collect Λ^{k^*} as the solution.

Algorithm S.5.1 iteratively decides whether to merge pairs of strata or not. The algorithm stops when no paiswise merging of existing strata reduces the worst-case loss.

We now study the properties of minimax matching in a small simulation study. We compare both the actual and worst-case losses under different stratifications. In the following model, we construct a bounded polyhedron G around $g^{(n)}$. We then calculate both the actual losses $L(\lambda|g^{(n)})$ and worst-case losses $\max_{h^{(n)}\in G} L(\lambda|h^{(n)})$ across different stratifications. We set $g(x) = x'\beta$ and

$$G = \{X^{(n)}\beta : \beta \in \mathcal{B}\}\,,$$

where $\mathcal{B}$ is a polyhedron such that $\beta \in \mathcal{B}$.

Model MM $2n = 24$; $p = 2$; $X_{i,1} = 0$ for $1 \leq i \leq 8$, $X_{i,2} = 1$ for $9 \leq i \leq 24$; $X_{i,2} \sim N(0,1)^2$ i.i.d. across i; $g(x) = x'\beta$, $\beta = (1,1)'$; $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2$, $\mathcal{B}_1 = \beta_1 + \gamma_1 \times [0,1]$, $\mathcal{B}_2 = \beta_2 + \gamma_2 \times [-1,1]$; $\gamma \in \{(0.5,0.5)', (2,2)', (0,2)', (2,0)'\}$.

We randomly generate $X^{(n)}$ in 100 replications and summarize

(a) ratios of the values of the actual loss against those under infeasible optimal stratifications.

(b) ratios of the values of the worst-case loss against those under size-bounded minimax stratifications with $k = 2$.

We consider the following stratifications:

Oracle infeasible optimal stratification in (20).

by1 : $\lambda_1 = \{i : 1 \leq i \leq 8\}, \lambda_2 = \{i : 9 \leq i \leq 24\}$.

by2 two strata separated by the sample median of $X_{i,2}$.

2by2 four strata as the cross product of **by1** and **by2**.

MP2 matching on $X_{i,2}$ only, i.e., stratification in (23) with $\hat{g}_m(x) = x_1$.

MPcell within each value of $X_{i,1}$, optimal matched-pair design using $X_{i,2}$.

MMpair the minimax matched-pair design in (S.54).

MMbdd the size-bounded minimax stratification in (S.52) with $k = 2$.

MMhier results from the hierarchical procedure in Algorithm S.5.1.

In Model MM, **MMpair** and **MMhier** have the same solution with **MMbdd** (which we know weakly dominates **MMpair**) most of the time, while other stratifications which do not incorporate minimax consideration sometimes generate much larger worst-case losses.

S.6 AEA RCT Registry

The following experiments in the AEA RCT Registry use matched-pair designs: AEARCTR-0000086, 0000171, 0000293, 0000443, 0000481, 0000550, 0000578, 0000587, 0000644, 0000688, 0000721, 0000983, 0000986, 0001034, 0001097, 0001218, 0001370, 0001591, 0001607, 0001712, 0001714, 0001778, 0001992, 0001995, 0002010, 0002125, 0002132, 0002282, 0002585, 0002622, 0002664, 0002750, 0002776, 0003056, 0003076, 0003524, 0003581, 0003629, 0003648, 0003779, 0003814, 0003933, 0003994, 0004024, 0004042, 0004022, 0006706.

		Oracle	**by1**	**by2**	**2by2**	**MP2**	**MPcell**	**MMpair**	**MMbdd**	**MMhier**
$\gamma = (0.5, 0.5)$	25%	**1.0000**	4.0109	2.3318	1.8158	1.0619	1.0256	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	7.2394	3.8858	2.9807	1.2631	1.5291	1.0000	1.0000	1.0000
	75%	**1.0000**	13.7890	7.7959	6.5012	1.8567	4.4629	1.0001	1.0001	1.0001
	Mean	**1.0000**	13.6242	7.0691	7.6378	1.7480	5.5226	1.0346	1.0346	1.0346
Worst-case	25%	1.0000	4.0109	2.3381	1.7832	1.0481	1.0243	1.0000	**1.0000**	1.0000
	50%	1.0000	6.9420	3.6908	2.8469	1.1858	1.4011	1.0000	**1.0000**	1.0000
	75%	1.0003	11.9445	6.7125	5.9020	1.6146	3.9388	1.0000	**1.0000**	1.0000
	Mean	1.0212	10.3183	5.4894	5.6169	1.4884	3.8240	1.0000	**1.0000**	1.0000
$\gamma = (2, 2)$	25%	**1.0000**	4.4595	2.7007	2.1185	1.0700	1.0994	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	9.2109	4.1580	3.5446	1.3348	1.8127	1.0096	1.0096	1.0096
	75%	**1.0000**	14.5268	6.8864	6.9304	1.8268	4.0986	1.3038	1.3038	1.3038
	Mean	**1.0000**	13.5257	7.3036	6.3795	1.7873	3.8773	1.2997	1.2997	1.2997
Worst-case	25%	1.0000	3.8897	2.2736	1.8008	1.0408	1.0315	1.0000	**1.0000**	1.0000
	50%	1.0126	6.2500	3.1816	2.6923	1.1604	1.4000	1.0000	**1.0000**	1.0000
	75%	1.2516	10.1048	5.0563	4.3542	1.6279	2.8357	1.0000	**1.0000**	1.0000
	Mean	1.2390	8.5778	4.9668	3.9661	1.4436	2.2735	1.0000	**1.0000**	1.0000
$\gamma = (0, 1)$	25%	**1.0000**	4.1720	2.5479	1.8497	1.0397	1.1857	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	7.4458	3.8469	3.3647	1.2599	1.7892	1.0135	1.0135	1.0135
	75%	**1.0000**	14.1891	7.6734	6.3794	1.7666	3.1199	1.1199	1.1199	1.1199
	Mean	**1.0000**	12.4138	6.8864	5.5793	1.8987	2.8784	1.1301	1.1301	1.1301
Worst-case	25%	1.0000	4.3021	2.3348	1.8989	1.0012	1.2292	1.0000	**1.0000**	1.0000
	50%	1.0077	7.2928	3.4658	3.6051	1.0450	1.5861	1.0000	**1.0000**	1.0000
	75%	1.1138	16.6540	6.7290	6.8655	1.2165	3.7622	1.0000	**1.0000**	1.0000
	Mean	1.1128	12.0228	5.8405	5.4142	1.2350	2.8276	1.0000	**1.0000**	1.0000
$\gamma = (1, 0)$	25%	**1.0000**	3.5310	2.1679	2.0152	1.0654	1.0985	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	8.5908	4.1682	3.8322	1.2567	1.9700	1.0481	1.0481	1.0481
	75%	**1.0000**	17.9252	8.6984	8.2448	1.8296	3.6598	1.5850	1.5850	1.5850
	Mean	**1.0000**	14.7115	8.3951	6.6366	1.6191	3.8705	1.7197	1.7197	1.7197
Worst-case	25%	1.0000	2.9528	2.4142	1.5418	1.1470	1.0000	1.0000	**1.0000**	1.0000
	50%	1.0435	4.6975	3.3215	2.1056	1.5634	1.0211	1.0000	**1.0000**	1.0000
	75%	1.6225	9.0650	5.4879	3.9089	2.6225	1.6384	1.0000	**1.0000**	1.0000
	Mean	1.6231	7.8535	6.4319	3.6442	2.3219	1.8804	1.0000	**1.0000**	1.0000

Table S.1: Ratios of values of the actual loss under all stratifications against those under the infeasible optimal stratifications (**Oracle**) and ratios of values of the worst-case loss under all stratifications against those under size-bounded minimax stratifications (**MMbdd**) in Model MM. Benchmarks are displayed in bold face.

References

BAI, Y., SHAIKH, A. and ROMANO, J. P. (2019). Inference in experiments with matched pairs. Working paper.

BERTSIMAS, D. and TSITSIKLIS, J. N. (1997). *Introduction to linear optimization*, vol. 6.

BOGACHEV, V. I. (2007). *Measure theory*. Springer, Berlin–New York.

BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, **113** 1784–1796.

BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, **10** 1747–1785.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and LUO, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, **86** 1911–1938.

GRÖTSCHEL, M. and WAKABAYASHI, Y. (1990). Facets of the clique partitioning polytope. *Mathematical Programming*, **47** 367–387.

MUNKRES, J. R. (1997). *Analysis on manifolds*. Westview Press.

RUDIN, W. (1976). *Principles of mathematical analysis*, vol. 3. McGraw-hill New York.

SPIVAK, M. (1965). *Calculus on manifolds*.

TABORD-MEEHAN, M. (2020). Stratification trees for adaptive randomization in randomized controlled trials. Working paper.