On randomization inference after inexact Mahalanobis matching

Kevin Guo Dominik Rothenhäsler

April 29, 2022

Abstract

In observational causal inference, matched-pairs studies are often analyzed using randomization tests. While intuitively appealing, these tests are not formally justified by the randomness in the treatment assignment process (the “design”) unless all matches are exact. This paper asks whether these tests can instead be justified by the random sampling of experimental units. We find that under fairly restrictive sampling assumptions, the paired randomization test based on a regression-adjusted test statistic may be asymptotically valid despite inexact matching. We propose a new randomization test based on matching with replacement that can be justified under weaker sampling assumptions.

1 Introduction

1.1 Motivation

We consider the problem of using a large observational dataset \( \{(X_i, Y_i, Z_i) : i = 1, \ldots, n\} \) to test whether a binary treatment \( Z_i \in \{0, 1\} \) has any causal effect on an outcome \( Y_i \in \mathbb{R} \). The vector \( X_i \in \mathbb{R}^d \) contains measured confounders whose effects must be controlled away.

A prominent tradition in causal inference posits that this problem should be approached using a combination of matching and randomization inference \([46, 47, 49, 50, 52, 19, 20, 11, 10]\). The simplest version of this strategy can be described as follows:

1. In the design phase, the data analyst controls for the effect of the \( X_i \)'s by matching each treated observation \( i \) with a similar control observation \( m(i) \), thereby simulating a paired randomized experiment. We focus on matching schemes that use the covariates directly.

2. In the analysis phase, the data analyst estimates a causal effect with the matched sample and assesses its statistical significance using the paired Fisher randomization test. In other words, the analysis phase is performed as if the data had actually come from a paired randomized experiment.

When all matches formed in the design phase are exact (i.e. \( X_i = X_{m(i)} \) for all treated observations \( i \)), this strategy provides a highly transparent approach to observational causal inference. Since pair matching perfectly controls for the overt bias due to measured confounders, the paired randomization test delivers a finite-sample exact p-value under no assumptions except that the study is free of hidden bias. Sensitivity to hidden bias can be probed using tools described in Rosenbaum \([52]\).

However, it is often not possible to find an exact match for every treated observation, e.g. if any of the measured covariates has a continuous distribution. In such cases, the analyst must worry about overt bias before turning to hidden bias.

In the methodological literature, it is well known that the standard design-based justification for the paired Fisher randomization test fails in the presence of inexact matching; see Rosenbaum \([47, \text{Section 6}]\) or Pashley et al \([41, \text{Section 5}]\). Indeed, when any of the measured confounders has a continuous distribution, it is known that no nontrivial test with assumption-free validity exists \([61]\). Rosenbaum \([46, 49]\) has proposed more sophisticated randomization tests that regain validity by making parametric assumptions on the propensity score. Nevertheless, it remains popular in practice to analyze inexactly-matched observational studies using the ordinary paired randomization test. This may be motivated by the belief that the resulting inferences will be approximately valid as long as the matches are close enough.
In this article, we assume that the experimental units are independent samples from a population $P$ and ask what needs to be assumed about $P$ to ensure that the paired Fisher randomization test has approximate (asymptotic) validity despite inexact matching. Said another way, we investigate whether there is a sampling-based justification for ignoring inexact matches even though no design-based justification exists. Since permutation tests have been shown to have large-sample validity in a variety of problems where their finite-sample validity fails [31, 12], it is natural to hope the same will hold in paired observational studies.

A second goal of this article is to study a simple modification of the basic matching-plus-randomization-inference strategy that improves trustworthiness in the face of inexact matching. In particular, we propose a new randomization test based on one-nearest-neighbor matching with replacement and investigate its large-sample properties.

1.2 Outline and overview of results

Here, we outline the rest of the paper and summarize our main findings.

The first part of this paper is dedicated to the asymptotic properties of the paired randomization test. Our results concern three aspects of this test:

1) **Bias.** To achieve asymptotic validity, the analyst must base the randomization test on a test statistic that is approximately unbiased under the null hypothesis. Using this fact, we show that the test based on the simple difference-of-means statistic is not generally valid, even under strong smoothness or functional form assumptions. The issue is that test statistic’s well-documented bias [53, 56, 1, 59, 62], which may persist even when all conventional balance tests pass. One way to remove the bias is to make parametric assumptions on the outcome model (e.g. $E[Y|X,Z] = \tau Z + \beta^T X$) and use a regression-adjusted test statistic instead.

2) **Variance.** After removing the bias, the next question is whether the paired randomization test consistently estimates the sampling variance of the test statistic. Surprisingly, we show that the answer depends on which regression-adjusted test statistic is applied. For the ordinary least-squares test statistic, the paired randomization test underestimates the variance unless matching successfully balances covariate means. Meanwhile, for the test statistic proposed by Tukey [63], the paired randomization test overestimates the variance unless matching successfully balances covariate means.

3) **Robustness.** Finally, the data analyst may be interested in whether the paired randomization test is robust to small violations of parametric assumptions. After all, gaining robustness is one of the primary statistical justifications for matching [54, 29]. We formalize this problem in terms of local misspecification [38, 8] and show that the robustness of the paired randomization test depends crucially on the distribution of the propensity score $\epsilon(X) = P(Z = 1 | X)$. If some units have propensity scores larger than one-half, then the test is sensitive to certain small departures from parametric assumptions.

These result show that the sampling properties of the paired randomization test will be satisfactory despite inexact matching when three conditions are satisfied: (i) the analyst uses a regression-adjusted test statistic based on an approximately correctly-specified model; (ii) the matching passes a balance test; (iii) there are no units with propensity scores larger than one-half. Absent any of these conditions, the paired randomization test may perform poorly even if the propensity score and outcome model are very smooth.

Some analysts may find these conditions too strong for their liking. For example, Sävje [62] points out that the condition $P\{\epsilon(X) > 0.5\} = 0$ is far more restrictive than the usual overlap assumption needed for identification, and Hill [28] asks why we should perform matching at all if we are willing to assume a correctly-specified parametric model. While it is possible to perform valid inference under considerably weaker assumptions using semiparametric methods [45] or bias-aware inference [6], these approaches depart from the matching-plus-randomization inference framework and may be less interpretable to some audiences. As Rosenbaum [52] cautions, “If adjustment for observed covariates becomes unnecessarily complex ... then the analysis may never engage the fundamental issue, namely possible biases from covariates that were not measured.”

Thus, the second part of this paper proposes a simple modification of the baseline pair matching procedure that yields asymptotically valid randomization inferences under weaker assumptions. Our goal was to gain
some (but not all) of the robustness of semiparametric methods without sacrificing the interpretability of matching. In particular, we set ourselves the constraint of changing standard practices as little as possible.

Our proposed test replaces pair matching by one-nearest-neighbor matching with replacement. This form of matching has been studied extensively by [1, 2, 3, 5, 40, 36]. However, it has not really caught on in the randomization inference community, possibly because it does not appear to simulate any randomized experiment.

Our main contribution in this part is to argue that matching with replacement can be viewed as simulating a certain stratified experiment with one untreated unit per stratum. Moreover, we show that applying the Fisher randomization test from a stratified experiment yields inferences that are asymptotically valid and robust to local misspecification, even when some units have propensity scores larger than one-half.

1.3 Setting

The setting of this article is the Neyman-Rubin causal model [39, 55] with an infinite superpopulation. We assume that experimental units \((X_i, Y_i(0), Y_i(1), Z_i)\) are independent samples from a common distribution \(P\) and that only \((X_i, Y_i, Z_i)\) is observed where \(Y_i = Y_i(Z_i)\). The problem of interest is to use the observed data to test Fisher’s sharp null hypothesis:

\[ Y_i(0) = Y_i(1) \text{ with probability one.} \tag{1} \]

We let \(H_0\) denote the set of distributions satisfying (1). Although there is growing interest in using randomization tests to study weaker null hypotheses [18, 67, 70], this paper focuses exclusively on the sharp null hypothesis.

Throughout, we will assume that the underlying distribution \(P\) satisfies a few conditions.

**Assumption 1.** The distribution \(P\) satisfies the following:

(a) Unconfoundedness. \(\{Y(0), Y(1)\} \perp Z \mid X\).

(b) Overlap. \(P(Z = 0 \mid X) \geq \delta > 0\).

(c) More controls than treated. \(0 < P(Z = 1) < 0.5\).

(d) Moments. \(\|X\|\text{ and } Y\) have more than four moments.

(e) Nonsingularity. \(\text{Var}(X \mid Z = 1)\) is nonsingular and \(\text{Var}(Y \mid X, Z) > 0\).

The unconfoundedness and overlap conditions are standard identifying assumptions. Meanwhile, the condition that \(0 < P(Z = 1) < 0.5\) ensures that treated observations exist and that it is eventually possible to find an untreated match for each treated observation. The last two conditions are needed for various technical reasons, e.g. to ensure that the Mahalanobis distance exists. They are generally weaker than the technical conditions used by Abadie & Imbens [1, 3] to study matching estimators.

The analysis in this paper is asymptotic. We say that a sequence of p-values \(p \equiv p_n(\{(X_i, Y_i, Z_i)\}_{i \leq n})\) is (pointwise) asymptotically valid at a particular distribution \(P \in H_0\) if (2) holds under independent sampling from \(P\).

\[
\limsup_{n \to \infty} P(p_n < \alpha) \leq \alpha \quad \text{for all } \alpha \in (0, 1) \tag{2}
\]

Various notions of uniform asymptotic validity will be considered in this paper as well. We adopt the classical asymptotic regime where the covariate dimension \(d\) stays fixed as the sample size grows large. In fact, following the advice of Rubin [57], we recommend thinking of \(d\) as a fairly small number (eight or less).

The independent sampling model used in this paper differs from several alternatives previously used in the matching literature. One alternative is the design-only framework which treats \((X_i, Y_i(0), Y_i(1))\) as constants and only models the randomness in \(Z_i\) [49]. However, this randomness does not obviously justify the paired randomization test in the presence of inexact matching so we look to sampling for alternative justifications. Another alternative is the sampling framework in which the number of untreated observations \(N_0\) grows at a faster rate than the number of treated observations \(N_1\). For example, Abadie & Imbens [4], Ferman [15] assume that \(N_0 \gg N_1^{d/2}\). This setting is mathematically favorable for matching, but the sample size requirement is stringent even when \(d = 5, N_1 = 100\). Sävje [62, Section 5] offers further discussion on the practical relevance of this asymptotic regime.
2 Asymptotics of the paired randomization test

2.1 The paired randomization test after optimal matching

In this section, we present our findings on the asymptotic validity of the Fisher randomization test after pair matching. The precise procedure we study is the following. First, matched pairs are formed using optimal Mahalanobis matching, as described in Rosenbaum [48]. For each treated unit $i$, this matching scheme finds a unique untreated unit $m(i)$ in such a way that the total Mahalanobis distance across pairs (3) is minimized.

$$\sum_{Z_i=1} \{(X_i - X_{m(i)})^\top \hat{\Sigma}^{-1}(X_i - X_{m(i)})\}^{1/2}$$

Ties may be broken arbitrarily. In (3), $\hat{\Sigma}$ is the sample covariance matrix of the covariates and we arbitrarily set $\hat{\Sigma}^{-1}$ to be the identity matrix when $\Sigma$ is singular. We also let $\mathcal{M} = \{i : Z_i = 1 \text{ or } i = m(j) \text{ for some treated unit } j\}$ be the set of matched units.

Given a matching $\mathcal{M}$, the paired Fisher randomization test computes a p-value for the sharp null hypothesis (1) in the following fashion: First, a test statistic $\hat{\tau} \equiv \hat{\tau}(\{(X_i, Y_i, Z_i)\}_{i \in \mathcal{M}})$ is computed from the matched data. Then, conditional on the original data $\mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i \leq n}$, pseudo-assignments $(Z'_i)_{i \in \mathcal{M}}$ are sampled by randomly permuting the true treatment assignments across matched pairs. Finally, the paired randomization test p-value is defined as $p = \mathbb{P}(|\hat{\tau}| \geq |\hat{\tau}| \mid \mathcal{D})$ where $\hat{\tau}$ is the test statistic evaluated using the pseudo-assignments instead of the actual ones. When the matching $\mathcal{M}$ is undefined, we arbitrarily set $p = 1$.

2.2 The issue of bias

The main concern when evaluating the approximate validity of the paired randomization test is the bias of the test statistic under the null hypothesis. Even biases that are small in absolute terms may dramatically degrade the level of the randomization test.

This may be explained by the following heuristic. For many test statistics, the reference distribution of the paired randomization test will be concentrated in an interval of width $O(1/\sqrt{n})$ around zero, since this is the desired behavior in paired randomized experiments with no treatment effect [17]. Therefore, any test statistic $\hat{\tau}$ whose bias does not vanish at least as fast as $O(1/\sqrt{n})$ under the null will appear atypical by the standards of the reference distribution, yielding small p-values and dramatically inflated Type I error.

One implication of this heuristic is that the paired randomization test based on the simple difference-of-means statistic $\hat{\tau} \equiv \hat{\tau}(\{(X_i, Y_i, Z_i)\}_{i \in \mathcal{M}}) = \sum_{Z_i=1}(Y_i - Y_{m(i)})/N_1$ will not generally be asymptotically valid, even if we are willing to make strong smoothness or functional-form assumptions on the outcome model and propensity score. Starting with Rubin [53, 54], numerous papers have shown that $\hat{\tau}$ may be considerably biased even in simple problems [56, 57, 58, 37, 1, 3, 59, 62]. While the consequences of this bias for randomization inference were not spelled out in these papers, the following examples (proved in the supplementary materials) show that they may be dramatic.

Example 1. (Asymptotic bias in an interval). Let $X \sim \text{Uniform}(0,1)$ and suppose that the propensity score and outcome model are both linear, $e(X) = \xi + \theta X$, $\mathbb{E}[Y \mid X, Z] = \gamma + \beta X$ with $\beta \neq 0$. Further suppose that Assumption 1 holds and that $\mathbb{P}(e(X) \geq 0.5) > 0$. Then Sävje [62, Proposition 1] shows that $\hat{\tau} \equiv \hat{\tau}(\{(X_i, Y_i, Z_i)\}_{i \in \mathcal{M}})$ is asymptotically biased and a slight extension of his analysis shows that $\mathbb{P}(p < \alpha) \rightarrow 1$ for every $\alpha \in (0, 1]$. Thus, in large samples, the paired randomization test makes a false discovery with probability tending to one.

Example 2. (Asymptotic bias on a disc). This example is based on the analysis of Rubin [56]. Suppose that $d = 2$ and let $X$ be uniformly distributed on the unit disc. Pick any unit vector $\theta \in \mathbb{R}^2$ and sample $Z \mid X \sim \text{Bernoulli}(0.35(1 + \theta^\top X))$, $Y \mid X, Z \sim N(\theta^\top X, \sigma^2)$, and finally set $Y(0) = Y(1) = Y$. This distribution satisfies Assumption 1 and Fisher’s sharp null hypothesis. However, $\hat{\tau}$ is asymptotically biased and $\mathbb{P}(p < \alpha) \rightarrow 1$ for every $\alpha \in (0, 1]$.

In the above examples, the paired randomization test fails because some parts of the covariate space have too many treated units and not enough untreated units. This leads to asymptotic bias, since any pairing
scheme must make some low-quality matches in regions with not enough untreated units. Recognizing this problem, some authors have suggested only using pair matching in problems where \( P\{e(X) < 0.5\} = 1 \) [69]. The next two examples show that this may work if \( X \) has only one continuous component, but it does not work in general.

**Example 3.** *(One continuous covariate).* Suppose that \( P \) satisfies Assumption 1, Fisher’s sharp null hypothesis, and \( P\{e(X) \leq 0.5 - \kappa\} = 1 \) for some \( \kappa > 0 \). Let \( \mu(x) = \mathbb{E}(Y \mid X = x) \) be the outcome regression. Then we show in the supplementary materials that the paired randomization test based on \( \hat{\tau}^{DM} \) is asymptotically valid if and only if the (conditional) bias decays at a rate faster than \( n^{-1/2} \):

\[
\frac{1}{N_1} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} = o_P(n^{-1/2}). \tag{4}
\]

When \( d = 1 \), Abadie & Imbens [4, Proposition 1] implies that (4) holds if \( \mu \) is Lipschitz-continuous and \( X \) has a positive density supported on a compact interval. Thus, the paired randomization test based on \( \hat{\tau}^{DM} \) is asymptotically valid if no units have propensity scores close to one-half and \( X \) has only one continuous component. Numerical evidence suggests that the same conclusion holds for up to three continuous components.

**Example 4.** *(More than three continuous covariates).* Once the number of continuous covariates is larger than three, the assumption that there are no large propensity scores no longer suffices for asymptotic validity, even if the outcome model is linear and the propensity score is extremely smooth. We illustrate this in a numerical example with four continuous covariates. For various sample sizes \( n \) between 200 and 2,000, we sampled data from the following linear/logistic model:

\[
\begin{align*}
X &\sim \text{Uniform}([-1, 1]^4) \\
Z \mid X &\sim \text{Bernoulli}[1/\{1 + \exp(1.1 - X_1)\}] \\
Y \mid X, Z &\sim N(3X_1, 1). \\
\end{align*}
\tag{5}
\]

This was repeated 2,000 times per sample size. In each simulation, we recorded the bias of \( \hat{\tau}^{DM} \) and the paired randomization test p-value\(^1\). The results are shown in Fig. 1. Although this distribution has \( P\{e(X) < 0.475\} = 1 \), the paired randomization test performs poorly. In fact, the Type I error of nominally level 5% tests appears to increase with the sample size. The issue is that the bias of the difference-of-means statistic converges to zero slower than (4) requires.

It is worth mentioning that the balance requirement (4) would not hold even in a completely randomized experiment [34], so the balance required for asymptotic validity of the paired randomization test cannot be certified by any balance test with a completely randomized reference distribution. This includes the two-sample t-test and all of the examples in Chapter 10 of [52]. Indeed, in each of our simulations from (5), we also performed a nominally level 10% balance test using Hotelling’s \( T^2 \). The balance test did not detect imbalance in any of our simulations. See [7, 25, 30, 10] for further discussion on the role of balance tests.

In our view, these examples show that no standard sampling assumptions (smoothness conditions or functional-form assumptions) will be able to justify the asymptotic validity of the paired randomization test based on the difference-of-means statistic, except when the number of continuous covariates is very small and no units have propensity scores close to one-half.

However, this does not mean that it is hopeless to apply the paired randomization test after matching outside of such problems. It is possible that the test would perform better with a different test statistic. In particular, test statistics that perform additional model-based covariance adjustment after matching can achieve approximate unbiasedness even \( d > 3 \) or \( P\{e(X) > 0.5\} > 0 \), provided the model is correctly-specified. For example, the analysis of covariance statistic \( \hat{\tau} \) from the regression (6) is exactly unbiased when the linear model \( \mathbb{E}(Y \mid X, Z) = \gamma + \tau Z + \beta^\top X \) holds. Thus, our heuristic does not rule out the paired randomization test based on this statistic or its variants.

\[
(\hat{\gamma}, \hat{\tau}, \hat{\beta}) = \arg\min_{\gamma, \tau, \beta} \sum_{i \in \mathcal{M}} (Y_i - \gamma - \tau Z_i - \beta^\top X_i)^2 \tag{6}
\]

\(^1\)We approximate this p-value using 1,000 randomly sampled permutations.
Figure 1: The left panel plots the average bias of the difference-in-means statistic in the distribution (5) at various sample sizes, on a log-log scale. The slope of the best-fit line is \( \approx -0.45 \), suggesting that the bias does not satisfy the \( o(n^{-1/2}) \) decay rate required by (4). The right panel shows the Type I error of nominally level 5% paired randomization tests based on the difference-of-means statistic.

Covariance-adjusted test statistics have been used in randomized experiments since Fisher [16], so employing such a statistic does not depart from the general philosophy of analyzing observational studies as randomized experiments.

2.3 The sampling and randomization variance

Suppose now that the data analyst bases the paired randomization test on a regression-adjusted test statistic and the true outcome model is linear. Then bias is not an issue, so the next question to ask is whether the randomization distribution consistently estimates the sampling variance of the test statistic.

In answering this question, we focus our analysis on two regression-adjusted test statistics that are widely used in randomized experiments:

(P1) The first is the analysis of covariance/ordinary least-squares test statistic proposed by Fisher [16]. We regress \( Y \) on \( Z \) and \( X \) in the matched sample as in Equation (6), and then use the regression coefficient \( \hat{\tau} \) on \( Z \) as our test statistic.

(P2) The second is the residual difference-of-means statistic proposed by Tukey [63] and Gail et al [21]. We regress \( Y \) on \( X \) (but not \( Z \)) in the matched sample to obtain fitted residuals \( \hat{\varepsilon}_i = Y_i - \hat{\gamma} - \hat{\beta}^\top X_i \) and then use \( \hat{\tau} = \frac{\sum_{Z_i=1} (\hat{\varepsilon}_i - \hat{\varepsilon}_{m(i)})}{N_1} \) as our test statistic.

These are not the only regression-adjusted test statistics based on linear models. Six or seven others are considered in [54, 43, 35, 4, 60, 17, 49]. In the interest of space, we do not provide results for these alternatives.

The following Proposition shows that if linearity holds and optimal matching successfully balances covariates in the matched sample, then the paired randomization based on either of these regression-adjusted test statistics is asymptotically valid. These assumptions are quite strong, but they have testable implications. In problems where the linearity assumption is found to be unreasonable (e.g. binary outcomes), other test statistics would be more appropriate.

**Proposition 1.** Suppose that \( \mathcal{P} \) satisfies Assumption 1, Fisher’s sharp null hypothesis, and the linear model \( \mathbb{E}(Y \mid X, Z) = \gamma + \beta^\top X \). Further suppose that \( ||X_i - \bar{X}_z|| = o_P(1) \), where \( \bar{X}_z \) is the average of \( X_i \) among matched samples with \( Z_i = z \). Then the paired randomization test p-value based on either (P1) or (P2) satisfies \( \mathbb{P}(p < \alpha) \rightarrow \alpha \) for all \( \alpha \in (0, 1) \).

The requirement that the matching scheme balances covariates may be surprising. After all, the other conditions of Proposition 1 allow for asymptotically valid model-based inference (e.g., linear regression in the matched sample followed by Eicker-Huber-White robust standard errors) even if covariate balance is not achieved. This type of inference has been advocated by Ho et al [29]. Examples 1 and 2 in the previous
subsection show that matching does not always balance covariates, so model-based inference may succeed in some examples where randomization inference does not.

The explanation is that the reference distribution of the paired randomization test assumes that \( Z \) is independent of \( X \) in the matched sample, since it permutes the treatment assignment uniformly at random across pairs without regard for \( X \). However, if the matching process does not successfully balance covariates, then \( Z \) and \( X \) are correlated even after matching. This correlation affects the sampling variance of \( \hat{\tau} \) but not the randomization variance of \( \tilde{\tau} \). Proposition 2 shows that this mismatch may result in overrejection or underrejection, depending on which test statistic is used.

**Proposition 2.** Suppose that \( P \) satisfies Assumption 1 and that \( Y \mid X, Z \sim N(\gamma + \beta^\top X, \sigma^2) \). Further assume that optimal matching fails to balance covariates, \( ||\bar{X}_1 - \bar{X}_0|| \neq o_P(1) \). Then for any \( \alpha \in (0, 1) \), the paired randomization test based on the analysis of covariance statistic \((P1)\) is not asymptotically valid \((\lim \sup P(p < \alpha) > \alpha)\) and the paired randomization test based on the residual difference-of-means statistic \((P2)\) is asymptotically conservative \((\lim \inf P(p < \alpha) < \alpha)\).

An implication of this result is that users should assess covariate balance after matching, even when they plan to use regression adjustment to account for residual imbalances. Unlike the stringent balance requirement \((4)\) required by the difference-of-means estimator, the balance requirement in Proposition 1 is mild and can be certified by standard balance tests. In simulations, we have found that the inflation in Type I error rate due to covariate imbalance is minor unless the imbalance is severe.

### 2.4 Robustness to local misspecification

Proposition 1 above justifies the paired randomization test under the assumption that the linear outcome model is correctly specified. In this section, we explore the extent to which the test is robust to local deviations from linearity.

To study this problem, we introduce a family of distributions where the linear model is misspecified. Let \( P \) be any baseline distribution satisfying Assumption 1, Fisher’s sharp null hypothesis, and the linear outcome model \( Y \mid X, Z \sim N(\gamma + \beta^\top X, \sigma^2) \). For any bounded nonlinear function \( g: \mathbb{R}^d \to \mathbb{R} \) and any \( h \in \mathbb{R} \), let \( P_{h,g} \) be the distribution of the vector obtained by adding \( hg(X) \) to both potential outcomes. Fisher’s sharp null hypothesis continues to hold under \( P_{h,g} \). However, the regression function now takes the form

\[
E_{h,g}[Y \mid X, Z] = \gamma + \beta^\top X + hg(X),
\]

so the linear outcome model is misspecified in the model \( P_g = \{P_{h,g} : h \in \mathbb{R}\} \).

To formalize that the misspecification is small, we follow the literature on local misspecification \([38, 8]\) and consider a triangular array where \( h \equiv h_n \) converges to zero at rate \( O(1/\sqrt{n}) \). Under this scaling, the linear model is nearly correctly specified in the sense that no test can consistently distinguish \( P \) from \( P_{h_n,g} \) on the basis of \( n \) samples. Thus, a robust test should remain valid even under sampling from \( P_{h_n,g} \). This motivates the following definition.

**Definition 1.** A sequence of p-values \( p_n \) is called robust to local nonlinearity near \( P \) if \((7)\) holds for every bounded nonlinear function \( g \) and every sequence \( h_n = O(1/\sqrt{n}) \).

\[
\lim \sup_{n \to \infty} P_{h_n,g}(p_n < \alpha) \leq \alpha \quad \text{for all } \alpha \in (0, 1),
\]

\[
(7)
\]

A robust p-value remains asymptotically valid under small deviations from the linear model. Equivalently, a robust p-value controls the asymptotic level locally uniformly over every model \( P_g \). Although this is not a genuinely nonparametric concept (since it is restricted to small neighborhoods of a linear model), it has been a familiar notion of robustness in semiparametric statistics since Beran \([9]\) or earlier. Standard results show that \((7)\) can be achieved using the Wald p-value based on any locally efficient estimator of the average treatment effect on the treated \([64, \text{Chapter 25}]\). In contrast, the Wald p-value based on linear regression without matching is typically not robust in this way.

The following proposition shows that the local robustness of the paired randomization test p-value depends crucially on the distribution of the propensity score. If there are units with propensity scores larger than one-half, then the paired randomization test remains sensitive to certain small deviations from linearity.
Proposition 3. Suppose that $P$ satisfies Assumption 1 and the linear model $Y \mid X, Z \sim N(\gamma + \beta^\top X, \sigma^2)$. Let $p$ be the paired randomization test $p$-value based on either of the regression-adjusted test statistics (P1) or (P2). Then $p$ is robust to local nonlinearity near $P$ if $\mathbb{P}\{e(X) < 0.5\} = 1$ but not if $\mathbb{P}\{e(X) > 0.5\} > 0$.

If some units have propensity scores larger than one-half, this Proposition says that even the paired randomization test based on the residual difference-of-means statistic (P2) is not entirely safe. Although Proposition 1 guarantees the asymptotic conservativeness of that test, it only does so under the assumption of exact linearity. Under only approximate linearity, misspecification may still lead to overrejection.

The intuition is the following. If some units have propensity scores larger than one-half, then there are functions $g$ that can never be successfully balanced by optimal matching or any other maximal pair matching scheme. For example, $g(x) = I\{e(x) > 0.5\}$ is one such function. If the linear model is slightly misspecified in the direction of one of these functions, the resulting bias affects the asymptotic distribution of $\hat{\tau}$ and spoils asymptotic validity. If $\mathbb{P}\{e(X) < 0.5\} = 1$, then all functions can be successfully balanced and this problem does not arise.

Finally, we make a comment on the interpretation of this result. Proposition 3 does not state that the paired randomization test always performs poorly when the true outcome regression is nonlinear. Actually, we have found the test based on either (P1) or (P2) to be remarkably robust to (non-adversarial) nonlinearity in the outcome model. However, Proposition 3 does provide a warning: if units with propensity scores larger than one-half exist, deviations from linearity that are hard to detect can still adversely affect the Type I error of the paired randomization test.

3 The stratified Fisher randomization test

3.1 Motivation

In the second part of this article, we introduce a new randomization test that repairs some of the weaknesses described in 2. Our goal with this test is to gain as much robustness as possible while respecting the following principles: (i) the test should be the Fisher randomization test from some clearly-articulated randomized experiment; (ii) any model-based adjustments going into this test must be as simple as the parametric adjustments currently used in the randomization inference community.

We start by giving some motivation for our proposal. Our results above show that the regression-adjusted paired randomization test mainly runs into trouble when some units have propensity scores larger than one-half. In such cases, point estimators may be asymptotically biased (Examples 1 and 2), the randomization variance and sampling variance may disagree (Proposition 2), and inferences may be sensitive to small violations of parametric assumptions (Proposition 3). The source of all these problems is that any pair matching scheme must make low-quality matches in regions of the covariate space with not enough untreated units.

In our view, the simplest solution to this problem is to allow the same untreated observation to be matched to multiple treated observations, i.e. do the matching with replacement. This way, the match for a treated observation $i$ is simply its nearest untreated neighbor:

$$m(i) = \arg\min_{j: Z_j = 0} (X_i - X_j)^\top \hat{\Sigma}^{-1} (X_i - X_j).$$

Matching with replacement produces closer pairs than optimal matching and — assuming overlap — it never runs out of controls in any region of the covariate space. In the R programming language [44], this type of matching is implemented by default in the Matching package [60].

One-nearest-neighbor matching with replacement has several advantages over other schemes that might be used to handle units with propensity scores larger than one-half. These include variable-ratio matching [37], full matching [48], matching for balance [66], and kernel optimal matching [32]. One major advantage is interpretability: for each treated observation $i$, the scheme (8) finds one and only one untreated observation $j$ to proxy for $i$’s counterfactual outcome. This enables a qualitative/narrative description of certain matched

---

2 Although our focus is on problems with continuous covariates, our theory will also apply when ties are possible as long as they are broken according to the randomized rule in [13].
Figure 2: An illustration of how one-nearest-neighbor matching with replacement simulates a stratified randomized experiment. Left: each treated unit is matched with an untreated unit. Right: each matched untreated unit defines a stratum, and the strata partition the set of matched units.

pairs, as discussed in [52, Section 22.3]. Matching with replacement is also computationally straightforward and analytically tractable.

Despite these advantages, matching with replacement has not caught on in the randomization inference community. One explanation is that it does not clearly simulate any randomized experiment. In contrast with pair matching, the matched set cannot be viewed as arising from a paired experiment since treatment assignment is correlated across pairs with the same control. Therefore, it is not clear how to perform randomization inference after matching with replacement. Below, we point out one way this can be done and establish its asymptotic validity under sampling assumptions.

3.2 Simulating a stratified experiment

Let $\mathcal{M} = \{i : Z_i = 1 \text{ or } i = m(j) \text{ for some treated unit } j\}$ be the set of observations matched by the nearest-neighbor Mahalanobis matching with replacement scheme. For each untreated observation $i \in \mathcal{M}$, define its stratum $S(i)$ by:

$$S(i) = \{i\} \cup \{j : m(j) = i\}.$$

Observe that these strata $\{S(i)\}_{i \in \mathcal{M}, Z_i = 0}$ form a partition of the matched set $\mathcal{M}$, and each stratum has exactly one untreated observation. Therefore, the one-nearest-neighbor matching process can be viewed as simulating a stratified experiment with strata $\{S(i)\}$ which treats all but one unit in each stratum. Figure 2 illustrates this in a simple two-dimensional example.

With this perspective in hand, it becomes natural to use the randomization test from the associated stratified randomized experiment to test the sharp null (1). We call this the stratified (Fisher) randomization test. This is the same as the paired randomization test, except the pseudo-assignments $(Z_i^*)_{i \in \mathcal{M}}$ are obtained by randomly permuting the true assignments within each stratum $S(i)$ instead of within each matched pair. This type of randomization is familiar from observational studies with variable-ratio matching. However, each of our strata has one untreated unit and (possibly) several treated units while in variable-ratio matching it is the other way round. The stratified experiment reference distribution can also be used for checking covariate balance, in line with the recommendations of Hansen & Bowers [26], Branson [10].

We have presented the stratified randomization test as a design-based test. However, neither the stratified randomization test nor the paired randomization test are actually justified by design when matches are inexact, so this framing is largely cosmetic. We are chiefly interested in the stratified randomization test because it strikes a reasonable balance between interpretability and favorable statistical properties under superpopulation sampling.
3.3 Test statistics

As in Section 2, our theoretical results for the stratified randomization test require the analyst to base the test on a regression-adjusted test statistic. Even though matching with replacement is typically less biased than pair matching, model-based bias reduction is still required for reliable inference when \( d > 3 \) \cite{1, 3}.

We study two types of test statistics, which generalize the ones from Section 2.3 to the case of more than one match:

(R1) The first is the analysis of covariance test statistic, obtained by extracting the coefficient \( \hat{\tau} \) in the weighted regression (9).

\[
\hat{\gamma}, \hat{\tau}, \hat{\beta} = \arg\min_{\gamma, \tau, \beta} \sum_{i \in M} W_i (Y_i - \gamma - \tau Z_i - \beta^T X_i)^2 \tag{9}
\]

Here, the weights are \( W_i = 1 \) if \( Z_i = 1 \) and \( W_i = |S(i)| - 1 \) if \( Z_i = 0 \). This weighting is natural if we think of untreated observations as being repeated once per match. In the randomization test, we would set \( W_i = 1 \) if \( Z_i = 1 \) and \( W_i = |S(i)| - 1 \) if \( Z_i = 0 \).

(R2) The second is the analogue of the residual difference-of-means statistic. First, we regress \( Y \) on \( X \) (but not \( Z \)) in the matched sample to obtain residuals \( \hat{\epsilon}_i = Y_i - \hat{\gamma} - \hat{\beta}^T X_i \). The regression may either be weighted or unweighted. Then, we use \( \hat{\tau} = \sum_{Z_i=1} \{ \hat{\epsilon}_i - \hat{\epsilon}_{m(i)} \} / N_1 \) as our test statistic. If the residuals \( \hat{\epsilon}_i \)'s are obtained from an unweighted regression, we would compute \( \hat{\tau} = \sum_{Z_i=1} \{ \hat{\epsilon}_i - \hat{\epsilon}_{m(i)} \} / N_1 \) in each permutation to build the reference distribution. If the initial regression was weighted, it would be more natural to re-estimate the residuals in each permutation as well.

If no two treated units are matched to the same control unit, then these test statistics reduce to their paired counterparts (P1) and (P2) and the stratified randomization test reproduces the paired randomization test.

3.4 Validity and robustness

We now show that when the stratified randomization test is performed using a regression-adjusted test statistic, the resulting p-value \( \hat{p} \) is uniformly asymptotically valid in certain local neighborhoods of \( H_0 \). This still requires approximately linearity, but that is probably unavoidable unless one is willing to perform complex nonparametric adjustment after matching (as in \cite{1, 36}) or dramatically alter the matching scheme (as in \cite{66}). However, we now allow for some or even all units to have propensity scores larger than one-half.

**Theorem 1.** Assume that \( P \) satisfies Assumption 1, except possibly the condition that \( P(Z = 1) < 0.5 \). Further assume that Fisher’s sharp null hypothesis holds and that the outcome model is linear, \( \mathbb{E}[Y|X, Z] = \gamma + \beta^T X \). Then for any smooth parametric model \( \{ P_h : h \in \mathbb{R}^k \} \subset H_0 \) with \( P_0 = P \) and any radius \( C < \infty \), we have:

\[
\sup_{||h|| \leq C/\sqrt{n}} \mathbb{P}_h(\hat{p} < \alpha) \to \alpha \quad \text{for all } \alpha \in (0, 1). \tag{10}
\]

In other words, \( \hat{p} \) is locally uniformly asymptotically valid in the semiparametric sense.

The uniformity guarantee (10) is stronger than the robustness-to-local-nonlinearity property discussed in Section 2.4, and may be unexpected. To our knowledge, all prior examples of tests satisfying (10) are based on regular asymptotically linear estimators. However, the test statistic \( \hat{\tau} \) is not asymptotically linear and prior work has suggested that its distribution is an irregular function of \( P \) \cite{5, 68}. In fact, it is not even known whether \( \sqrt{n} \hat{\tau} \) has any asymptotic distribution at all when \( d \geq 2 \). When \( d = 1 \), \( \sqrt{n} \hat{\tau} \) is asymptotically normal but the nonparametric bootstrap is inconsistent \cite{2}. This is often taken as a sign of poor uniform performance.

Intuitively, this robustness arises because the regression-adjusted test statistics (R1) and (R2) have two chances at consistency. The first chance comes from the linear model, which is uniformly \( \sqrt{n} \)-consistent in contiguous neighborhoods of a linear model, but not locally uniformly asymptotically unbiased. The second chance comes from the matching process itself.
Lemma 1. Under Assumption 1, both (R1) and (R2) are consistent for the average treatment effect on the treated \( \mathbb{E}\{Y(1) - Y(0) \mid Z = 1\} \). No smoothness assumptions on the outcome model or propensity score are required.

The convergence rate in Lemma 1 may be slow, but any rate of convergence suffices to cancel away the local asymptotic bias of the model-based estimator. In contrast, the pair matching estimators (P1) and (P2) only have one chance at consistency when \( \mathbb{P}\{e(X) > 0.5\} > 0 \), since the point estimator is not guaranteed to be consistent unless the linear model holds (Examples 1 and 2). Thus, there is no cancellation of local bias and (10) does not always hold for the paired randomization test.

3.5 A numerical example

In this section, we numerically compare the robustness of several tests discussed so far in this paper in a stylized example. For each of several values of \( h \in [0, 0.25] \), we sampled \( n = 2,000 \) observations from the following distribution:

\[
\begin{align*}
X &\sim \text{Uniform}([-1, 1]^4) \\
Z \mid X &\sim \text{Bernoulli}(1/[1 + \exp\{1 - 2X_1 - 2.5X_2X_3\}]) \\
Y \mid X, Z &\sim N(X_1 + hX_2X_3, 1).
\end{align*}
\]

(11)

The \( X_2X_3 \) interaction term induces nonlinear confounding which is not addressed by linear adjustment alone. Roughly one-third of the observations in a typical sample from the distribution (11) will be treated, but some of them will have propensity scores larger than one-half.

In each of our simulations, we computed test statistics and p-values from each of the following procedures:

(i) The paired Fisher randomization test based on the difference-of-means statistic \( \hat{\tau}_{\text{DM}} \)
(ii) The paired Fisher randomization test based on the analysis of covariance statistic (P1)
(iii) The stratified Fisher randomization test based on the residual difference-of-means statistic (R2)
(iv) The model-based p-value from a linear model fit to the entire dataset with no matching step, based on Eicker-Huber-White robust standard errors.

For the stratified randomization test, we use (R2) rather than (R1) only because its reference distribution is easier to compute; the two statistics yield tests that are asymptotically equivalent under the null hypothesis. Pairings and stratifications were computed using the \texttt{optmatch} and \texttt{Matching} packages, respectively [27, 60].

Figure 3 plots the bias of test statistics and the Type I error rate of nominally level 5% tests as a function of \( h \). These are are estimated using 2,000 replications per value of \( h \). Results for the difference-of-means statistic are not plotted, as it has dramatically larger bias (> 0.15 for all \( h \)) and Type I error (> 0.75 for all \( h \)) than the other methods. For small values of \( h \), all regression-adjusted methods have low bias and approximately nominal Type I error. However, for larger values of \( h \), regression after optimal pair matching results in lower bias and Type I error than regression without matching, and regression after matching with replacement is even more robust. Only the stratified Fisher randomization test controls Type I error at nearly the nominal level for all considered values of \( h \).

4 Discussion

This paper studied the behavior of randomization tests under independent sampling. We found that the paired randomization test based on an approximately correctly-specified regression model is asymptotically valid despite inexact matching if no units have propensity scores larger than one-half. However, if \( \mathbb{P}\{e(X) > 0.5\} > 0 \) or the unadjusted difference-of-means statistic is used, the test may have poor Type I error control. We also proposed a new randomization test that achieves approximate validity even when some units have large propensity scores.

After addressing overt bias due to measured confounders, the analyst must turn to hidden bias. Although we have not proposed a specific sensitivity analysis for the stratified randomization test, we expect that existing sensitivity analysis methods for full matching or variable-ratio matching could be applied with minimal modification [51]. A more interesting question is the extent to which existing methods for sensitivity
Figure 3: The left panel shows estimated bias of various test statistics. The right panel shows the estimated Type I error of nominally level 5% tests. EHW is linear regression with Eicker-Huber-White standard errors and no matching. Paired is the paired randomization test with the analysis of covariance test statistic (P1). Stratified is the stratified randomization test with the residual difference-of-means test statistic (R2).

analysis remain valid in the presence of inexact matching. As discussed in Pashley et al [41, Section 5], the usual justification for design-based sensitivity analysis also not apply in the presence of inexact matching. The extent to which regression adjustment or alternative matching schemes can make up for this is an important question for future work.

Acknowledgement

We are grateful to Sky Cao, John Cherian, Peng Ding, Colin Fogarty, Isaac Gibbs, Samir Khan, Sam Pimentel, Fredrik Sävje and seminar participants at the 2021 Stanford Causal Science Center conference for helpful comments and discussions.
References

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

[2] Alberto Abadie and Guido W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

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

[4] Alberto Abadie and Guido W. Imbens. A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843, 2012.

[5] Alberto Abadie and Guido W. Imbens. Matching on the Estimated Propensity Score. *Econometrica*, 84(2):781–807, 2016.

[6] Timothy B. Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177, 2021.

[7] Peter C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12):2037–49, 2008.

[8] Anil K. Bera and Mann J. Yoon. Specification testing with locally misspecified alternatives. *Econometric Theory*, 9(4):649–658, 1993.

[9] Rudolf Beran. An efficient and robust adaptive estimator of location. *Annals of Statistics*, 6:292–313, 1978.

[10] Zach Branson. Randomization tests to assess covariate balance when designing and analyzing matched datasets. *Observational Studies*, 2021.

[11] Zach Branson and Marie-Abéle Bind. Randomization-based inference for bernoulli trial experiments and implications for observational studies. *Statistical Methods in Medical Research*, 28(5):1378–1398, 2019. PMID: 29451089.

[12] EunYi Chung and Joseph P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484 – 507, 2013.

[13] Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gabor Lugosi. On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. *The Annals of Statistics*, 22(3):1371 – 1385, 1994.

[14] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

[15] Bruno Ferman. Matching estimators with few treated and many control observations. *Journal of Econometrics*, 225, 08 2021.

[16] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[17] Colin B. Fogarty. Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000, 06 2018.

[18] Colin B. Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, 115(531):1518–1530, 2020.

[19] Colin B. Fogarty, Mark E. Mikkelsen, David F. Gaieski, and Dylan S. Small. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514):447–458, 2016.
[20] Colin B. Fogarty, Pixu Shi, Mark E. Mikkelsen, and Dylan S. Small. Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. *Journal of the American Statistical Association*, 112(517):321–331, 2017.

[21] M. H. Gail, W. Y. Tan, and S. Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1):57–64, 1988.

[22] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

[23] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

[24] P. Hall and C.C. Heyde. *Martingale Limit Theory and Its Application*. Probability and mathematical statistics. Academic Press, 1980.

[25] Ben B. Hansen. The essential role of balance tests in propensity-matched observational studies: Comments on ‘a critical appraisal of propensity-score matching in the medial literature between 1996 and 2003’ by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27:2050–2054, May 2008.

[26] Ben B. Hansen and Jake Bowers. Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science*, 23(2):219 – 236, 2008.

[27] Ben B. Hansen and Stephanie Olsen Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

[28] Jennifer Hill. Comment on ‘covariance adjustment in randomized experiments and observational studies’ by Paul R. Rosenbaum. *Statistical Science*, 17:304–327, 08 2002.

[29] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

[30] Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171, 12 2007.

[31] Arnold Janssen. Studentized permutation tests for non-i.i.d. hypotheses and the generalized behrens-fisher problem. *Statistics & Probability Letters*, 36(1):9–21, 1997.

[32] Nathan Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020.

[33] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of hogdes’ estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

[34] Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

[35] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman’s critique. *Ann. Appl. Stat.*, 7(1):295–318, 03 2013.

[36] Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect, 2021.

[37] Kewei Ming and Paul R. Rosenbaum. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56(1):118–124, 2000.

[38] Whitney K. Newey. Generalized method of moments specification testing. *Journal of Econometrics*, 29(3):229–256, 1985.
[39] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 11 1923.

[40] Taisuke Otsu and Yoshiyasu Rai. Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732, 2017.

[41] Nicole E. Pashley, Guillaume W. Basse, and Luke W. Miratrix. Conditional as-if analyses in randomized experiments. *Journal of Causal Inference*, 9(1):264–284, 2021.

[42] Valentin V. Petrov. *Classical-Type Limit Theorems for Sums of Independent Random Variables*, pages 1–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[43] Dana Quade. Nonparametric analysis of covariance by matching. *Biometrics*, 38(3):597–611, 1982.

[44] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[45] James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421, 2008.

[46] Paul R. Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

[47] Paul R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

[48] Paul R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

[49] Paul R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 08 2002.

[50] Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, 2002.

[51] Paul R. Rosenbaum. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63 2:456–64, 2007.

[52] Paul R. Rosenbaum. *Design of Observational Studies*. Springer, 2010.

[53] Donald B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183, 1973.

[54] Donald B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, 1973.

[55] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[56] Donald B. Rubin. Multivariate matching methods that are equal percent bias reducing, ii: Maximums on bias reduction for fixed sample sizes. *Biometrics*, 32(1):121–132, 1976.

[57] Donald B. Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, 36(2):293–298, 1980.

[58] Donald B. Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

[59] Joseph Schafer and Joseph Kang. Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological methods*, 13:279–313, 01 2009.

[60] Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52, 2011.
[61] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

[62] Fredrik Sävje. On the inconsistency of matching without replacement. *Biometrika*, 06 2021. asab035.

[63] John W. Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14(4):266–285, 1993.

[64] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[65] Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the $r$th Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299 – 303, 1965.

[66] Yixin Wang and José R. Zubizarreta. Large sample properties of matching for balance. *Statistica Sinica*, 2019.

[67] Jason Wu and Peng Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 0(0):1–16, 2020.

[68] Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020. PMID: 33088006.

[69] Ruoqi Yu, Jeffrey H. Silber, and Paul R. Rosenbaum. Rejoinder: Matching Methods for Observational Studies Derived from Large Administrative Databases. *Statistical Science*, 35(3):371 – 374, 2020.

[70] Anqi Zhao and Peng Ding. Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294, 2021. Themed Issue: Treatment Effect 1.
A Proofs of main results

In this section: for any integer $k \geq 1$, $e_k$ is the $k$-th standard basis vector (where the ambient dimension will be clear from context) and $[k]$ is the set $\{1, \ldots, k\}$; for two symmetric matrices $A, B$, we write $A \succeq B$ (or $B \preceq A$) if $A - B$ is positive semidefinite.

Given a sequence of random distributions $\hat{Q}_n$, we say that $\hat{Q}_n$ converges weakly in probability to $Q$ (denoted $\hat{Q}_n \Rightarrow Q$) if $\rho(\hat{Q}_n, Q) = o_P(1)$ for some distance $\rho$ metrizing weak convergence. Often, we will apply this with $\hat{Q}_n$, the distribution of a statistic $H_*$ conditional on the original data $\{(X_i, Y_i, Z_i)\}_{i \leq n}$. In such cases, we will also write $H_* \Rightarrow H$ where $H \sim Q$.

For any distribution $P$ satisfying Assumption 1, we use $\mu_z(x)$ to denote the regression function $E[Y|X = x, Z = z]$ and $\sigma^2_z(X)$ to denote the conditional variance $\text{Var}(Y|X = x, Z = z)$. If $P$ is also assumed to satisfy Fisher’s null hypothesis, then we will typically drop the subscript $z$ and simply write $\mu(x), \sigma^2(x)$ since the conditional distribution of $Y$ given $(X, Z)$ will not depend on $Z$.

A.1 Technical lemmas

Lemma 2. (Berry-Esseen Theorem, [42]). Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed random variables with mean zero and three finite moments. Let $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \text{Var}(S_n)$. Then we have:

$$
\sup_{t \in \mathbb{R}} |\mathbb{P}(S_n/\sigma \leq t) - \Phi(t)| \leq \frac{C \sum_{i=1}^n \mathbb{E}[|X_i|^3]}{\sigma^3}
$$

where $C < \infty$ is an absolute constant.

Remark 1. We will typically apply this theorem conditionally on the dataset $D = \{(X_i, Y_i, Z_i)\}_{i \leq n}$ to establish weak convergence in probability. It is important to note that $\sup_{t \in \mathbb{R}} |\mathbb{P}(H_* \leq t | D) - \mathbb{P}(H \leq t)| = o_P(1)$ implies $H_* \Rightarrow H$ as convergence in the Kolmogorov distance is stronger than weak convergence.

Lemma 3. (Randomization Slutsky theorem, [12]). If $A_n \Rightarrow A$ and $B_n = A_n + o_P(1)$, then $B_n \Rightarrow A$ as well.

Lemma 4. Let $K_{i,n}$ be the number of treated observations matched with the $i$-th observation under matching with replacement (with the estimated Mahalanobis distance). If $P(Z = 0|X) \geq \delta > 0$ and $\text{Var}(X)$ is finite and invertible, then for every $q \geq 0$, $\sup_{n \geq 1} \mathbb{E}[K_{i,n}^q]$ is bounded by a constant $\kappa_q$ depending only on $q, \delta$ and $d$.

Proof. This result is very similar to [1, Lemma 3.(iii)], so we defer the proof to Appendix B. The proof is quite different from theirs, as we do not make as many regularity assumptions.

Lemma 5. For any $q > 0$, $\max_{i \leq n} K_{i,n} = o_P(n^q)$.

Proof. It suffices to prove the result for $q < 1$, in which case we can show convergence not only in probability but also in $L^{1/q}$. Start by writing the following:

$$
\mathbb{E} \left[ \left( \frac{\max_{i \leq n} K_{i,n}}{n^q} \right)^{1/q} \right] = \frac{1}{n} \mathbb{E}[\max_{i \leq n} K_{i,n}^{1/q}] = \int_0^\infty \frac{1}{n} \mathbb{P}(\max_{i \leq n} K_{i,n}^{1/q} > t) \, dt \quad : = f_n(t).
$$

Observe that $f_n(t) \to 0$ as $n \to \infty$ for each fixed $t$. Moreover, by the union bound, Markov’s inequality and Lemma 4, we have the bound $f_n(t) \leq \kappa_{2/q}/t^2 \equiv f(t)$, and $f(\cdot)$ is integrable. Therefore, by the dominated convergence theorem, $\int_0^\infty f_n(t) \, dt \to 0$. 

\[ \square \]
A.2 Details for Example 1

Lemma 6. Suppose that $X \sim \text{Uniform}(0,1)$ and $Z \mid X \sim \text{Bernoulli}(\xi + \theta X)$. If $P(Z = 1) < 0.5$ and $P(e(X) \geq 0.5) > 0$, then the covariate imbalance under optimal Mahalanobis matching satisfies:

$$
\frac{1}{N_i} \sum_{i=1}^{N_i} \{X_i - X_{m(i)}\} \overset{P}{\to} \frac{(2(\theta + \xi) - 1)^3}{3\theta^2(\theta + 2\xi)} \neq 0.
$$

Proof. Since the propensity score in our example is a linear function of the one-dimensional covariate $X$, optimal Mahalanobis matching is equivalent to optimal Euclidean matching which is in turn equivalent to optimal propensity score matching. Thus, we may use results from [62] in our analysis.

We will need some notation from [62]. Let $U = [n], T = \{i : Z_i = 1\}, C = \{i : Z_i = 0\}$.

Further define $M^*$ by:

$$
M^* = \begin{cases}
    \{m(i) : i \in T\} & \text{if } N_1 \leq N_0 \\
    C & \text{if } N_1 > N_0
\end{cases}
$$

Finally, for each set $A \in \{U, T, C, M^*\}$, let $A_+ = \{i \in A : e(X_i) \geq p^*\}$ and $A_- = \{i \in A : e(X_i) < p^*\}$.

Using this notation, we may write the covariate imbalance as the sum of several terms:

$$
\bar{X}_1 - \bar{X}_0 = \frac{1}{np} \sum_{i=1}^{np} \{X_i - X_{m(i)}\} + o_P(1)
$$

$$
= \frac{1}{np} \sum_{i \in T} X_i - \frac{1}{np} \sum_{i \in M^*_+} X_i + \frac{1}{np} \sum_{i \in C} X_i - \frac{1}{np} \sum_{i \in M^*_+} Y_i + \frac{1}{np} \sum_{i \in \tau_+} X_i - \frac{1}{np} \sum_{i \in \tilde{C}} X_i + o_P(1).
$$

The proof Lemma S7 in [62] gives a vanishing bound on the expectation of the term $i$.

$$
\mathbb{E}[[ii]] \leq \frac{2}{p} \left( \mathbb{E} \left[ \frac{|M^*_+| - |C|}{n} \right] + \mathbb{E} \left[ \frac{[M^*_+] - |C|}{n} \right] \right) \to 0.
$$

Hence, by Markov’s inequality, $i = o_P(1)$. Similarly, the proof of Lemma S12 in [62] gives $ii \leq \frac{1}{np} \sum_{i \in \tau_+} \{X_i - X_{m^*(i)}\}$ where $\tau_+$ and $m^*$ are defined in that proof. Lemmas S8, S11, S12 and S13 of [62] prove that the expectation of this upper bound vanishes, so $ii = o_P(1)$ again by Markov’s inequality. Finally, the term $iii$ converges to the claimed limit:

$$
iii = \frac{1}{np} \sum_{i=1}^{n} X_i I\{Z_i = 1, e(X_i) \geq p^*\} - \frac{1}{np} \sum_{i=1}^{n} X_i I\{Z_i = 0, e(X_i) \geq p^*\}
$$

$$
\overset{P}{\to} \frac{1}{p} \left( \mathbb{E}[X|Z = 1, e(X) \geq p^*] - \mathbb{E}[X|Z = 0, e(X) \geq p^*] \right)
$$

$$
= \frac{(2(\theta + \xi) - 1)^3}{3\theta^2(\theta + 2\xi)}.
$$

To show that this limit is not zero, it suffices to check that $2(\theta + \xi) - 1 \neq 0$. If $2(\theta + \xi) - 1 = 0$ then $\xi + \theta = 0.5$. However, if $\theta + \xi = 0.5$, then $P(e(X) < 0.5) = P(\xi + \theta X < 0.5) = 1$, which violates our assumption that $P(e(X) \geq 0.5) > 0$.

Lemma 7. Assume that $P$ satisfies Assumption 1, and let $\hat{q}_{1-\alpha}$ denote the 100(1−\alpha)% quantile of the randomization distribution of the difference-of-means statistic:

$$
\hat{q}_{1-\alpha} = \inf\{t \in \mathbb{R} : P(|\hat{z}^{DM}_i| \leq t \mid \{(X_i, Y_i, Z_i)\}_{i \leq n}) \geq 1 - \alpha\}.
$$

Then $\hat{q}_{1-\alpha} = O_P(1/\sqrt{n})$. 

Proof. Conditional on the original data $D = \{(X_i, Y_i, Z_i)\}_{i \leq n}$, the permuted difference-of-means statistic $\hat{\tau}_{DM}^*$ has mean zero and variance bounded as follows:

$$\text{Var}(\hat{\tau}_{DM}^* | D) = \frac{1}{N_1^2} \sum_{Z_i=1} \{Y_i - Y_m(i)\}^2 \leq \frac{1}{N_1^2} \sum_{Z_i=1} 2\{Y_i^2 + Y_m^2(i)\} \leq \frac{1}{N_1^2} \sum_{i=1}^n 2Y_i^2 =: \tilde{\sigma}^2.$$

Therefore, Chebyshev’s inequality gives $P(|\hat{\tau}_{DM}^*| \leq \frac{\tilde{\sigma}}{\sqrt{n}} | D) \geq 1 - \alpha$, which implies $\hat{\tau}_{1-\alpha} \leq \frac{\sigma}{\sqrt{n}}$. Since $\frac{1}{N_1} \sum_{i=1}^n 2Y_i^2 = O_P(1)$ by the law of large numbers and the assumption that $\text{Var}(Y) < \infty$, we conclude $\tilde{\sigma} = O_P(1/\sqrt{n})$ and the conclusion follows.

Lemma 8. Under the conditions of Example 1, the paired randomization test $p$-value $p$ based on the difference-of-means statistic after optimal Mahalanobis matching satisfies $P(p < \alpha) \to 1$ for all $\alpha \in (0, 1)$.

Proof. Let $\varepsilon_i = Y_i - \gamma - \beta X_i$. Observe that $\sum_{Z_i=1} \{\varepsilon_i - \varepsilon_m(i)\} = o_P(1)$ by Chebyshev’s inequality applied conditionally. Thus, we may apply Lemma 6 to show that $\hat{\tau}_{DM}$ converges to a nonzero limit:

$$\hat{\tau}_{DM} = \frac{1}{N_1} \sum_{Z_i=1} \beta\{X_i - X_m(i)\} + \frac{1}{N_1} \sum_{Z_i=1} \{\varepsilon_i - \varepsilon_m(i)\} \xrightarrow{P} \frac{\beta(2(\theta + \xi) - 1)^3}{3\theta^2(\theta + 2\xi)}.$$

Moreover, Lemma 7 implies the randomization critical value converges to zero at rate $O_P(1/\sqrt{n})$. Therefore, Slutsky’s theorem allows us to write:

$$P(p < \alpha) = P(P(|\hat{\tau}_{DM}^*| \geq |\hat{\tau}_{DM}^*| | D) < \alpha)$$

$$= P(P(|\hat{\tau}_{DM}^*| < |\hat{\tau}_{DM}^*| | D) \geq 1 - \alpha)$$

$$\leq P(\hat{\tau}_{1-\alpha} < |\hat{\tau}_{DM}^*|) \to 1.$$

\[\square\]

A.3 Details for Example 2

Lemma 9. Under the conditions of Example 2, the paired randomization test $p$-value $p$ based on the difference-of-means statistic after optimal Mahalanobis matching satisfies $P(p < \alpha) \to 1$ for all $\alpha \in (0, 1)$.

Proof. Write $Y = \theta^T X + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ independently of $(X, Z)$. By rotation invariance, the distribution of $\eta = \theta^T X$ is the same for every unit vector $\theta$. In particular, by picking $\theta = (1,0)$, we conclude that the law of $\theta^T X$ is the distribution of the first coordinate of a uniformly random point on the disc, which has density $f_\eta(h) = \frac{1}{2} \sqrt{1 - h^2}$. Using this fact, we may compute the conditional density of $\eta$ given $Z = z$:

$$f_{\eta|Z=1}(h) = \frac{2}{\pi} (1 + h) \sqrt{1 - h^2} \quad \text{and} \quad f_{\eta|Z=0}(h) = \frac{2}{\pi} \left(1 - \frac{7}{13} h\right) \sqrt{1 - h^2}.$$

Therefore, the law of large numbers and Slutsky’s theorem give:

$$\frac{1}{N_1} \sum_{Z_i=1} \theta^T X_i \xrightarrow{P} \mathbb{E}[\eta | Z = 1] = \int_{-1}^1 h f_{\eta|Z=1}(h) dh = \frac{1}{4}.$$

We will show that in large samples, $\frac{1}{N_1} \sum_{Z_i=1} \theta^T X_m(i)$ will be asymptotically bounded away from $1/4$. We do this by showing that $- \theta^T X_i$, with high probability in large samples — the average of even the $N_1$ largest values of $\theta^T X_i$ among untreated units will still be smaller than $1/4$. This is roughly the approach taken in [53] and [56].

Let $L$ be the indices of the $N_1$ untreated observations with the largest values of $\theta^T X_i$, with $L = \emptyset$ when there are fewer than $N_1$ untreated observations. We claim that, with high probability in large enough
samples, $\mathcal{L}$ will include all untreated values of $\theta^T X_i$ larger than $-0.22$. This follows from the following calculation:

$$\frac{\sum_{Z_i=0} \mathbb{I}\{\theta^T X_i \geq -0.22\}}{N_1} = \frac{\sum_{Z_i=0} \mathbb{I}\{\theta^T X_i \geq -0.22\}}{N_1/N_0} \overset{P}{\to} P(\eta \geq -0.22 | Z = 0) \frac{0.35/0.65}{0.533 < 0.539 < 1}.$$

The above implies that $\sum_{Z_i=0} \mathbb{I}\{\theta^T X_i \geq -0.22\}/N_1 \leq 1 + \zeta$ with probability approaching one for some small $\zeta > 0$, so $\sum_{Z_i=0} \mathbb{I}\{\theta^T X_i \geq -0.22\} < N_1$ with probability approaching one. On this event, the $N_1$ largest untreated units must contain all untreated units with $\theta^T X_i \geq -0.22$. Thus, we may write:

$$\frac{1}{N_1} \sum_{Z_i=1} \theta^T X_m(i) \leq \frac{1}{N_1} \sum_{Z_i=1} \theta^T X_i \leq \frac{\sum_{Z_i=0} \theta^T X_i \mathbb{I}\{\theta^T X_i \geq -0.22\}}{\sum_{Z_i=0} \mathbb{I}\{\theta^T X_i \geq -0.22\}} \overset{P}{\to} E[\eta | Z = 0, \eta \geq -0.22] < 0.242.$$

Thus, we may conclude:

$$\frac{1}{N_1} \sum_{Z_i=1} \{Y_i - Y_{m(i)}\} = \frac{1}{N_1} \sum_{Z_i=1} \theta^T X_i - \frac{1}{N_1} \sum_{Z_i=1} \theta^T X_m(i) + \frac{1}{N_1} \sum_{Z_i=1} \{\varepsilon_i - \varepsilon_{m(i)}\} \geq 0.08 - o_P(1).$$

Now $P(\theta < \alpha) \to 1$ may be derived using the same calculations as in the proof of Lemma 8.

\[\square\]

A.4 Details for Example 3

The main goal of this section is to follow Proposition, which was stated in the Example. Along the way, we will build up several important technical tools that will be used repeatedly in subsequent proofs.

**Proposition 4.** Suppose that $P$ satisfies Assumption 1, Fisher’s sharp null hypothesis, and $P(\epsilon(X) < 0.5) = 1$. Then the paired randomization test based on the difference-of-means statistic is asymptotically valid if and only if $\sum_{Z_i=1} (\mu(X_i) - \mu(X_{m(i)}))/N_1 = o_P(n^{-1/2}).$

A.4.1 Preparation

**Lemma 10.** Suppose that $P$ satisfies Assumption 1 and $P(\epsilon(X) < 0.5) = 1$. Then the average matching discrepancy under optimal Mahalanobis matching vanishes, $\frac{1}{N_1} \sum_{Z_i=1} ||X_i - X_{m(i)}|| = o_P(1).$ As a result, $||\bar{X}_1 - \bar{X}_0|| = o_P(1).$

**Proof.** Let $\varepsilon > 0$ be arbitrary. We begin by proving that there exists a (possibly suboptimal) matching scheme $\bar{m}$ satisfying $\frac{1}{N_1} \sum_{Z_i=1} ||X_i - X_{m(i)}|| \leq \sqrt{2\varepsilon E[||X||^2]/p^2} + o_P(1)$, where $p = P(Z = 1)$. To construct this matching, we use a similar approach to the proof of Proposition 2 in [3]. Let $B$ be a number so large that $P(X \notin [-B, B]^d) \leq \varepsilon/4$. Then, partition $[-B, B]^d$ into $N_n = (2B)^d/n^d$ cubes $C_1, \ldots, C_{N_n}$ of side length $r_n$ where $r_n \sim 1/\log(n)$. Call a set $C_i$ “good” if $P(X \in C_i) \geq r_n^2$ and let $\mathcal{G}$ be the union of the $C_i$’s. A simple volume argument shows that $P(X \notin \mathcal{G}) \leq \varepsilon/4 + o(1).$

$$P(X \notin \mathcal{G}) = P(X \notin [-B, B]^d) + \sum_{i : P(C_i) < r_n^2} P(C_i)$$

$$\leq \varepsilon/4 + N_n r_n^2d$$

$$\leq \varepsilon/4 + (2B)^d r_n^d$$

$$= \varepsilon/4 + o(1).$$

Since $\epsilon(X) < \frac{1}{2}$ almost surely, we may find $\delta > 0$ so small that the set $\mathcal{X}_0 = \{x : \epsilon(x) \leq \frac{1}{2} - \delta\}$ has probability at least $1 - \varepsilon/4$. Now define $D_i = C_i \cap \mathcal{X}_0$ and call $D_i$ “good” if $C_i$ was good. Let $\mathcal{G}_0$ be the union of the good $D_i$’s. Then $P(X \notin \mathcal{G}_0) \leq P(X \notin \mathcal{X}_0) + P(X \notin \mathcal{G}) \leq \varepsilon/2 + o(1)$. For all large $n$, this is less than $\varepsilon$. 

20
We will prove that, with high probability, all good $D_j$'s have more untreated than treated observations in them. First, we show this holds in expectation for each such $D_j$.

$$\Delta_{j,n} = E \left[ \sum_{i=1}^{n} I\{X_i \in D_j, Z_i = 1\} - \sum_{i=1}^{n} I\{X_i \in D_j, Z_i = 0\} \right]$$

$$= nP(X \in D_j, Z = 1) - P(X \in D_j, Z = 0)$$

$$= n \int_{D_j} \{2e(x) - 1\} dP_X(x)$$

$$\leq n \int_{D_j} -2\delta P_X(x)$$

$$\leq -n\delta r^2 \frac{d}{n}$$

Hoeffding’s inequality and the union bound then imply:

$$\mathbb{P}(D_j \text{ has more treated than untreated for any good } D_j) \leq N_n e^{-n\delta^2 r^2 d^2/2}.$$  

This upper bound tends to zero, so we may conclude that all good sets $D_j$ have an abundance of untreated units with probability tending to one.

Finally, we are ready to define the matching $\bar{m}$. It begins by sequentially matching each treated observation in a “good” region $D_j$ to an untreated observation in the same region by some arbitrary rule. With high probability, this is feasible without re-using any untreated observations. Since $P(Z = 1) < \frac{1}{2}$, leftover untreated observations exist with high probability even after matching all treated observations in good regions. We use these to provide matches for the remaining treated observations not belonging to good regions.

Under this matching scheme, we have:

$$\frac{1}{N_1} \sum_{Z_i = 1} ||X_i - X_{\bar{m}(i)}|| \leq \frac{1}{N_1} \sum_{Z_i = 1} X_i \in G_0 ||X_i - \bar{m}(i)|| + \frac{1}{N_1} \sum_{X_i \notin G_0} ||X_i - \bar{m}(i)||$$

$$\leq \frac{1}{N_1} \sum_{Z_i = 1} X_i \in G_0 \sum_{i=1}^{n} Z_i ||X_i - \bar{m}(i)||$$

$$\leq \sqrt{\delta r_n} + \left( \frac{1}{N_1} \sum_{Z_i = 1} X_i \in G_0 \right)^{1/2} \left( \frac{1}{N_1} \sum_{i=1}^{n} Z_i ||X_i - \bar{m}(i)|| \right)^{1/2}$$

$$\leq o(1) + (\varepsilon/p + o_P(1))^{1/2} \left( \frac{1}{N_1} \sum_{i=1}^{n} 2Z_i ||X_i||^2 + ||\bar{m}(i)||^2 \right)^{1/2}$$

$$= o(1) + (\varepsilon/p + o_P(1))^{1/2} \left( \frac{1}{N_1} \sum_{i=1}^{n} 2||X_i||^2 \right)^{1/2}$$

$$= \sqrt{2\varepsilon \mathbb{E}[||X||^2]/p^2} + o_P(1).$$

Hence, $\frac{1}{N_1} \sum_{Z_i = 1} ||X_i - \bar{m}(i)|| \leq \sqrt{2\varepsilon \mathbb{E}[||X||^2]/p^2} + o_P(1)$.

Since the sample covariance matrix $\mathbf{\Sigma}$ converges to the population covariance matrix $\Sigma = \text{Var}(X)$, all Euclidean and (estimated) Mahalanobis distances are within a factor of $2\kappa(\Sigma)$ of one another with probability approaching one. Here, $\kappa(\Sigma) = \lambda_{\text{max}}(\Sigma)/\lambda_{\text{min}}(\Sigma)$ is the condition number $\Sigma$. Thus, for the matching $\bar{m}$ defined above, we have:

$$\frac{1}{N_1} \sum_{Z_i = 1} d_M(X_i, X_{\bar{m}(i)}) \leq 2\kappa(\Sigma)\sqrt{2\varepsilon \mathbb{E}[||X||^2]/p^2} + o_P(1).$$
Since the optimal Mahalanobis matching achieves smaller total Mahalanobis discrepancy than \( \bar{m} \), the same upper bound holds for \( \frac{1}{N_1} \sum_{Z_i=1} d_M(X_i, X_{m(i)}) \). Since \( \varepsilon \) is arbitrary, this implies \( \frac{1}{N_1} \sum_{Z_i=1} d_M(X_i, X_{m(i)}) = o_P(1) \). Hence, the average Mahalanobis discrepancy from optimal matching vanishes. Since Mahalanobis discrepancy and Euclidean discrepancy are comparable, the average Euclidean discrepancy vanishes as well.

Finally, since \( ||X_1 - X_0|| \leq \frac{1}{N_1} \sum_{Z_i=1} ||X_i - X_{m(i)}|| \), we may also conclude that optimal matching successfully balances covariates under these assumptions.

\[ \text{Lemma 11.} \quad \text{Assume that } P \text{ satisfies Assumption 1 and } P(\varepsilon(X) < 0.5) = 1. \text{ For any } q \in [1, \infty), \text{ let } g : \mathbb{R}^d \to \mathbb{R} \text{ be some function with } \mathbb{E}[|g(X)|^q] < \infty. \text{ Then under optimal Mahalanobis matching, we have:} \]

\[
\frac{1}{N_1} \sum_{Z_i=1} |g(X_i) - g(X_{m(i)})|^q \xrightarrow{P} 0.
\]

**Proof.** Let \( \varepsilon > 0 \) be arbitrary. Since bounded Lipschitz functions are dense in \( L^q \), there exists a bounded \( L \)-Lipschitz function \( h \) such that \( \mathbb{E}[|g(X) - h(X)|^q] < \varepsilon \). By the AM-GM inequality, there exists a constant \( C(q) < \infty \) such that the following hold:

\[
\frac{1}{N_1} \sum_{Z_i=1} |g(X_i) - g(X_{m(i)})|^q = \frac{1}{N_1} \sum_{Z_i=1} |h(X_i) - h(X_{m(i)})| + |g(X_i) - h(X_i)| + |h(X_{m(i)}) - g(X_{m(i)})|^q \\
\leq C(q) \frac{1}{N_1} \sum_{Z_i=1} |h(X_i) - h(X_{m(i)})|^q \\
+ C(q) \left( \frac{1}{N_1} \sum_{Z_i=1} |g(X_i) - h(X_i)|^q + \frac{1}{N_1} \sum_{Z_i=1} |g(X_{m(i)}) - h(X_{m(i)})|^q \right) \\
\leq C(q) \frac{1}{N_1} \sum_{Z_i=1} L||X_i - X_{m(i)}|| + C(q) \frac{1}{N_1} \sum_{Z_i=1} |g(X_i) - h(X_i)|^q \\
\leq C(q) \varepsilon/p + o_P(1).
\]

Since \( \varepsilon \) is arbitrary, this proves the result.

\[ \text{Lemma 12.} \quad \text{Assume that } P \text{ satisfies Assumption 1 and } P(\varepsilon(X) < 0.5) = 1. \text{ Let } \sigma_\ast = \frac{1}{N_1} \sum_{i \in M} \sigma^2(X_i). \text{ Let } \hat{q}_{1-\alpha} := \inf \{ t \in \mathbb{R} : P(\hat{\tau}_\ast \leq t \mid D) \geq 1 - \alpha \} \text{ be the randomization critical value. Then we have:} \]

\[
\hat{q}_{1-\alpha} = \frac{Z_{1-\alpha} \sigma^2}{\sqrt{N_1}} + o_P(1/\sqrt{n})
\]

**Proof.** Let \( \sigma_1, \sigma_2, \ldots \) be i.i.d. Rademacher(\( \frac{1}{2} \)) random variables independent of the original data \( D \). Conditional on \( D \), \( \hat{\tau}_\ast \) has the same distribution as \( \frac{1}{N_1} \sum_{Z_i=1} \sigma_i (Y_i - Y_{m(i)}) \). Let \( \varepsilon_i = Y_i - \mu(X_i) \). Then we have the following:

\[
N_1 \text{Var}(\hat{\tau}_\ast \mid D) = \frac{1}{N_1} \sum_{Z_i=1} \{ Y_i - Y_{m(i)} \}^2 \\
= \frac{1}{N_1} \sum_{Z_i=1} \{ \varepsilon_i - \varepsilon_{m(i)} \}^2 + \frac{2}{N_1} \sum_{Z_i=1} \{ \mu(X_i) - \mu(X_{m(i)}) \} \{ \varepsilon_i - \varepsilon_{m(i)} \} \\
= \frac{1}{N_1} \sum_{Z_i=1} \{ \varepsilon_i - \varepsilon_{m(i)} \}^2 + o_P(1) \text{ by Lemma 11} \\
= \frac{1}{N_1} \sum_{i \in M} \sigma^2(X_i) + \frac{1}{N_1} \sum_{Z_i=1} \{ \varepsilon_i - \varepsilon_{m(i)} \}^2 - \sigma^2(X_i) - \sigma^2(X_{m(i)}) + o_P(1) \\
= \sigma^2 + o_P(1) \text{ by Chebyshev conditional on } \{(X_i, Z_i) \}_{i \leq n}
\]
Since \( \sigma_*^2 \geq \frac{1}{N_1} \sum Z_i = 1 \sigma^2(X_i) \overset{P}{\to} \mathbb{E}[\sigma^2(X)|Z = 1] > 0 \), the quantity \( N_1 \text{Var}(\hat{\tau}^\text{DM}|D) \) is bounded from below in probability. In particular, \( 1/(N_1 \text{Var}(\hat{\tau}^\text{DM}|D))^{3/2} = O_P(1) \). Therefore, the Berry-Esseen theorem gives:

\[
\sup_{t \in \mathbb{R}} |\mathbb{P}(\hat{\tau}^\text{DM} \leq t \sqrt{\text{Var}(\hat{\tau}^\text{DM}|D)}) - \Phi(t)| \leq \frac{C}{N_1} \left( \frac{1}{N_1 \text{Var}(\hat{\tau}^\text{DM}|D))^{3/2}} \right)
\leq \frac{C}{N_1} \frac{1}{N_1 \text{Var}(\hat{\tau}^\text{DM}|D)} \times \frac{4n}{N_1 \sum_{i=1}^n |Y_i|^3}
\leq O_P(n^{-1/2}) \times O_P(1) \times O_P(1)
\overset{P}{\to} 0.
\]

Therefore, \( \hat{\tau}^\text{DM}/\sqrt{\text{Var}(\hat{\tau}^\text{DM}|D)} \sim H \) where \( H \sim N(0, 1) \). By the continuous mapping theorem for weak convergence in probability, \( |\hat{\tau}^\text{DM}/\sqrt{\text{Var}(\hat{\tau}^\text{DM}|D)}| \sim |H| \). Weak convergence in probability to a limit distribution with a positive density at its \((1 - \alpha)\)-th quantile implies convergence of that quantile (see [64, Lemma 23.3]), so the \((1 - \alpha)\)-th quantile of the randomization distribution of the studentized statistic \( \hat{\tau}^\text{DM}/\sqrt{\text{Var}(\hat{\tau}^\text{DM}|D)} \) is equal to \( z_{1-\alpha/2} + o_P(1) \). Hence, we have:

\[
\hat{q}_{1-\alpha} = \sqrt{\text{Var}(\hat{\tau}^\text{DM}|D)}(z_{1-\alpha/2} + o_P(1))
= \sqrt{N_1 \text{Var}(\hat{\tau}^\text{DM}|D)} \left( \frac{z_{1-\alpha/2}}{\sqrt{N_1}} + o_P(1/\sqrt{n}) \right)
= (\sigma_* + o_P(1)) \left( \frac{z_{1-\alpha/2}}{\sqrt{N_1}} + o_P(1/\sqrt{n}) \right)
= \frac{z_{1-\alpha/2} \sigma_*}{\sqrt{N_1}} + o_P(1/\sqrt{n}).
\]

\( \square \)

### A.4.2 Proof of the Proposition

**Proof.** Now, we are ready to prove Proposition 4.

First, we prove “sufficiency.” Let \( \sigma_* \) be as in Lemma 12. Assume that \( \sum_{Z_i = 1} \{\mu(X_i) - \mu(X_{m(i)})\}/N_1 = o_P(n^{-1/2}) \). Let \( \epsilon_i = Y_i - \mu(X_i) \). Then we may write:

\[
\sqrt{N_1 \hat{\tau}^\text{DM}}/\sigma_* = \frac{1}{\sqrt{N_1 \sigma_*}} \sum_{Z_i = 1} \{Y_i - Y_{m(i)}\}
= \frac{1}{\sqrt{N_1 \sigma_*}} \sum_{Z_i = 1} \{\mu(X_i) - \mu(X_{m(i)})\} + \frac{1}{\sqrt{N_1 \sigma_*}} \sum_{Z_i = 1} \{\epsilon_i - \epsilon_{m(i)}\}
= o_P(1) + \frac{1}{\sqrt{N_1 \sigma_*}} \sum_{Z_i = 1} \{\epsilon_i - \epsilon_{m(i)}\}.
\]

Now, we study the term \( \sum_{Z_i = 1} \{\epsilon_i - \epsilon_{m(i)}\}/\sqrt{N_1 \sigma_*} \). Observe that this term has mean zero and variance one conditionally on \( \{(X_i, Z_i)\}_{i \leq n} \). Therefore, the Berry-Esseen theorem (Lemma 2) gives:

\[
\mathbb{P} \left( \left| \frac{1}{\sqrt{N_1 \sigma_*}} \sum_{Z_i = 1} \{\epsilon_i - \epsilon_{m(i)}\} \right| \leq t \left| \{(X_i, Z_i)\}_{i \leq n} \right| \right) - \Phi(t) \leq \frac{C}{\sqrt{N_1 \sigma_*^2 n}} \sum_{Z_i = 1} \mathbb{E}[\epsilon_i - \epsilon_{m(i)}]^3 |\{(X_i, Z_i)\}_{i \leq n}|
\leq C' \frac{1}{\sqrt{N_1 \sigma_*^2 n}} \sum_{i=1}^n \mathbb{E}[|\epsilon_i|^3 |X_i, Z_i]
= O_P(1/\sqrt{n}) O_P(1) O_P(\mathbb{E}[|\epsilon_i|^3])
\overset{P}{\to} 0.
\]
Thus, $\Pr\left(\frac{1}{\sqrt{N_t}} \sum_{i=1}^{\infty} \{\varepsilon_i - \varepsilon_{m(i)}\} \leq t \big| \{(X_i, Z_i)\}_{i \leq n}\right) \xrightarrow{P} \Phi(t)$. Taking expectations on both sides proves that $\frac{1}{\sqrt{N_t}} \sum_{i=1}^{\infty} \{\varepsilon_i - \varepsilon_{m(i)}\} \xrightarrow{L} N(0, 1)$ unconditionally. Therefore, by Slutsky’s theorem, $\sqrt{N_t} \hat{\tau}_{DM}/\sigma_* \xrightarrow{D} \mathcal{N}(0, 1)$ as well. Lemma 15 and Slutsky’s theorem then give $\Pr(p < \alpha) \rightarrow \alpha$. This establishes the asymptotic validity of the paired randomization test under the assumption $\frac{1}{\sqrt{N_t}} \sum_{i=1}^{\infty} \{\mu(X_i) - \mu(X_{m(i)})\} = o_P(1/\sqrt{n})$.

Next, we prove “necessity.” Assume that $\frac{1}{\sqrt{N_t}} \sum_{i=1}^{\infty} \{\mu(X_i) - \mu(X_{m(i)})\} \neq o_P(n^{-1/2})$. Since $\sqrt{N_t} \hat{q}_{1-\alpha}/\sigma_* \xrightarrow{D} z_{1-\alpha}/2$ by Lemma 12, $\sqrt{N_t} \hat{q}_{1-\alpha}/\sigma_* \leq z_{1-\alpha}/2 + \zeta$ with probability approaching one for any $\zeta > 0$. Thus, we have:

$$\Pr\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} > \frac{\sqrt{N_t} \hat{q}_{1-\alpha}}{\sigma_*}\right) \geq \Pr\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} > z_{1-\alpha}/2 + \zeta \right) - o_P(1)$$

Taking expectations on both sides gives:

$$\mathbb{P}(\hat{\tau}_{DM} > \hat{q}_{1-\alpha}) = \mathbb{E}\left[\Pr\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} > \frac{\sqrt{N_t} \hat{q}_{1-\alpha}}{\sigma_*}\right)\right]$$

$$\geq \mathbb{E}\left[\Pr\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} > z_{1-\alpha}/2 + \zeta \right)\right] - o(1)$$

$$= \mathbb{E}\left[\mathbb{E}\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} \sum_{Z_i=1} \{\varepsilon_i - \varepsilon_{m(i)}\} > z_{1-\alpha}/2 + \zeta \right) \right] - \frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \left\{(X_i, Z_i)\}_{i \leq n}\right\} - \frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \left\{(X_i, Z_i)\}_{i \leq n}\right\} - o(1)$$

Since we have shown above that $\Pr(\frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\varepsilon_i - \varepsilon_{m(i)}\} < t \big| \{(X_i, Z_i)\}_{i \leq n})$ is uniformly close to $\Phi(t)$, we may use the bounded convergence theorem to further simplify the above expression as follows:

$$\mathbb{P}(\hat{\tau}_{DM} > \hat{q}_{1-\alpha}) \geq \mathbb{E}\left[\mathbb{E}\left(\frac{\sqrt{N_t} \hat{\tau}_{DM}}{\sigma_*} \sum_{Z_i=1} \{\varepsilon_i - \varepsilon_{m(i)}\} > z_{1-\alpha}/2 + \zeta \right) \right] - \frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \left\{(X_i, Z_i)\}_{i \leq n}\right\} - \frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \left\{(X_i, Z_i)\}_{i \leq n}\right\} - o(1)$$

In the last line, $\mathbb{H}$ is a standard normal random variable independent of everything else. Now we use the fact that $t \rightarrow \Pr(|\mathbb{H} + t| > c)$ is symmetric and strictly larger than $\Pr(|\mathbb{H}| > c)$ for any $t \neq 0$. By our assumption that $\frac{1}{\sqrt{N_t}} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \neq o_P(n^{-1/2})$ and the fact that $1/\sigma_* = O_P(1)$, we may deduce that $\frac{1}{\sqrt{N_t}} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} \neq o_P(1)$. Thus, there exists $\eta > 0$ such that $\Pr(\frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} > \eta) > \eta$ for arbitrarily large values of $n$. Thus, with some constant probability, $\Pr(|\mathbb{H} + \frac{1}{\sqrt{N_t} \sigma_*} \sum_{Z_i=1} \{\mu(X_i) - \mu(X_{m(i)})\} > z_{1-\alpha}/2 + \zeta) > \eta$ and it is never smaller. Thus, we have:

$$\liminf \Pr(\hat{\tau}_{DM} > \hat{q}_{1-\alpha}) \geq \Pr(|\mathbb{H}| > z_{1-\alpha}/2 + \zeta) + \nu$$

for some $\nu > 0$ depending only on $\eta$. Now take $\zeta$ down to zero to conclude $\Pr(\hat{\tau}_{DM} > \hat{q}_{1-\alpha}) > \alpha$. \qed
A.5 Proof of Proposition 1

A.5.1 Preparation

**Lemma 13.** Suppose that $P$ satisfies Assumption 1. Let $\mathbf{G} = \frac{1}{N_1} \sum_{i \in M} (Z_i, 1, X_i)(Z_i, 1, X_i)^\top$ be the design matrix used in the regression (6). Then there exists finite, positive constants $\lambda_{\min}, \lambda_{\max}$ such that $[\lambda_{\min}(\mathbf{G}), \lambda_{\max}(\mathbf{G})] \subset [\Lambda_{\min}/\max_{i \leq n} K_{i,n}, \Lambda_{\max}]$ with probability tending to one. Here, $K_{i,n}$ is the number of treated observations $j$ for which observation $i$ is the nearest untreated neighbor in the Mahalanobis distance $d_M$.

**Proof.** We start with the upper bound, which is much easier. Set $p = P(Z = 1)$ and write:

$$\hat{\mathbf{G}} \preceq \frac{1}{N_1} \sum_{i=1}^n (Z_i, 1, X_i)(Z_i, 1, X_i)^\top \overset{p}{\rightarrow} \frac{1}{p} \begin{bmatrix} p & p & \mathbb{E}[ZX]^\top \\ p & 1 & \mathbb{E}[X]^\top \\ \mathbb{E}[ZX] & \mathbb{E}[X] & \mathbb{E}[XX^\top] \end{bmatrix}$$

Hence, by the continuous mapping theorem, $\lambda_{\max}(\hat{\mathbf{G}})$ is bounded by twice the largest eigenvalue of the matrix on the right-hand side of the above display with probability tending to one.

Now we consider the lower bound, which is more complicated. For each treated observation $i$, let $\hat{m}(i) = \arg\min_{j; Z_j = 0} d_M(X_i, X_j)$ be its nearest untreated neighbor in the Mahalanobis distance. It is not hard to see that if an untreated observation is the nearest untreated neighbor of some treated observation, then it will be selected by the optimal Mahalanobis matching scheme. Thus, we may write:

$$\mathbf{G} \succeq \mathbf{G} = \frac{1}{N_1} \sum_{i=1}^n (Z_i, 1, X_i)(Z_i, 1, X_i)^\top + (Z_{\hat{m}(i)}, 1, X_{\hat{m}(i)})(Z_{\hat{m}(i)}, 1, X_{\hat{m}(i)})^\top$$

$$\succeq \frac{1}{N_1} \sum_{i=1}^n Z_i + (1 - Z_i)K_{i,n} \frac{(Z_i, 1, X_i)(Z_i, 1, X_i)^\top}{\max_{i \leq n} K_{i,n}}$$

$$= \frac{1}{N_1} \sum_{i=1}^n Z_i(1, 1, X_i)(1, 1, X_i)^\top + \frac{1}{N_1} \sum_{i=1}^n (1 - Z_i)K_{i,n}(0, 1, X_i)(0, 1, X_i)^\top$$

The two matrices in the lower bound both converge, the first by the law of large numbers and the second by Lemma 20. In particular, their sum converges to the following limit:

$$\frac{1}{N_1} \sum_{i=1}^n (Z_i + (1 - Z_i)K_{i,n})(Z_i, 1, X_i)(Z_i, 1, X_i)^\top \overset{p}{\rightarrow} \begin{bmatrix} 1 & 1 & \mathbb{E}[X^\top|Z = 1] \\ 1 & 2 & 2\mathbb{E}[X^\top|Z = 1] \\ \mathbb{E}[X|Z = 1] & 2\mathbb{E}[X|Z = 1] & 2\mathbb{E}[XX^\top|Z = 1] \end{bmatrix}$$

This limiting matrix is invertible, since it twice the outer-product matrix $\mathbb{E}[(Z^*, 1, X)(Z^*, 1, X)^\top|Z = 1]$ where $Z^* \sim \text{Bernoulli}(\frac{1}{2})$ independently of $(X, Z)$. Under Assumption 1, this is positive definite, and its smallest eigenvalue will be called $2\Lambda_{\min}$. Again by the continuous mapping theorem, we obtain $\lambda_{\min}(\mathbf{G}) \geq \Lambda_{\min}/\max_{i \leq n} K_{i,n}$ with probability tending to one.

**Lemma 14.** Assume the conditions of Proposition 1. Let $\hat{\tau}$ be either the analysis of covariance test statistic (P1) or the residual difference-of-means statistic (P2). Then $\sqrt{N_1} \hat{\tau} = \frac{1}{\sqrt{N_1}} \sum_{i=1}^n \{\varepsilon_i - \varepsilon_{m(i)}\} + o_P(1)$.

**Proof.** First, we prove the result for the analysis of covariance test statistic (P1). For convenience, we reparameterize $Z_i$ to take values in $\{\pm \frac{1}{2}\}$ rather than in $\{0, 1\}$ in this part of the proof, with $Z_i = +\frac{1}{2}$ indicating that the $i$-th observation is treated. This has no effect on the test statistic or its asymptotic distribution, but will simplify our calculations.

Let $\varepsilon_i = Y_i - \gamma - \beta^\top X_i$. By Lemma 13, the matched design matrix $\mathbf{G}$ is invertible with probability tending to one (and this remains true after our reparameterization), so standard linear regression formulas give the following formula for $\hat{\tau}$:

$$\sqrt{N_1} \hat{\tau} = \epsilon_1^\top \mathbf{G}^{-1} \frac{1}{\sqrt{N_1}} \sum_{i \in M} \varepsilon_i (Z_i, 1, X_i)^\top.$$
To get a better handle on this expression, write the matrix $\hat{G}^{-1}$ as follows:

$$
\hat{G}^{-1} = \begin{bmatrix}
\frac{1}{2} & 0 & \frac{1}{2}(\bar{X}_1 - \bar{X}_0)^T \\
0 & 2 & \frac{1}{2}(\bar{X}_1 + \bar{X}_0)^T \\
\frac{1}{2}(\bar{X}_1 - \bar{X}_0) & (\bar{X}_1 + \bar{X}_0) & \frac{1}{N} \sum_{i \in M} X_i X_i^T
\end{bmatrix}^{-1}
$$

Let $H$ be the submatrix of $\hat{G}$ obtained by deleting the first row and column. Then $H > \frac{1}{N} \sum_{i=1}^{N} (1, X_i)(1, X_i)^T = \mathbb{E}[(1, X)(1, X)^T | Z = 1] + o_P(1)$, which is positive definite by the nonsingularity condition in Assumption 1. Hence, $H^{-1}$ exists with high probability and $\lambda_{\text{max}}(H^{-1}) = O_P(1)$. Applying the partitioned matrix inversion formula then gives:

$$
\hat{G}^{-1} = \begin{bmatrix}
\frac{1}{2} & 0 & \frac{1}{2}(\bar{X}_1 - \bar{X}_0) \\
0 & 2 & \frac{1}{2}(\bar{X}_1 + \bar{X}_0) \\
\frac{1}{2}(\bar{X}_1 - \bar{X}_0) & (\bar{X}_1 + \bar{X}_0) & \frac{1}{N} \sum_{i \in M} X_i X_i^T
\end{bmatrix}^{-1}
$$

where $s = (\frac{1}{2} - (0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0))^T H^{-1} (0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0))(0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0))^T H^{-1}$, Under the assumption that $||\bar{X}_1 - \bar{X}_0|| = o_P(1)$, $s$ converges in probability to $2$ and $-s(0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0))H^{-1}$ converges in probability to $(0, 0)$. Hence, we have:

$$
\hat{G}^{-1} \varepsilon_1 = \begin{pmatrix}
-s(0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0))H^{-1} \\
-s(0, \frac{1}{2}(\bar{X}_1 - \bar{X}_0)) + \frac{1}{\sqrt{N}} \mathbb{E}(2, 0, (Z_i, 1, X_i)) \to (\frac{2}{0})
\end{pmatrix}
$$

Combining this with our representation for $\sqrt{N_1} \tau$ (and the fact that $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i(Z_i, 1, X_i) = O_P(1)$ by Chebyshev’s inequality conditional on $\{(X_i, Z_i)\}_{i \leq n}$ gives:

$$
\sqrt{N_1} \tau = \frac{1}{\sqrt{N_1}} \sum_{i \in M} \varepsilon_i((2, 0), (Z_i, 1, X_i)) + o_P(1)
$$

Next, we prove the result for the residual difference-of-means statistic. We return to the original parameterization where $Z_i \in \{0, 1\}$. We begin by showing that, in the regression of $Y$ on $X$ without $Z$, the estimated coefficient $\hat{\beta}$ is $\sqrt{n}$-consistent for the true vector $\beta$. Again, we use standard linear regression formulas to write:

$$
\begin{pmatrix}
\hat{\beta} - \gamma \\
\hat{\beta} - \beta
\end{pmatrix} = \left( \frac{1}{N} \sum_{i \in M} (1, X_i)(1, X_i)^T \right)^{-1} \frac{1}{N_1} \sum_{i \in M} \varepsilon_i(1, X_i)^T
$$

The matrix $\frac{1}{N} \sum_{i \in M} (1, X_i)(1, X_i)^T$ is at least as large as $\frac{1}{N} \sum_{i=1}^{N} (1, X_i)(1, X_i)^T$, and this lower bound converges to the invertible matrix $\mathbb{E}[(1, X)(1, X)^T | Z = 1]$ as the sample size grows large. Hence, the inverted matrix in the preceding display exists with probability tending to one and has operator norm of size $O_P(1)$. Meanwhile, $\frac{1}{N} \sum_{i \in M} \varepsilon_i(1, X_i)^T = O_P(1/\sqrt{N_1}) = O_P(1/\sqrt{n})$ by Chebyshev’s inequality applied conditionally on $\{(X_i, Z_i)\}_{i \leq n}$. Here, we need the condition in Assumption 1 that $||X||$ and $Y$ have more than four moments. Combining these two estimates gives $||\hat{\beta} - \beta|| = O_P(1/\sqrt{n})$.

Now we may write:

$$
\begin{align*}
\sqrt{N_1} \tau &= \frac{1}{\sqrt{N_1}} \sum_{Z_i=1} \{\varepsilon_i - \bar{\epsilon}_{m(i)}\} \\
&= \frac{1}{\sqrt{N_1}} \sum_{Z_i=1} Y_i - Y_{m(i)} - \hat{\beta}^T \{X_i - X_{m(i)}\} \\
&= \frac{1}{\sqrt{N_1}} \sum_{Z_i=1} \{\varepsilon_i - \bar{\epsilon}_{m(i)}\} + \sqrt{N_1(\hat{\beta} - \beta)^T (\bar{X}_1 - \bar{X}_0)} \\
&\quad = O_P(1)
\end{align*}
$$
Thus, both test statistics have the claimed asymptotically linear expansion.

\[ \hat{q}_{1-\alpha} = \inf\{t \in \mathbb{R} : \mathbb{P}(\hat{r}_* \leq t \mid \mathcal{D}) \geq 1 - \alpha\}. \]

Further define \( \sigma_*^2 = \sum_{i \in \mathcal{M}} \sigma^2(X_i)/N_1^2 \). Then we have \( \hat{q}_{1-\alpha} = z_{1-\alpha/2}\sigma_* + o_P(1/\sqrt{n}) \).

**Proof:** At a high level, this result follows by applying the Berry-Esseen theorem (conditionally on \( \mathcal{D} \)) to a linearization of \( \hat{r}_* \).

First, the linearization. Let \( \hat{X}_* = \frac{1}{N_1} \sum_{i \in \mathcal{M}} Z_i X_i \). Then \( \|\hat{X}_* - X_o\| = o_P(1) \) by Chebyshev’s inequality conditional on the original data. By repeating the proof of Lemma 14 with \( Z_i^* \) in place of \( Z_i \) and \( X_*^* \) in place of \( X_* \), we may conclude:

\[ \hat{r}_* = \frac{1}{N_1} \sum_{i=1}^{N_1} (2Z_i^* - 1)\{\varepsilon_i - \varepsilon_{m(i)}\} + o_P(1/\sqrt{n}). \]

This expansion holds for either of the two test statistics under consideration.

For each \( i \) with \( Z_i = 1 \), let \( \xi_i = (2Z_i^* - 1)\{\varepsilon_i - \varepsilon_{m(i)}\} \). Then conditional on the original observations \( \mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i \leq n} \), the random variables \( \xi_i \) are mean zero and independent. Therefore, the sum \( \frac{1}{N_1} \sum_{i=1}^{N_1} \xi_i \) has (conditional) variance \( \hat{\sigma}_*^2 \), where \( \hat{\sigma}_*^2 \) satisfies:

\[ N_1 \hat{\sigma}_*^2 = \frac{1}{N_1} \sum_{Z_i = 1} \{\varepsilon_i - \varepsilon_{m(i)}\}^2 \]

\[ = \frac{1}{N_1} \sum_{Z_i = 1} \varepsilon_i^2 + \varepsilon_{m(i)}^2 + \frac{1}{N_1} \sum_{Z_i = 1} 2\varepsilon_i \varepsilon_{m(i)} \]

\[ = \frac{1}{N_1} \sum_{i \in \mathcal{M}} \varepsilon_i^2 + \frac{1}{N_1} \sum_{Z_i = 1} 2\varepsilon_i \varepsilon_{m(i)} \]

\[ = \frac{1}{N_1} \sum_{i \in \mathcal{M}} \sigma^2(X_i) + o_P(1) \]

where the last step follows by applying Chebyshev’s inequality conditionally on \( \{(X_i, Z_i)\}_{i \leq n} \). In particular, since \( \sigma^2(X_i) \) is bounded away from zero, \( 1/(N_1 \hat{\sigma}_*^2) = O_P(1) \). This also implies \( \hat{\sigma}_*/\sigma_* = 1 + o_P(1) \).

Now, we apply the Berry-Esseen theorem (Lemma 2) to the sum \( \sum_{Z_i = 1} \xi_i / N_1 \) conditionally on \( \mathcal{D} \) to obtain:

\[ \sup_{t \in \mathbb{R}} \left| \mathbb{P} \left( \frac{1}{N_1} \sum_{Z_i = 1} \xi_i \leq \sigma_* t \mid \mathcal{D} \right) - \Phi(t) \right| \leq \frac{C \sum_{Z_i = 1} \mathbb{E}[|\xi_i|^3]\{D\}}{\hat{\sigma}_*^3} \]

\[ = \frac{C}{\sqrt{N_1}} \frac{1}{\sqrt{N_1 \hat{\sigma}_*^3}} \frac{1}{N_1} \sum_{Z_i = 1} |\varepsilon_i - \varepsilon_{m(i)}|^3 \]

\[ \leq \frac{C}{\sqrt{N_1}} \frac{1}{\sqrt{N_1 \hat{\sigma}_*^3}} \frac{1}{N_1} \sum_{i=1}^{n} 4|\varepsilon_i|^3 \]

\[ \xrightarrow{P} 0. \]

Hence, \( \frac{1}{N_1} \sum_{Z_i = 1} \xi_i / \hat{\sigma}_* \sim_N N(0, 1) \). By the randomization Slutsky theorem (Lemma 3), we may also conclude \( \hat{r}_*/\sigma_* \sim_N N(0, 1) \). The continuous mapping theorem then gives \( |\hat{r}_*/\sigma_*| \sim_N |H| \) where \( H \sim N(0, 1) \).
Since weak convergence in probability to a limit distribution implies convergence of all quantiles at which the density of the limit distribution is strictly positive, we conclude:

$$\hat{q}_{1-\alpha} = \inf\{t \in \mathbb{R} : P(|\hat{\tau}| \leq t | D) \geq 1 - \alpha\}$$

$$= \sigma_* \inf\{t \in \mathbb{R} : P(|\hat{\tau}/\sigma_*| \leq t | D) \geq 1 - \alpha\}$$

$$= \sigma_*(z_{1-\alpha/2} + o_P(1))$$

$$= z_{1-\alpha/2} \sigma_* + o_P(1/\sqrt{n}).$$

\[ \square \]

### A.5.2 Proof of the Proposition

**Proof.** In this proof, $\hat{\tau}$ denotes either of the two test statistics mentioned in the Proposition.

First, we prove the asymptotic normality of $\hat{\tau}/\sigma_*$, where $\sigma_*$ is defined in Lemma 15. Since $\sigma_* = O_P(1/\sqrt{n})$, Lemma 14 allows us to write:

$$\hat{\tau}/\sigma_* = \frac{1}{N_1} \sum_{Z_i=1} \{\xi_i - \varepsilon_{m(i)}\}/\sigma_* + o_P(1).$$

For each $i$ with $Z_i = 1$, let $\xi_i = \{\xi_i - \varepsilon_{m(i)}\}/N_1 \sigma_*$. Conditional on $\{(X_i, Z_i)\}_{i \leq n}$, the $\xi_i$’s are mean zero and independent. Therefore, the Berry-Esseen theorem (Lemma 2) gives:

$$P\left(\frac{1}{N_1} \sum_{Z_i=1} \xi_i \leq t \right| \{(X_i, Z_i)\}_{i \leq n}) - \Phi(t) \leq C \sum_{Z_i=1} \mathbb{E}[|\xi_i|^3] \{(X_i, Z_i)\}_{i \leq n}$$

$$= \frac{C}{\sqrt{N_1} (\sqrt{N_1} \sigma_*)^3} \frac{1}{N_1} \sum_{Z_i=1} \mathbb{E}[|\xi_i - \varepsilon_{m(i)}|^3] \{(X_i, Z_i)\}_{i \leq n}$$

$$\leq \frac{C}{\sqrt{N_1} (\sqrt{N_1} \sigma_*)^3} \frac{1}{N_1} \sum_{i=1}^n 4\mathbb{E}[|\xi_i|^3 | X_i, Z_i]$$

$$= O_P(1/\sqrt{n}) O_P(1) O_P(\mathbb{E}[|\xi|^3])$$

(Markov)

$$\xrightarrow{P} 0$$

Thus, $P(\sum_{Z_i=1} \xi_i/N_1 \leq t | \{(X_i, Z_i)\}) \xrightarrow{P} \Phi(t)$ for each $t$, and by the in-probability dominated convergence theorem, we may also conclude $P(\sum_{Z_i=1} \xi_i/N_1 \leq t) \rightarrow \Phi(t)$. Since this holds for every $t$, we have shown $\sum_{Z_i=1} \xi_i/N_1 \rightarrow N(0, 1)$, so by Slutsky’s theorem, $\hat{\tau}/\sigma_* \rightarrow N(0, 1)$ as well.

Finally, we combine this with the asymptotics of the randomization critical value given in Lemma 15 to get asymptotic validity. For any $\alpha \in (0, 1)$, we have:

$$\mathbb{P}(p < \alpha) \leq \mathbb{P}(|\hat{\tau}| > \hat{q}_{1-\alpha})$$

$$= \mathbb{P}(|\hat{\tau}/\sigma_*| > \hat{q}_{1-\alpha}/\sigma_*)$$

$$= \mathbb{P}(|\hat{\tau}/\sigma_*| > z_{1-\alpha/2} + o_P(1))$$

$$\rightarrow \mathbb{P}_{H \sim N(0, 1)}(|H| > z_{1-\alpha/2})$$

$$= \alpha.$$

\[ \square \]

### A.6 Proof of Proposition 2

### A.6.1 Preparation

**Lemma 16.** Let $F_n = \{(X_i, Z_i)\}_{i \leq n}$. Assume that $\Delta_n \mid F_n \sim N(0, V(F_n))$ and that $\hat{q}_n \xrightarrow{P} z_{1-\alpha} \sqrt{\tau} > 0$. Then the following hold:
• $V(F_n) \geq v$ a.s. and $\mathbb{P}(V(F_n) \geq v + \eta) \geq \eta$ for some $\eta > 0$ imply $\limsup \mathbb{P}(|\Delta_n| > \hat{q}_n) > \alpha$.
• $V(F_n) \leq v$ a.s. and $\mathbb{P}(V(F_n) \leq v - \eta) \geq \eta$ for some $\eta > 0$ imply $\liminf \mathbb{P}(|\Delta_n| > \hat{q}_n) < \alpha$.

Proof. For any $\varepsilon > 0$, we may write the following:

\[
\mathbb{P}(|\Delta_n| > \hat{q}_n|F_n) \geq \mathbb{P}(|\Delta_n| > z_{1-\alpha} \sqrt{v} - \varepsilon|F - n) - \mathbb{P}(|\hat{q}_n - z_{1-\alpha} \sqrt{v}| > \varepsilon|F_n)
\]

\[
= \mathbb{P}(|\Delta_n| > z_{1-\alpha} \sqrt{v} - \varepsilon|F_n)\mathbb{I}(V(F_n) \geq v + \eta) + \mathbb{P}(|\Delta_n| > z_{1-\alpha} \sqrt{v} - \varepsilon|F_n)\mathbb{I}(V(F_n) < v + \eta)
\]

\[
+ \mathbb{P}(|\hat{q}_n - z_{1-\alpha} \sqrt{v}| > \varepsilon|F_n)
\]

\[
geq \mathbb{P}_{H \sim N(0,v+\eta)}(|H| > z_{1-\alpha} \sqrt{v} - \varepsilon)\mathbb{I}(V(F_n) \geq v + \eta)
\]

\[
+ \mathbb{P}_{H \sim N(0,v)}(|H| > z_{1-\alpha} \sqrt{v} - \varepsilon)\mathbb{I}(V(F_n) < v + \eta)
\]

\[
- \mathbb{P}(|\hat{q}_n - z_{1-\alpha} \sqrt{v}| > \varepsilon|F_n)
\]

In the last step, we used the fact that the probability that a centered normal random variable escapes the interval $[\pm z_{1-\alpha} \sqrt{v} - \varepsilon]$ is an increasing function of the variance. Now taking expectations on both sides gives:

\[
\mathbb{E}(|\Delta_n| > \hat{q}_n) \geq \mathbb{E}_{H \sim N(0,v+\eta)}(|H| > z_{1-\alpha} \sqrt{v} - \varepsilon)\mathbb{P}(V(F_n) \geq v + \eta)
\]

\[
+ \mathbb{E}_{H \sim N(0,v)}(|H| > z_{1-\alpha} \sqrt{v} - \varepsilon)\mathbb{P}(V(F_n) < v + \eta)
\]

\[
- \mathbb{E}(|\hat{q}_n - z_{1-\alpha} \sqrt{v}| > \varepsilon|F_n)
\]

For small enough $\varepsilon$, this lower bound is larger than $\alpha$. This proves the first claim. The proof of the second claim is analogous.

A.6.2 Proof of the Proposition

Proof. Assume that $||X_1 - \bar{X}_0|| \neq o_P(1)$. Then there exists $\eta > 0$ such that $P(||X_1 - \bar{X}_0|| > \eta) > \eta$ for arbitrarily large $n$. By passing to a subsequence if necessary, we may assume that this holds for all $n$.

Lemma 15 implies that the randomization critical value satisfies $\sqrt{N_{1}\hat{q}_{1-\alpha}} \overset{P}{\rightarrow} z_{1-\alpha} \sqrt{2\sigma^2}$, so by Lemma 16, it suffices to prove the following: (1) the conditional variance of the analysis of covariance test statistic (P1) is always at least $2\sigma^2$ and exceeds $2\sigma^2$ by a constant amount with constant probability; (2) the conditional variance of the residual difference-of-means statistic (P2) is always at most $2\sigma^2$ and is a constant amount below $2\sigma^2$ with constant probability.

We start with the analysis of covariance test statistic. Standard least-squares theory shows that $\sqrt{N_{1} \hat{r}} | \{(X_i, Z_i)\}_{i \leq n} \sim N(0, \sigma^2 \mathbf{1}_1 \mathbf{G}^{-1} \mathbf{1}_1)$ where $\mathbf{G}$ is the matched design matrix defined in Lemma 13. If we additionally define the matrix $\mathbf{H}$ by

\[
\mathbf{H} = \begin{bmatrix}
2 & (\bar{X}_1 + \bar{X}_0)^\top \\
(\bar{X}_1 + \bar{X}_0)^\top & \frac{1}{N_1} \sum_{i \in \mathcal{M}} X_i X_i^\top
\end{bmatrix},
\]

then the proof of Lemma 14 proves the following identity using the partitioned matrix inversion formula:

\[
e_1^\top \mathbf{G}^{-1} e_1 = \frac{2}{1 - \frac{1}{2} \lambda_{min}(H^{-1}) \eta^2} = \frac{2}{1 - \frac{1}{2} \eta^2 / \lambda_{max}(H)}.
\]

Note that this is always at least 2, and on the event where $||X_1 - \bar{X}_0|| \geq \eta$, we have the inequality:

\[
e_1^\top \mathbf{G}^{-1} e_1 \geq \frac{2}{1 - \frac{1}{2} \lambda_{min}(H^{-1}) \eta^2} = \frac{2}{1 - \frac{1}{2} \eta^2 / \lambda_{max}(H)}.
\]

The largest eigenvalue of $\mathbf{H}$ may be controlled as follows:

\[
\lambda_{max}(\mathbf{H}) \leq \text{Tr}(\mathbf{H}) = 2 + \text{Tr} \left( \frac{1}{N_1} \sum_{i \in \mathcal{M}} X_i X_i^\top \right) \leq 2 + \frac{n}{N_1} \text{Tr} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right) \overset{p}{\rightarrow} 2 + \frac{1}{p} \text{Tr}(\mathbb{E}[XX^\top]).
\]
Call this upper bound \( \bar{\lambda} \). Then with probability \( 1 - \eta - o(1) \), \( e_1^T \hat{G}^{-1} e_1 \geq 2/(1 - \eta^2/3 \bar{\lambda}) \). Hence, the conditional variance of \( \sqrt{N_1} \hat{\tau} \) is always at least \( 2\sigma^2 \) and, with nonvanishing probability, it is at least \( 2\sigma^2/(1 - \eta^2/3 \lambda) > 2\sigma^2 \). Hence, Lemma 16 implies that the paired randomization test based on the analysis of covariance test statistic is not asymptotically valid.

For the residual difference-of-means test statistic, we introduce some additional notation. Let \( X \in \mathbb{R}^{2N_1 \times (1+d)} \) be the matrix with rows \((1, X_i)\) for \( i \in \mathcal{M} \), let \( Z = (Z_i)_{i \in \mathcal{M}} \) (still keeping the parameterization \( Z_i \in \{ \pm \frac{1}{2} \}) \), and let \( e = (\varepsilon_i)_{i \in \mathcal{M}} \) for \( \varepsilon_i = Y_i - \gamma - \beta^T X_i \). Then we may write:

\[
\hat{\tau} = \frac{2}{N_1} Z^T (I - X(X^T X)^{-1} X^T) Y \\
= \frac{2}{N_1} Z^T (I - X(X^T X)^{-1} X^T)(X(\gamma, \beta^T)^T + e) \\
= \frac{2}{N_1} Z^T (I - X(X^T X)^{-1} X^T)e.
\]

Observe that \( X^T Z/N_1 = \frac{1}{2}(0, \bar{X}_1 - \bar{X}_0) \). Thus, we may write:

\[
\sqrt{N_1} \hat{\tau} \mid \{(X_i, Z_i)\}_{i \leq n} \sim N\left(0, \frac{4\sigma^2}{N_1} Z^T (I - X(X^T X)^{-1} X^T) Z \right)
\]

\[
= N\left(0, \frac{4\sigma^2}{N_1} Z^T Z - \frac{N_2^2}{4} (0, \bar{X}_1 - \bar{X}_0)^T (X^T X)^{-1} (0, \bar{X}_1 - \bar{X}_0) \right)
\]

\[
= N\left(0, 2\sigma^2 - \sigma^2 (0, \bar{X}_1 - \bar{X}_0)^T (X^T X/N_1)^{-1} (0, \bar{X}_1 - \bar{X}_0) \right).
\]

The quantity \( \sigma^2 (0, \bar{X}_1 - \bar{X}_0)^T (X^T X/N_1)^{-1} (0, \bar{X}_1 - \bar{X}_0) \) is never negative, and on the event where \( \| \bar{X}_1 - \bar{X}_0 \| > \eta \), it is at least \( \sigma^2 \lambda_{\min}(X^T X/N_1)^{-1}\eta^2 = \sigma^2 \eta^2 / \lambda_{\max}(X^T X/N_1) \). By the same argument that controls \( \lambda_{\max}(H) \), it can be shown that there exists \( \lambda' < \infty \) such that \( P(\lambda_{\max}(X^T X/N_1) \leq \lambda') \to 1 \). Thus, the conditional variance of \( \sqrt{N_1} \hat{\tau} \) is always at least \( 2\sigma^2 \), and with probability \( 1 - \eta - o(1) \) it is less than \( 2\sigma^2 (1 - \eta^2 / \lambda') \). Hence, Lemma 16 implies that the paired randomization test based on the residual difference-of-means test statistic is asymptotically conservative. \( \square \)

A.7 Proof of Proposition 3

A.7.1 Preparation

**Lemma 17.** For any bounded function \( g : \mathbb{R}^d \to \mathbb{R} \) and any sequence \( h_n \) with \( \sqrt{n} h_n \to h \), the sequences \( \{P_n\}_{n \geq 1} \) and \( \{P_{h_n,g}^n\}_{n \geq 1} \) are mutually contiguous.

Proof. This follows from a standard likelihood ratio calculation. See, e.g. [33]. \( \square \)

**Lemma 18.** Assume \( P \) satisfies Assumption 1 and \( P(\varepsilon(X) > 0.5) > 0 \). Let \( g(x) = I\{e(x) > 0.5\} \). Then there exists \( \eta > 0 \) such that \( \frac{1}{N_1} \sum_{Z_i=1} g(X_i) - g(X_{m(i)}) \geq \eta \) with probability approaching one for any pair matching scheme.

Proof. Let \( X_0 = \{ x : e(x) > 0.5 \} \) so that \( g(x) = I\{x \in X_0\} \). Then we have:

\[
\frac{1}{N_1} \sum_{Z_i=1} g(X_i) - g(X_{m(i)}) \geq \frac{1}{N_1} \sum_{Z_i=1} g(X_i) - \frac{1}{N_1} \sum_{Z_i=0} g(X_i) \\
\geq P_{\frac{1}{N_1} \sum_{Z_i=1} g(X_i)} P(Z = 1, X \in X_0) - P(Z = 0, X \in X_0) \\
= \frac{1}{p} \int_{X_0} \{e(x) - [1 - e(x)]\} dP(x).
\]

Since \( e(x) - [1 - e(x)] > 0 \) everywhere on \( X_0 \), the final line of this display is a strictly positive quantity, which we call \( 2\eta \). Therefore, \( \frac{1}{N_1} \sum_{Z_i=1} g(X_i) - g(X_{m(i)}) \geq \eta \) with probability approaching one. \( \square \)
A.7.2 Proof of the Proposition

Proof. First, we show that if \( P(\epsilon(X) < 0.5) = 1 \), then the paired randomization test based on either test statistic is robust to local nonlinearity. Let \( g \) be any bounded function and let \( h_n = O(1/\sqrt{n}) \). We will show that \( \mathbb{P}_{h_n,g}(p < \alpha) \rightarrow \alpha \) by proving that every subsequence \( \{n_k\} \) has a further subsequence \( \{n_{k_j}\} \) along which this convergence occurs. For notational convenience, we write \( P_n \) in place of \( P_{h_n,g} \).

Let \( \{n_k\} \) be given. Since \( \sqrt{n}h_n \) is bounded, we may extract a subsequence \( \{n_{k_j}\} \) along which \( \sqrt{n}h_n \to h \in \mathbb{R} \). Along this subsequence, Lemmas 15 and 17 give:

\[
\hat{q}_{1-\alpha} = z_{1-\alpha}/2\sigma \sqrt{2} + o_{P_n}(1/\sqrt{n})
\]

for either test statistic. Therefore, it suffices to show that \( \sqrt{N_1}\tau \xrightarrow{P_{n_{k_j}}} N(0,2\sigma^2) \) along the subsequence.

Let \( \varepsilon_i = Y_i - \gamma - \beta^T X_i \). By Lemma 14 and contiguity, the following expansions hold along the subsequence for either test statistic (P1) or (P2):

\[
\sqrt{N_1}\tau = \frac{1}{\sqrt{N_1}} \sum_{Z_{i,1}} \{ (Y_i - \gamma - \beta^T X_i) - (Y_{m(i)} - \gamma - \beta^T X_{m(i)}) \} + o_{P_n}(1) \quad (1)
\]

\[
= \frac{1}{\sqrt{N_1}} \sum_{Z_{i,1}} \sum_{Z_{i,1}} \{ (Y_i - \gamma - \beta^T X_i - h_n g(X_i)) - (Y_{m(i)} - \gamma - \beta^T X_{m(i)} - h_n g(X_{m(i)})) \}
\]

\[
+ \frac{1}{\sqrt{N_1}} \sum_{Z_{i,1}} h_n \{ g(X_i) - g(X_{m(i)}) \} + o_{P_n}(1) \quad (2)
\]

\[
= \frac{1}{\sqrt{N_1}} \sum_{Z_{i,1}} \{ (Y_i - \gamma - \beta^T X_i - h_n g(X_i)) - (Y_{m(i)} - \gamma - \beta^T X_{m(i)} - h_n g(X_{m(i)})) \}
\]

\[
\sim N(0,2\sigma^2)
\]

\[
\Rightarrow \quad P_{n_{k_j}}(1) \xrightarrow{\text{by Lemma 11 and contiguity}} N(0,2\sigma^2)
\]

Thus, by Slutsky’s theorem, we conclude that \( \mathbb{P}_{h_n,g}(p < \alpha) \to \alpha \) along the subsequence. Since the initial subsequence \( \{n_k\} \) was arbitrary, this convergence must also hold along the entire sequence \( n = 1, 2, 3, \ldots \) and we conclude that the paired randomization test is robust to local nonlinearity.

Next, we prove that if \( P(\epsilon(X) > 0.5) > 0 \), then the randomization test based on either (P1) or (P2) is not robust to local nonlinearity. The argument differs depending on which test statistic is used.

For the analysis of covariance statistic (P1), Proposition 2 implies that the test is not asymptotically valid even for the case \( h_n \equiv 0 \) unless \( ||X_1 - \bar{X}_0|| = O_P(1) \), so we may assume that this condition holds. Under this assumption, we can show that asymptotic validity fails along the sequence \( g(x) = 1(\epsilon(x) > 0.5), h_n = 1/\sqrt{n} \). The same arguments from above yield:

\[
\sqrt{N_1}\tau = \frac{1}{\sqrt{N_1}} \sum_{Z_{i,1}} \{ (Y_i - \gamma - \beta^T X_i - h_n g(X_i)) - (Y_i - \gamma - \beta^T X_i) \}
\]

\[
+ \frac{h}{\sqrt{N_1}} \sum_{Z_{i,1}} \{ g(X_i) - g(X_{m(i)}) \} + o_{P_n}(1).
\]

Since \( \frac{1}{N_1} \sum_{Z_{i,1}} \{ g(X_i) - g(X_{m(i)}) \} \geq \eta \) with probability approaching one for some \( \eta > 0 \) by Lemma 18, \( \sqrt{N_1}\tau \) is asymptotically biased. By an argument parallel to the one used in the proof of Proposition 4, this implies \( \limsup \mathbb{P}_{h_n,g}(p < \alpha) > \alpha \) and hence the paired randomization test is not robust to local nonlinearity.

For the residual difference-of-means test statistic (P2), Proposition 2 does not establish the failure of asymptotic validity when \( ||X_1 - \bar{X}_0|| \neq O_P(1) \) so we cannot assume \( ||X_1 - \bar{X}_0|| = O_P(1) \). Therefore, we do
Lemma 19. If \( \text{Var}(X) \) is finite and nonsingular, \( P(Z = 0 | X) \geq \delta > 0 \), and \( P(Z = 1) > 0 \), then \( ||X_1 - X_{m(1)}|| = o_P(1) \). If \( ||X|| \) has more than two moments and \( ||X_1 - X_{m(1)}|| \) is interpreted to be zero when there are no untreated units, then also \( \mathbb{E}[||X_1 - X_{m(1)}||] \to 0 \).

Proof. First, we prove the result under Euclidean nearest-neighbor matching by mimicking the proof of Lemma 6.1 in [22]. Let \( \varepsilon > 0 \) be arbitrary. Then we may write:

\[
P(||X_1 - X_{m(1)}|| > \varepsilon | Z_1 = 1) \leq P(Z_1 ||X_1 - X_{m(1)}|| > \varepsilon | Z_1 = 1) = \int_{\mathbb{R}^d} P(||x - X_{m(1)}|| > \varepsilon | Z_1 = 1, X_1 = x) dP(x | Z = 1)
\]

\[
\leq \int_{\mathbb{R}^d} \prod_{i=2}^n P(Z_i = 1 \text{ or } X_i \notin \bar{B}_\varepsilon(x)) dP(x | Z = 1)
\]

\[
= \int_{\mathbb{R}^d} P(Z_i = 1 \text{ or } X_i \notin \bar{B}_\varepsilon(x))^{n-1} dP(x | Z = 1).
\]

For each fixed \( x \) in the support of \( X | Z = 1 \), the probability \( P(Z_i = 1 \text{ or } X_i \notin \bar{B}_\varepsilon(x)) \) is strictly less than one by overlap and the definition of support. Therefore, \( P(Z_i = 1 \text{ or } X_i \notin \bar{B}_\varepsilon(x))^{n-1} \to 0 \) for \( P(X \in \cdot | Z = 1) \)-almost every \( x \). Hence, by the dominated convergence theorem, \( P(Z_1 ||X_1 - X_{m(1)}|| > \varepsilon) \to 0 \). Since \( \varepsilon \) is arbitrary, this means \( Z_1 ||X_1 - X_{m(1)}|| = o_P(1) \) under Euclidean matching. To extend this result to Mahalanobis matching, we reason as follows:

\[
||X_1 - X_{m(1)}||^2 \leq \lambda_{\min}(\hat{\Sigma}^{-1})^{-1} Z_1 d_M(X_1, X_{m(1)})^2 \\
= \lambda_{\min}(\hat{\Sigma}^{-1})^{-1} \min_{j:Z_j=0} Z_1 d_M(X_1, X_j)^2 \\
\leq \lambda_{\max}(\hat{\Sigma}^{-1}) \min_{j:Z_j=0} Z_1 ||X_1 - X_j||^2 \\
= O_P(1) \times o_P(1) \\
= o_P(1).
\]

Hence, \( Z_1 ||X_1 - X_{m(1)}|| \xrightarrow{L^1} 0 \) under Mahalanobis matching as well.

Now, we improve this to \( L^1 \) convergence under the additional assumption that \( ||X|| \) has more than two moments. We use the exchangeability of the observations to show that \( \mathbb{E}[Z_1 ||X_1 - X_{m(1)}||^2] \) is uniformly
bounded:
\[
\mathbb{E}[||X_1 - X_{m(1)}||^2] \leq \mathbb{E} \left[ \sum_{j \neq 1} ||X_1 - X_j||^2 I\{m(1) = j\} \right] \\
\leq \mathbb{E} \left[ \sum_{j \neq 1} (2||X_1||^2 + 2||X_j||^2) I\{m(1) = j\} \right] \\
\leq 2\mathbb{E}[||X_1||^2] + \sum_{j \neq 1} 2\mathbb{E}[||X_j||^2 I\{m(1) = j\}] \\
= 2\mathbb{E}[||X_1||^2] + 2(n-1)\mathbb{E}[||X_2||^2 I\{m(1) = 2\}] \\
= 2\mathbb{E}[||X_1||^2] + 2(n-1)\frac{1}{n-1} \sum_{j \neq 2} \mathbb{E}[||X_2||^2 I\{m(j) = 2\}] \\
= 2\mathbb{E}[||X_1||^2] + 2\mathbb{E}[||X_2||^2] \sum_{j \neq 2} I\{m(j) = 2\} \\
\leq 2\mathbb{E}[||X_1||^2] + 2\mathbb{E}[||X_2||^2 K_{2,n}] \\
\]
where in the last line, \(K_{2,n}\) is the number of times observation 2 is used as a match. Since \(||X_2||^2\) has more than one moment \(K_{2,n}\) has infinitely many uniformly bounded moments (Lemma 4), we conclude \(\mathbb{E}[||X_1 - X_{m(1)}||^2]\) is uniformly bounded in \(n\) by Hölder’s inequality. Hence, the sequence \(\{||X_1 - X_{m(1)}||\}\) is uniformly integrable and \(\mathbb{E}[||X_1 - X_{m(1)}||] \to 0\) by Vitali’s convergence theorem.

\[\Box\]

**Lemma 20.** Assume that \(P\) satisfies Assumption 1, except possibly that \(P(Z = 1) < 0.5\). Let \(f : \mathbb{R}^d \to \mathbb{R}\) be any function with more than one moment. Then under one-nearest neighbor Mahalanobis matching with replacement, we have:
\[
\frac{1}{N_1} \sum_{Z_i = 1} f(X_i) - \frac{1}{N_1} \sum_{Z_i = 1} f(X_{m(i)}) \overset{P}{\to} 0.
\]

**Proof.** Throughout this proof, \(K_{i,n}\) denotes the number of times the \(i\)-th observation is used as an untreated match (so \(K_{i,n} = 0\) if \(Z_i = 1\)).

Let \(\eta > 0\) be arbitrary. As in the proof of Lemma 11, let \(g\) be an \(L\)-Lipschitz function such that \(\mathbb{E}[|g(X) - f(X)|^{1+\rho}] < \eta\). Then we may write:
\[
\left| \frac{1}{N_1} \sum_{Z_i = 1} \{f(X_i) - f(X_{m(i)})\} \right| \leq \left| \frac{1}{N_1} \sum_{Z_i = 1} |g(X_i) - g(X_{m(i)})| \right| + \left| \frac{1}{N_1} \sum_{i=1}^n (1 + K_{i,n}) |f(X_i) - g(X_i)| \right|
\leq L \frac{1}{N_1} \sum_{Z_i = 1} ||X_i - X_{m(i)}|| \\
+ \left( \frac{1}{N_1} \sum_{i=1}^n (1 + K_{i,n})^{(1+\rho)/\rho} \right)^{\rho/(1+\rho)} \left( \frac{1}{N_1} \sum_{i=1}^n |f(X_i) - g(X_i)|^{1+\rho} \right)^{1/(1+\rho)}.
\]

By exchangeability, \(\mathbb{E}[\frac{1}{n} \sum_{i=1}^n Z_i||X_i - X_{m(i)}||] = \mathbb{E}[Z_1||X_1 - X_{m(1)}||]\), which vanishes by Lemma 19. Hence, \(L \frac{1}{N_1} \sum_{Z_i = 1} ||X_i - X_{m(i)}|| = O_P(1)\) by Markov’s inequality. Meanwhile, \(\frac{1}{N_1} \sum_{i=1}^n (1 + K_{i,n})^{(1+\rho)/\rho} = O_P(1)\) by Lemma 4 and Markov. Crucially, this \(O_P(1)\) does not have any dependence on \(\eta\). Moreover, \(\frac{1}{N_1} \sum_{i=1}^n |f(X_i) - g(X_i)|^{1+\rho} \leq \eta/p + o_P(1)\) by the law of large numbers. Thus, we conclude that:
\[
\frac{1}{N_1} \sum_{Z_i = 1} \{f(X_i) - f(X_{m(i)})\} \leq o_P(1) + \eta^{1/(1+\rho)}O_P(1).
\]

Since \(\eta\) is arbitrary and the \(O_P(1)\) term does not depend on \(\eta\), this proves the result.

\[\Box\]
A.8.2 Proof of the lemma

Proof. Let \( \mu_z(x) = \mathbb{E}[Y | X = x, Z = z] \) and set \( \varepsilon_i = Y_i - \mu_{Z_i}(X_i) \). We start with the analysis of covariance statistic (R1). Parameterize \( Z_i \in \{-\frac{1}{2}, \frac{1}{2}\} \) rather than \( \{0, 1\} \). Observe that the weighted design matrix can be written as follows:

\[
\frac{1}{N_1} \sum_{i \in \mathcal{M}} W_i(Z_i, 1, X_i)(Z_i, 1, X_i)^\top = \frac{1}{N_1} \sum_{Z_i > 0} (\frac{1}{2}, 1, X_i)(\frac{1}{2}, 1, X_i)^\top + \frac{1}{N_1} \sum_{Z_i < 0} K_{i,n}(-\frac{1}{2}, 1, X_i)(-\frac{1}{2}, 1, X_i)^\top
\]

\[
\xrightarrow{P} \begin{bmatrix}
\frac{1}{2} \\ \frac{1}{2} \\ 1
\end{bmatrix}
\begin{bmatrix}
\frac{1}{2} \mathbb{E}[X | Z = 1]^\top \\
\mathbb{E}[X | Z = 1]^\top \\
\mathbb{E}[X | Z = 1]^\top
\end{bmatrix}
\]

\[
+ \begin{bmatrix}
\frac{1}{4} \\ -\frac{1}{2} \\ 1
\end{bmatrix}
\begin{bmatrix}
\frac{1}{2} \mathbb{E}[X | Z = 1]^\top \\
\mathbb{E}[X | Z = 1]^\top \\
\mathbb{E}[X | Z = 1]^\top
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\frac{1}{2} \\ 0 \\ 2
\end{bmatrix}
\begin{bmatrix}
\mathbb{E}[X | Z = 1]^\top \\
2 \mathbb{E}[X | Z = 1]^\top \\
2 \mathbb{E}[X | Z = 1]^\top
\end{bmatrix}
\]

Here, the convergence step follows from Lemma 20. Next, observe that we may write the quantity \( \frac{1}{N_1} \sum_{i \in \mathcal{M}} W_i Y_i(Z_i, 1, X_i) \) as follows:

\[
\frac{1}{N_1} \sum_{i \in \mathcal{M}} W_i Y_i(Z_i, 1, X_i) = \frac{1}{N_1} \sum_{Z_i > 0} \mu_1(X_i)(\frac{1}{2}, 1, X_i) + \frac{1}{N_1} \sum_{Z_i < 0} \varepsilon_i(\frac{1}{2}, 1, X_i)
\]

\[
\xrightarrow{P} \mathbb{E}[\mu_1(X)(\frac{1}{2}, 1, X) | Z = 1] \text{ by LLN, conditionally} + \mathbb{E}[\mu_1(X) + \mu_0(X) | Z = 1] \text{ by Lemma 5}
\]

\[
\xrightarrow{P} \begin{bmatrix}
\mathbb{E}[\mu_1(X) - \mu_0(X) | Z = 1] \\
\mathbb{E}[\mu_1(X) + \mu_0(X) | Z = 1]
\end{bmatrix}
\]

Finally, we may combine these results with standard linear regression formulas to establish the consistency of \( \hat{\tau} \):

\[
\hat{\tau} = e_1^\top \left( \frac{1}{N_1} \sum_{i \in \mathcal{M}} W_i(Z_i, 1, X_i)(Z_i, 1, X_i) \right)^{-1} \frac{1}{N_1} \sum_{i \in \mathcal{M}} W_i Y_i(Z_i, 1, X_i)
\]

\[
\xrightarrow{P} e_1^\top \begin{bmatrix}
\frac{1}{2} \\ 0 \\ 2
\end{bmatrix}
\begin{bmatrix}
\mathbb{E}[\mu_1(X) - \mu_0(X) | Z = 1] \\
\mathbb{E}[\mu_1(X) + \mu_0(X) | Z = 1]
\end{bmatrix}
\]

\[
= \mathbb{E}[\mu_1(X) - \mu_0(X) | Z = 1].
\]

Under unconfoundedness, this limit is equal to \( \mathbb{E}[Y(1) - Y(0) | Z = 1] \). Thus, we have proved consistency of the analysis of covariance statistic.

Next, we prove the result for the residual-difference-of-means statistic (R2). We return to our original
The proof is quite long, so we give a high-level plan as follows:

\[
\begin{align*}
\|\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \|_2 &= \left\| \left( \frac{1}{N_1} \sum_{i \in M} W_i(1, X_i)(1, X_i)^T \right)^{-1} \frac{1}{N_1} \sum_{i \in M} W_i Y_i(1, X_i) \right\|_2 \\
&\leq \lambda_{\max} \left( \left( \frac{1}{N_1} \sum_{i \in M} (1, X_i)(1, X_i)^T \right)^{-1} \right)^{1/2} \left\| \frac{1}{N_1} \sum_{i \in M} W_i Y_i(1, X_i) \right\|_2 \\
&\leq \left[ \frac{1}{N_1} \sum_{i \in M} \lambda_{\min} \left( (1, X_i)(1, X_i)^T \right)^{-1} \right]^{1/2} O_P(1) \\
&= \left[ \lambda_{\min} \left( \mathbb{E}[1] \right) \right]^{1/2} O_P(1) \\
&= O_P(1).
\end{align*}
\]

The same arguments apply in the case of the unweighted version of (R2). Thus, for either version of this test statistic, we have:

\[
\hat{\tau} = \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ Y_i - \hat{\beta}^T X_i - Y_m(i) + \hat{\beta}^T X_m(i) \right\} \\
= \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ \mu_1(X_i) - \mu_0(X_m(i)) \right\} + \frac{1}{N_1} \sum_{i \in M} \left\{ \varepsilon_i - \varepsilon_m(i) \right\} + \frac{\hat{\beta}^T}{O_P(1)} \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ X_m(i) - X_i \right\} \\
= \frac{1}{N_1} \sum_{i = 1}^{N_1} \mu_1(X_i) - \frac{1}{N_1} \sum_{i = 1}^{n} (1 - Z_i) K_{i,n} \mu_0(X_i) + o_P(1) \\
\xrightarrow{P} \mathbb{E}[\mu_1(X)|Z = 1] \text{ by LLN } \\
\xrightarrow{P} \mathbb{E}[\mu_0(X)|Z = 1] \text{ by Lemma 20.}
\]

The same arguments apply in the case of the unweighted version of (R2). Thus, for either version of this test statistic, we have:

\[
\hat{\theta} = \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ Y_i - \hat{\beta}^T X_i - Y_m(i) + \hat{\beta}^T X_m(i) \right\} \\
= \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ \mu_1(X_i) - \mu_0(X_m(i)) \right\} + \frac{1}{N_1} \sum_{i \in M} \left\{ \varepsilon_i - \varepsilon_m(i) \right\} + \frac{\hat{\beta}^T}{O_P(1)} \frac{1}{N_1} \sum_{i = 1}^{N_1} \left\{ X_m(i) - X_i \right\} \\
= \frac{1}{N_1} \sum_{i = 1}^{N_1} \mu_1(X_i) - \frac{1}{N_1} \sum_{i = 1}^{n} (1 - Z_i) K_{i,n} \mu_0(X_i) + o_P(1) \\
\xrightarrow{P} \mathbb{E}[\mu_1(X)|Z = 1] \text{ by LLN } \\
\xrightarrow{P} \mathbb{E}[\mu_0(X)|Z = 1] \text{ by Lemma 20.}
\]

\section*{A.9 Proof of Theorem 1}

The proof is quite long, so we give a high-level plan as follows:

1. In Section A.9.1, we establish the locally uniform asymptotic normality of the studentized test statistic \( \hat{\tau}/\sigma \), where \( \sigma \) is (approximately) the conditional variance of \( \hat{\tau} \) given the matching. The proof combines the martingale argument in [4] with an application of Le Cam’s third lemma [64, Theorem 6.7].

2. In Section A.9.2, we develop build up the technical tools needed to study the stratified randomization distribution. The culmination of this analysis is Proposition 6, which is a law of large numbers for the stratified randomization distribution.

3. In Section A.9.3, we use this randomization law of large numbers to derive the asymptotic distribution of the permuted test statistic \( \hat{\tau}_* \). In particular, we show that the randomization critical value \( \hat{q}_{1-\alpha} := \inf\{ t \in \mathbb{R} : \mathbb{P}[|\hat{\tau}| \leq t | D] \geq 1 - \alpha \} \) satisfies \( \hat{q}_{1-\alpha} = z_{1-\alpha}/2 + o_P(1/\sqrt{n}) \) for \( \sigma \) the same quantity introduced in Section A.9.1.

4. Finally, in Section A.9.4, we combine the locally uniform asymptotic normality of \( \hat{\tau}/\sigma \) with the asymptotics of the randomization critical value to prove Theorem 1.
A.9.1 Asymptotic normality of the test statistic

Lemma 21. Assume the conditions of Theorem 1. Let \( \varepsilon_i = Y_i - \gamma - \beta^\top X_i \). Then for any square-integrable function \( f(x, y, z) \), we have:

\[
\frac{1}{n} \sum_{i=1}^{n} \left( Z_i - (1 - Z_i) K_{i,n} \right) \varepsilon_i f(X_i, Y_i, Z_i) \xrightarrow{P} E[\phi(X, Y, Z) f(X, Y, Z)]
\]

where \( \phi \) is the efficient influence function for the ATT derived in [23].

Proof. Under the conditions of Theorem 1, the efficient influence function \( \phi \) has the following expression:

\[
\phi(x, y, z) = \frac{1}{p} \left( z - (1 - z) \frac{e(x)}{1 - e(x)} \right) (y - \gamma - \beta^\top x)
\]

where \( p = P(Z = 1) \). Now write the quantity of interest as the sum of three terms:

\[
\frac{1}{n} \sum_{i=1}^{n} \left( Z_i - (1 - Z_i) K_{i,n} \right) \varepsilon_i f(X_i, Y_i, Z_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i \varepsilon_i}{p} f(X_i, Y_i, Z_i) - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) K_{i,n}}{p} E[\varepsilon f(X, Y, 0) | X_i, Z_i] - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) K_{i,n}}{p} (\varepsilon_i f(X_i, Y_i, 0) - E[\varepsilon f(X, Y, 0) | X_i, Z_i])
\]

The first term on the right-hand side of the preceding display converges to \( E[Z \varepsilon f(X, Y, Z)/p] \) by the law of large numbers. By Lemma 20, the second term has the following limit:

\[
\frac{1}{np} \sum_{i=1}^{n} (1 - Z_i) K_{i,n} E[\varepsilon f(X, Y, 0) | X_i, Z_i] \xrightarrow{P} E[E[\varepsilon f(X, Y, 0) | X, Z = 0] | Z = 1]
\]

\[
= \frac{1}{p} E[E[\varepsilon f(X, Y, 0) | X, Z = 0]]
\]

\[
= \frac{1}{p} E \left[ (1 - Z) \frac{e(x)}{1 - e(x)} \varepsilon f(X, Y, 0) \right]
\]

\[
= E \left[ (1 - Z) \frac{e(x)}{1 - e(x)} \varepsilon f(X, Y, Z) \right].
\]

Finally, we claim the third term tends to zero in probability. Observe \( E[f(X_i, Y_i, 0)^2] < \infty \) (by overlap), and we have assumed \( e \) has more than two moments. Therefore, the product \( \varepsilon_i f(X_i, Y_i, 0) \) has more than one moment. Let \( \rho \in (0, 1) \) be a number small enough so that \( E[|\varepsilon f(X, Y, 0)|^{1+\rho}] < \infty \), and hence \( E[|\varepsilon f(X, Y, 0) - E[\varepsilon f(X, Y, 0) | X, Z]|^{1+\rho}] < \infty \). Now apply the von Bahr-Esseen inequality [65] conditional on \( \{(X_i, Z_i, U_i)\}_{i \leq n} \) (where \( U_i \) is the randomness used to break ties in the matching process) to conclude:

\[
E \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) K_{i,n}}{p} (\varepsilon_i f(X_i, Y_i, 0) - E[\varepsilon f(X, Y, 0) | X_i, Z_i]) \right|^{1+\rho/2} \right] \leq 2C \frac{n^{\rho/2}}{n \rho^{2}} \sum_{i=1}^{n} \frac{(1 - Z_i) K_{i,n}^{1+\rho/2}}{p} E[|\varepsilon_i f(X_i, Y_i, 0) - E[\varepsilon f(X, Y, 0) | X_i, Z_i]|^{1+\rho/2} | \{(X_i, Z_i, U_i)\}_{i \leq n}]
\]

\[
\leq \frac{C}{n \rho^{2/\rho}} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) K_{i,n}}{p} \right)^{2(\rho+2)/\rho} \frac{1}{n} \sum_{i=1}^{n} \frac{E[|\varepsilon_i f(X_i, Y_i, 0)|^{1+\rho}] \{(X_i, Z_i, U_i)\}}{^{2(\rho+2)}/(2+2\rho)} = O_P(1) \text{ by Lemma 4 + Markov}
\]

\[
= o_P(1)
\]

36
Hence, applying Markov’s inequality conditionally implies the third term tends to zero in probability. Combining these three terms gives:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{(Z_i - (1 - Z_i)K_{i,n})\varepsilon_i}{p} f(X_i, Y_i, Z_i) \overset{p}{\to} \mathbb{E} \left[ \frac{Z\varepsilon}{p} f(X, Y, Z) \right] - \mathbb{E} \left[ \frac{(1 - Z)}{1 - e(X)} \varepsilon f(X, Y, Z) \right] = \mathbb{E}[\phi(X, Y, Z)f(X, Y, Z)].
\]

Lemma 22. Assume the conditions of Theorem 1. Let \( \varepsilon_i = Y_i - \gamma - \beta^\top X_i \). Then the analysis of covariance statistic \((R1)\) and either the weighted or unweighted residual difference-of-means statistic \((R2)\) satisfy:

\[
\hat{\tau} = \frac{1}{np} \sum_{i=1}^{n} (Z_i - (1 - Z_i)K_{i,n})\varepsilon_i + o_P(1/\sqrt{n})
\]

Proof. First, we prove the result for the analysis of covariance statistic \((R1)\). As in the proof of Lemmas 15 and 14, we consider a parameterization where \( Z_i \) takes values in \( \{\pm \frac{1}{2}\} \) rather than \( \{0, 1\} \). Then we may write:

\[
\hat{\tau} = \varepsilon_1^\top \hat{G}^{-1} \frac{1}{N_1} \sum_{Z_i > 0} \left\{ \varepsilon_i \begin{pmatrix} Z_i \\ 1 \\ X_i \end{pmatrix} + \varepsilon_{m(i)} \begin{pmatrix} Z_{m(i)} \\ 1 \\ X_{m(i)} \end{pmatrix} \right\}
\]

where \( \hat{G} \) is the matched design matrix. By Lemma 20 and the law of large numbers, \( \hat{G} \) has the following limit:

\[
\hat{G} = \frac{1}{N_1} \sum_{Z_i > 0} \left( \frac{1}{2}, 1, X_i \right) \left( \frac{1}{2}, 1, X_i \right)^\top + \frac{1}{N_1} \sum_{Z_i > 0} \left( -\frac{1}{2}, 1, X_{m(i)} \right) \left( -\frac{1}{2}, 1, X_{m(i)} \right)^\top
\]

\[
\overset{p}{\to} \begin{bmatrix}
\frac{1}{2} & 0 & 2E[X^\top | Z = 1] \\
0 & 2E[X | Z = 1] & 2E[X^\top X | Z = 1]
\end{bmatrix}
\]

Hence, \( \hat{G}^{-1}\varepsilon_1 \overset{p}{\to} (2, 0) \). In addition, we can show that \( \frac{1}{N_1} \sum_{Z_i > 0} \{\varepsilon_i(Z_i, 1, X_i) + \varepsilon_{m(i)}(Z_{m(i)}, 1, X_{m(i)})\} = O_P(1/\sqrt{n}) \) by applying Chebyshev’s inequality conditionally on \( \{(X_i, Z_i, U_i)\}_{i \leq n} \) (where \( U_i \) is the independent randomness used in breaking ties):

\[
\text{Var} \left( \frac{1}{N_1} \sum_{Z_i > 0} \varepsilon_i(Z_i, 1, X_i) + \varepsilon_{m(i)}(Z_{m(i)}, 1, X_{m(i)}) \middle| \{(X_i, Z_i, U_i)\}_{i \leq n} \right) = \text{Var} \left( \frac{1}{N_1} \sum_{Z_i > 0} (I[Z_i > 0] + K_{i,n}I[Z_i < 0]) \varepsilon_i \middle| \{(X_i, Z_i, U_i)\}_{i \leq n} \right)
\]

\[
= \frac{1}{N_1^2} \sum_{Z_i > 0} (I[Z_i > 0] + K_{i,n}^2I[Z_i < 0])\sigma^2(X_i)
\]

\[
\leq \frac{1}{N_1} \left( \frac{1}{N_1} \sum_{i=1}^{n} (1 + K_{i,n}^2) \right)^{1/2} \left( \frac{1}{N_1} \sum_{i=1}^{n} \sigma^4(X_i) \right)^{1/2}
\]

\[
= O_P(1/n)
\]

where the last step follows by Lemma 4 and Markov’s inequality. Combining \( \hat{G}^{-1}\varepsilon_1 = (2, 0) + o_P(1) \) with \( \frac{1}{N_1} \sum_{Z_i > 0} \{\varepsilon_i(Z_i, 1, X_i) + \varepsilon_{m(i)}(Z_{m(i)}, 1, X_{m(i)})\} = O_P(1/\sqrt{n}) \) gives \( \hat{\tau} = \frac{1}{N_1} \sum_{Z_i > 0} (Z_i - (1 - Z_i)K_{i,n})\varepsilon_i + o_P(1/\sqrt{n}) \). Now returning to the original parameterization where \( Z_i \in \{0, 1\} \) allows us to write this as \( \hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{n} (Z_i - (1 - Z_i)K_{i,n})\varepsilon_i + o_P(1/\sqrt{n}) \). Since \( N_1/np = 1 + o_P(1) \), the same conclusion holds with \( N_1 \) replaced by \( np \).
Next, we consider the two versions of the Tukey’s residual difference-of-means statistic \((R2)\). We begin by showing that \(\hat{\beta} - \beta = O_P(1/\sqrt{n})\). Consider first the unweighted case:

$$\left\| \left( \frac{1}{N_1} \sum_{Z_i=1} \{ (1, X_i)(1, X_i)^T \} \right)^{-1} \frac{1}{N_1} \sum_{Z_i=1} \{ (1, X_i)\xi_i \} \right\| \leq \lambda_{\max} \left( \left( \frac{1}{N_1} \sum_{Z_i=1} \{ (1, X_i)(1, X_i)^T \} \right)^{-1} \right) \left\| \frac{1}{N_1} \sum_{Z_i=1} \{ (1, X_i)\xi_i \} \right\| \leq \lambda_{\min}^{-1} \left( \frac{1}{N_1} \sum_{Z_i=1} X_iX_i^T \right) \left\| \frac{1}{N_1} \sum_{Z_i=1} \{ (1, X_i)\xi_i \} \right\| \xrightarrow{P} O_P(1/\sqrt{n})$$

by Chebyshev condition on \((X_i, Z_i, U_i)\) and Lemma 22.

The weighted case follows by an identical calculation. Finally, write:

$$\hat{\beta} = \frac{1}{N_1} \sum_{Z_i=1} \{ (Y_i - \hat{\gamma} - \hat{\beta}^T X_i) - (Y_{m(i)} - \hat{\gamma} - \hat{\beta}^T X_{m(i)}) \}$$

$$= \frac{1}{N_1} \sum_{Z_i=1} \{ \xi_i - \xi_{m(i)} \} + (\beta - \hat{\beta})^T \left( \frac{1}{N_1} \sum_{Z_i=1} X_i - X_{m(i)} \right) = O_P(1/\sqrt{n})$$

by Lemma 20.

As before, we may replace \(N_1\) in the denominator by \(np\) since \(np/N_1 \xrightarrow{P} 1\) and \(\frac{1}{N_1} \sum_{Z_i=1} \{ \xi_i - \xi_{m(i)} \} = O_P(1/\sqrt{n})\).

**Corollary 1.** Assume the conditions of Theorem 1. Let \(\sigma^2 = \frac{1}{(np)^2} \sum_{i=1}^n (Z_i + (1 - Z_i)K_{i,n})\sigma^2(X_i)\). Then \(1/(\sqrt{n}\sigma) = O_P(1)\) and the following expansion holds:

$$\frac{\hat{\beta}}{\sigma} = \frac{1}{\sqrt{n}\sigma} \sqrt{n} \sum_{i=1}^n (Z_i + (1 - Z_i)K_{i,n})\xi_i + o_P(1).$$

**Proof.** The definition of \(\sigma^2\) gives \(na^2 \geq \frac{1}{n} \sum_{i=1}^n Z_i\sigma^2(X_i) = E[\sigma^2(X)Z] - o_P(1)\). Under our assumptions, \(E[\sigma^2(X)Z] > 0\), so \(1/(\sqrt{n}\sigma) \leq 1/\sqrt{E[\sigma^2(X)|Z=1]} = o_P(1) = O_P(1)\). Now, the conclusion follows by algebra and Lemma 22.

**Proposition 5.** (Locally uniform asymptotic normality)

Assume that \(P\) satisfies the conditions of Theorem 1. Let \(\{P_h\} \subset H_0\) be a smooth parametric model with \(P_0 = P\). Then we have:

$$\frac{\hat{\beta}}{\sigma} \xrightarrow{P_n} N(0,1)$$

whenever \(\{h_n\}\) satisfies \(\sqrt{n}h_n \to h \in \mathbb{R}^k\).

**Proof.** First, we set some notation. Let \(\ell(x, y, z)\) be the score function (quadratic-mean derivative) in the model \(\{P_h\}\) at \(h = 0\). It will also be convenient to write \(\ell\) as the sum of two terms:

$$\ell(x, y, z) = \mathbb{E}[\ell(X, Y, Z)]|X = x, Z = z] + \ell(x, y, z) - \mathbb{E}[\ell(X, Y, Z)|X = x, Z = z]$$

$$:= \ell_{X,x}(x, z) + \ell_{Y|x, z}(y|x, z)$$

38
Let \( \phi(x, y, z) \) be the efficient influence function for the ATT derived in [23] (see Lemma 21 for its formula). Since the model \( \{P_n\} \) is a subset of the null hypothesis \( H_0 \), \( \ell \) belongs to the nuisance parameter tangent space and hence \( \mathbb{E}_P[\ell(Y, X, Z)\phi(X, Y, Z)] = 0 \).

We now derive the joint asymptotic distribution of \( \hat{\tau}/\sigma \) and \( L_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell(X_i, Y_i, Z_i) \) under \( P \). We use the Cr\'amer-Wold device. Let \( a, b \) be any real numbers, and consider the \( a\hat{\tau}/\sigma + bL_n \). Using the expansion from Corollary 1 gives:

\[
a\hat{\tau}/\sigma + bL_n = \sum_{i=1}^n \frac{1}{\sqrt{n}} \left\{ \frac{1}{\sqrt{n}\sigma} a(Z_i - (1 - Z_i)K_{i,i,n}) \varepsilon_i + b\ell(X_i, Y_i, Z_i) \right\} + o_P(1). \tag{14}
\]

As in [3], we study the linear term on the right-hand side of (14) using martingale methods. First, we write it as the sum of a martingale difference sequence. For each \( i \) by Corollary 1, it suffices to prove by:

\[
\mathbb{E}_P\left[ \ell(Y, X, Z) \phi(X, Y, Z) \right] = 0.
\]

Clearly, the first condition is that the accumulated conditional variance \( \sum_{i=1}^n \ell(X, Y, Z) \phi(X, Y, Z) \) converges to some limit. We prove this by writing the accumulated variance as the sum of four terms:

\[
\sum_{i=1}^{2n} \mathbb{E} \left[ \xi_i^2 | \mathcal{F}_{i-1,n} \right] = \sum_{i=1}^n \frac{b^2\mathbb{E}[\ell_{X,Z}(X, Z)^2]}{n} + \frac{1}{n\sigma^2} \sum_{i=1}^n (Z_i - (1 - Z_i)K_{i,i,n})^2 \varepsilon_i^2 + \mathbb{E}[\ell_{Y|X,Z}(Y|X, Z)^2]
\]

The first term converges to \( b^2\mathbb{E}[\ell_{X,Z}(X, Z)^2] \), the second is exactly equal to \( a^2 \), and the third term converges to \( b^2\mathbb{E}[\ell_{Y|X,Z}(Y|X, Z)^2] \). I claim the final term converges in probability to zero. Since \( ab/(\sqrt{n}\sigma) = O_P(1) \) by Corollary 1, it suffices to prove \( \sum_{i=1}^n (Z_i - (1 - Z_i)K_i Y_i, Z_i) \mathbb{E}[\phi(X, Y, Z) \ell_{Y|X,Z}(Y_i|X_i, Z_i)/n = o_P(1) \). This follows easily from Lemma 21:

\[
\frac{1}{n} \sum_{i=1}^n (Z_i - (1 - Z_i)K_{i,i,n}) \varepsilon_i \ell_{Y|X,Z}(Y_i|X_i, Z_i) \mathbb{E}[\phi(X, Y, Z) \ell_{Y|X,Z}(Y|X, Z)]
\]

\[
= \mathbb{E}[\phi(X, Y, Z) \ell(Y, X, Z)] = 0.
\]
In summary, we have established that \( \sum_{i=1}^{2n} E[\xi_i^2 | \mathcal{F}_{i-1,n}] \xrightarrow{L} a^2 + b^2(\mathbb{E}[\ell_{X,Z}(X,Z)^2 + \ell_{Y|X,Z}(Y|X,Z)^2]) = a^2 + b^2\mathbb{E}[\ell(X,Y,Z)^2] \).

The other condition that we need to check to apply the martingale central limit theorem is the conditional Lindeberg condition \( \sum_{i=1}^{2n} E[\xi_i^2 | \{ | \xi_i | > \varepsilon \}] | \mathcal{F}_{i-1,n} | = o_p(1) \) for every \( \varepsilon > 0 \). Let \( \varepsilon > 0 \) be given. For each \( i \in [n] \), define \( \xi_{n+i,n}^{(a)} \) and \( \xi_{n+i,n}^{(b)} \) as follows:

\[
\xi_{n+i,n}^{(a)} = \frac{1}{\sqrt{n}\sigma} \frac{a(Z_i - (1 - Z_i)K_{i,n})\varepsilon_i}{\sqrt{n}\rho} \quad \text{and} \quad \xi_{n+i,n}^{(b)} = \frac{b\ell_{Y|X,Z}(Y_i|X_i, Z_i)}{\sqrt{n}}
\]

Then we may write:

\[
\sum_{i=1}^{2n} E[\xi_i^2 \mathbb{1}\{|\xi_i | > \varepsilon \}] \leq \sum_{i=1}^{n} E[|\xi_i |^2 \mathbb{1}|\{ |\xi_i | > \varepsilon \}] | \mathcal{F}_{i-1,n} | + \sum_{i=n+1}^{2n} E[|\xi_i |^2 \mathbb{1}|\{ |\xi_i | > \varepsilon / 2 \}] | \mathcal{F}_{i-1,n} | \leq o_p(1)
\]

Since the terms \( \xi_{1,n}, \ldots, \xi_{n,n} \) are independent and identically distributed, (15) tends to zero by the verification of the Lindeberg condition for the standard i.i.d. central limit theorem (i.e. by the dominated convergence theorem). For the other terms, let \( \rho > 0 \) be small enough so that \( \mathbb{E}[|\varepsilon |^{2+\rho}] < \infty \) and argue as follows:

\[
(16) \leq \sum_{i=1}^{n} E[|\xi_i |^{2+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(2+\rho/2)}
\]

\[
= \frac{1}{n^{\rho/2}} \left( \frac{1}{n^{\rho/2}} \right)^{2+\rho/2} \frac{1}{n} \sum_{i=1}^{n} (Z_i + (1 - Z_i)K_{i,n}^{2+\rho/2}) E[\xi_i^{2+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(2+\rho/2)}
\]

\[
\leq \max_{i \leq n} (1 + K_{i,n}^{2+\rho/2}) \frac{1}{n^{\rho/2}} \sum_{i=1}^{n} E[\xi_i^{2+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(2+\rho/2)}
\]

\[
= o_p(1)
\]

\[
(17) \leq \sum_{i=1}^{2n} 2E[|\xi_i |^{1+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(1+\rho/2)}
\]

\[
= \frac{2}{n^{\rho/2}} \left( \frac{a}{\sqrt{n}\sigma} \right)^{1+\rho/2} \frac{1}{n} \sum_{i=1}^{n} (Z_i + (1 - Z_i)K_{i,n}^{1+\rho/2}) E[|\xi_i |^{1+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(1+\rho/2)}
\]

\[
\leq \max_{i \leq n} 2(1 + K_{i,n}^{1+\rho/2}) \frac{1}{n^{\rho/2}} \sum_{i=1}^{n} E[|\xi_i |^{1+\rho/2} | \mathcal{F}_{i-1,n} | (\varepsilon / 2)^{-(1+\rho/2)}
\]

\[
= o_p(1)
\]
Proof.
By Lemma 19 and the continuity of $\ell_Y|X,Z(Y_i|X_i, Z_i)\|\ell_Y|X,Z(Y_i|X_i, Z_i) > \varepsilon \sqrt{n/b} \mid \mathcal{F}_{i-1,n}$

$= O_P(1)\left(\sum_{i=1}^{n} \mathbb{E}\left[(Z_i - (1 - Z_i))K_{i,n} \varepsilon_i \ell_Y|X,Z(Y_i|X_i, Z_i)\|\ell_Y|X,Z(Y_i|X_i, Z_i) > \varepsilon \sqrt{n/b}\right]\right)$

$= O_P(1)\left(\sum_{i=1}^{n} \mathbb{E}\left[(Z_i - (1 - Z_i))K_{i,n} \varepsilon_i \ell_Y|X,Z(Y_i|X_i, Z_i)\|\ell_Y|X,Z(Y_i|X_i, Z_i) > \varepsilon \sqrt{n/b}\right]\right) = O(1)\text{ by Hölder + Lemma 4}$

Thus, all the terms (15) – (19) are vanishing, and the conditional Lindeberg condition holds.

Thus, the martingale central limit theorem implies that the linear term in (14) converges in distribution to $N(0, a^2 + b^2\mathbb{E}[\ell(X,Y,Z)^2])$. By Slutsky’s theorem, $a\hat{r} + bL_n$ converges weakly to the same limit distribution. Since $a, b$ are arbitrary, this gives:

$\frac{\hat{r}/\sigma}{L_n} \leadsto N\left(\begin{array}{c}0 \\ 0 \end{array}\right)\cdot \begin{bmatrix}1 \\ \mathbb{E}[\ell(X,Y,Z)^2]\end{bmatrix}$.}

Thus, by Le Cam’s third lemma [64, Theorem 6.7] and the likelihood expansion of a smooth parametric model [64, Theorem 7.2], $\hat{r}/\sigma \leadsto N(0,1)$ also under $P_{h_n}$.

A.9.2 Technical results on stratified randomization

Lemma 23. For any function $f : \mathbb{R}^d \to \mathbb{R}$ and any $q > 1$, we have $\mathbb{E}[f(X_m(1))\mathbb{I}\{N_0 > 0\}] \leq C(q)\mathbb{E}[f(X_1)^q]^{1/q}$ for some constant $C(q) < \infty$ depending only on $q$ and the underlying distribution $P$.

Proof. The proof is based on Lemma 6.3 in [22]. By the exchangeability of the observations, we may write:

$\mathbb{E}[f(X_m(1))\mathbb{I}\{N_0 > 0\}] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[f(X_m(j))\mathbb{I}\{N_0 > 0\}]$

$\leq \frac{1}{n}\sum_{j=1}^{n} \sum_{i=1}^{n} \mathbb{E}[f(X_i)\mathbb{I}\{m(j) = i\}]$

$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f(X_i)\sum_{j=1}^{n} \mathbb{I}\{m(j) = i\}\right]$

$= \mathbb{E}[f(X_i)K_{i,n}]$.}

Now the conclusion follows from Hölder’s inequality and Lemma 4.

Lemma 24. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with $\mathbb{E}[|f(X)|^r] < \infty$. Then for any $q < r$, we have $\mathbb{E}[|f(X_1) - f(X_m(1))|^{q}\mathbb{I}\{N_0 > 0\}] \to 0$.

Proof. By Lemma 19 and the continuity of $f$, $|f(X_1) - f(X_m(1))|^{q}\mathbb{I}\{N_0 > 0\} \xrightarrow{P} 0$. Moreover, Lemma 23 implies a uniform bound on a higher moment. Hence, the conclusion follows from Vitali’s convergence theorem.

Lemma 25. For any untreated observation $i$, let $S(i) = \{i\} \cup \{j : m(j) = i\}$ be the “stratum” associated with observation $i$. Let $f$ be any continuous function with $\mathbb{E}[|f(X)|^r] < \infty$. Then for any $q < r$, we have:

$\mathbb{E}\left[\max_{j \in S(i)} |f(X_j) - f(X_i)|^{q}\mathbb{I}\{K_{i,n} > 0\}\right] \to 0$
Let \( f \) be a continuous function such that \( f \) has more than one moment. Then we have:

\[
\mathbb{E} \left[ \max_{j \in \mathcal{S}(1)} |f(X_j) - f(X_1)|^q \mathbb{I}\{K_{1,n} > 0\} \right] \leq \mathbb{E} \left[ \sum_{i=1}^n |f(X_i) - f(X_1)|^q \mathbb{I}\{m(j) = 1\} \right]
\]

\[
= (n-1) \mathbb{E}[|f(X_2) - f(X_1)|^q \mathbb{I}\{m(2) = 1\}]
\]

\[
= (n-1) \mathbb{E}[|f(X_2) - f(X_{m(2)})|^q \mathbb{I}\{m(2) = 1\}]
\]

\[
= \sum_{i \neq 2} \mathbb{E}[|f(X_2) - f(X_{m(2)})|^q \mathbb{I}\{m(2) = i\}]
\]

\[
= \mathbb{E}\left[|f(X_2) - f(X_{m(2)})\sum_{i \neq 2} \mathbb{I}\{m(2) = i\}\right]
\]

\[
\leq \mathbb{E}[|f(X_2) - f(X_{m(2)})|^q \mathbb{I}\{N_0 > 0\}] 
\]

This upper bound tends to zero by Lemma 24.

\[\square\]

**Corollary 2.** Suppose that \( f : \mathbb{R}^d \to \mathbb{R} \) is continuous and satisfies \( \mathbb{E}[|f(X)|^r] < \infty \). Then for any \( q \in (1, r) \), we have:

\[
\mathbb{E} \left[ f(X_i) - \frac{1}{K_{i,n} + 1} \sum_{j \in \mathcal{S}(i)} f(X_j) \mathbb{I}\{K_{i,n} > 0\} \right]^q \to 0.
\]

**Proof.** Bound the average discrepancy by the maximum discrepancy, then apply Lemma 25. \[\square\]

**Corollary 3.** Suppose that \( f : \mathbb{R}^d \to \mathbb{R} \) is continuous and satisfies \( \mathbb{E}[|f(X)|^r] < \infty \) for some \( r > 2 \). Then we have:

\[
\mathbb{E} \left[ f(X_i)^2 - \frac{1}{K_{i,n}(K_{i,n} + 1)} \sum_{j,k \in \mathcal{S}(i)} f(X_j)f(X_k) \mathbb{I}\{K_{i,n} > 0\} \right] \to 0.
\]

**Proof.** For any \( j, k \) we have \( |f(X_i)^2 - f(X_j)f(X_k)|\mathbb{I}\{K_{i,n} > 0\} \leq \max_{j \in \mathcal{S}(i)} |f(X_i) - f(X_j)|^2 \mathbb{I}\{K_{i,n} > 0\} \), so we may apply Lemma 25 with \( q = 2 \) to prove the result. \[\square\]

**Proposition 6.** (Stratified randomization law of large numbers)

Let \( f \) be a continuous function such that \( f \) has more than one moment. Then we have:

\[
\frac{1}{N_1} \sum_{Z_{i,1}^*} f(X_i) \overset{P}{\to} \mathbb{E}[f(X)|Z = 1] \quad (20)
\]

\[
\frac{1}{N_1} \sum_{Z_{i,1}^*} f(X_{m(i)}) \overset{P}{\to} \mathbb{E}[f(X)|Z = 1] \quad (21)
\]

**Proof.** First, we prove the result assuming that \( f \) is uniformly bounded by one. With some abuse of notation, we will let \( \mathcal{M} \) denote the information used in forming matches (formally, \( \mathcal{M} = \sigma\{(X_i, Z_i, U_i)\}_{i \leq n} \)) where \( U_i \) is the randomness used in breaking ties.

We start with (20). The quantity \( \frac{1}{N_1} \sum_{Z_{i,1}^*} f(X_i) \) can be written as follows:

\[
\frac{1}{N_1} \sum_{Z_{i,1}^*} f(X_i) = \frac{1}{N_1} \sum_{K_{i,n} > 0} \sum_{j \in \mathcal{S}(i)} Z_j^* f(X_j).
\]

\[42\]
For each \( j \in S(i) \), we have \( \mathbb{P}(Z_j^*|\mathcal{M}) = 1/(K_{i,n} + 1) \). Therefore, the conditional expectation of the preceding display satisfies:

\[
\mathbb{E} \left[ \frac{1}{N_1} \sum_{Z_j^*=1} f(X_i) \bigg| \mathcal{M} \right] = \frac{1}{N_1} \sum_{K_{i,n} > 0} \sum_{j \in S(i)} \frac{K_{i,n}}{K_{i,n} + 1} f(X_j) = \frac{1}{N_1} \sum_{i=1}^{n} \frac{K_{i,n}}{K_{i,n} + 1} f(X_j)
\]

\[
= \frac{1}{N_1} \sum_{i=1}^{n} (1 - Z_i)K_{i,n}f(X_i) + \frac{1}{N_1} \sum_{i=1}^{n} \left( \frac{1}{K_{i,n} + 1} \sum_{j \in S(i)} f(X_j) - f(X_i) \right) \mathbb{I}\{K_{i,n} > 0\}
\]

\[
= \mathcal{O}(1) \text{ by Corollary 2 + Markov}
\]

\[
\mathbb{E}[f(X)|Z = 1]
\]

where the final line is by Lemma 20. Moreover, the conditional variance is vanishing:

\[
\text{Var} \left( \frac{1}{N_1} \sum_{Z_j^*=1} f(X_i) \bigg| \mathcal{M} \right) = \frac{1}{N_1^2} \sum_{K_{i,n} > 0} \text{Var} \left( \sum_{j \in S(i)} Z_j^* f(X_j) \bigg| \mathcal{M} \right)
\]

\[
\leq \frac{1}{N_1^2} \sum_{K_{i,n} > 0} \mathbb{E} \left[ \left( \sum_{j \in S(i)} Z_j^* f(X_j) \right)^2 \bigg| \mathcal{M} \right]
\]

\[
\leq \frac{1}{N_1^2} \sum_{i=1}^{n} (K_{i,n} + 1)^2
\]

\[
= \mathcal{O}(1)
\]

where the final line is by Lemma 4. Hence, by Chebyshev’s inequality, (20) holds in the bounded case.

Next, we prove (21), still in the bounded case. Start by \( \frac{1}{N_1} \sum_{Z_j^*=1} f(X_{m^*(i)}) \) as follows:

\[
\frac{1}{N_1} \sum_{Z_j^*=1} f(X_{m^*(i)}) = \frac{1}{N_1} \sum_{K_{i,n} > 0} \sum_{j \in S(i)} (1 - Z_j^*)K_{i,n}f(X_j).
\]

A short computation shows that the conditional expectation of this quantity given \( \mathcal{M} \) is exactly the same as the conditional expectation of \( \frac{1}{N_1} \sum_{Z_j^*=1} f(X_i) \). Therefore, the same calculation above gives:

\[
\mathbb{E} \left[ \frac{1}{N_1} \sum_{Z_j^*=1} f(X_{m^*(i)}) \bigg| \mathcal{M} \right] \xrightarrow{P} \mathbb{E}[f(X)|Z = 1].
\]

The conditional variance is also vanishing:

\[
\text{Var} \left( \frac{1}{N_1} \sum_{Z_j^*=1} f(X_{m^*(i)}) \bigg| \mathcal{M} \right) = \frac{1}{N_1^2} \sum_{i=1}^{n} K_{i,n}^2 \text{Var} \left( \sum_{j \in S(i)} (1 - Z_j^*)f(X_j) \bigg| \mathcal{M} \right)
\]

\[
\leq \frac{1}{N_1^2} \sum_{i=1}^{n} K_{i,n}^2 \mathbb{E} \left[ \left( \sum_{j \in S(i)} (1 - Z_j^*)f(X_j) \right)^2 \bigg| \mathcal{M} \right]
\]

\[
\leq \frac{1}{N_1^2} \sum_{i=1}^{n} K_{i,n}^2 (K_{i,n} + 1)^2
\]

\[
= \mathcal{O}(1)
\]
Hence, Chebyshev’s inequality implies (21) holds in the bounded case.
Both (20) and (21) can be extended to unbounded $f$ via a routine truncation argument.

A.9.3 The randomization critical value

**Lemma 26.** Under the assumptions of Theorem 1, the permuted test statistics based on either (P1) or (P2) satisfy:

$$\sqrt{N_1} \hat{\tau} = \frac{1}{\sqrt{N_1}} \sum_{Z^*_i=1} \{ \varepsilon_i - \varepsilon_{m^*(i)} \} + o_P(1)$$

**Proof.** The proofs are identical to the proof of Lemma 22, but we use the stratified randomization law of large numbers of Proposition 6 in place of the ordinary law of large numbers and Lemma 20.

**Lemma 27.** Assume the conditions of Theorem 1. Let $\sigma^2$ be as in Corollary 1. Let $\mathcal{D} = \{(X_i, Y_i, Z_i, U_i)\}_{i \leq n}$ be the original data and the randomness used by the matching scheme when breaking ties. Let $\sigma^2_* = \text{Var}(\frac{1}{N_1} \sum_{Z^*_i=1} \{ \varepsilon_i - \varepsilon_{m^*(i)} \} | D)$. Then $\sigma_* / \sigma = 1 + o_P(1)$.

**Proof.** We start by writing $\frac{1}{\sqrt{N_1}} \sum_{Z^*_i=1} \{ \varepsilon_i - \varepsilon_{m^*(i)} \}$ as the sum of (conditionally) independent terms:

$$\frac{1}{\sqrt{N_1}} \sum_{Z^*_i=1} \{ \varepsilon_i - \varepsilon_{m^*(i)} \} = \sum_{K_i, n > 0} \sum_{j \in S(i)} \frac{1}{\sqrt{N_1}} (Z^*_j - (1 - Z^*_j) K_{i,n}) \varepsilon_j$$

Since $Z^*_j \mid D \sim \text{Bernoulli}(K_{i,n}/(K_{i,n} + 1))$, it is easy to check that $E[\xi_i | D] = 0$ for every $i$. Moreover, for any $j, k \in S(i)$ with $j \neq k$, we have $\text{Var}(Z^*_j - (1 - Z^*_j) K_{i,n} | D) = K_{i,n}$ and $\text{Cov}(Z^*_j - (1 - Z^*_j) K_{i,n}, Z^*_k - (1 - Z^*_k) K_{i,n} | D) = -1$. Thus, we may compute:

$$\text{Var}(\xi_i | D) = \frac{1}{N_1} \sum_{j \in S(i)} \varepsilon_j^2 K_{i,n} - \frac{1}{N_1} \sum_{j, k \in S(i)} \varepsilon_j \varepsilon_k$$

$$\text{Var} \left( \frac{1}{\sqrt{N_1}} \sum_{Z^*_i=1} \{ \varepsilon_i - \varepsilon_{m^*(i)} \} \mid D \right) = \frac{1}{N_1} \sum_{K_i, n > 0} \sum_{j \in S(i)} K_{i,n} \varepsilon_j^2 - \frac{1}{N_1} \sum_{K_i, n > 0} \sum_{j, k \in S(i)} \varepsilon_j \varepsilon_k.$$
The term \( i \) is the “main term” in the conditional variance, and has the following asymptotics:

\[
\begin{align*}
    i &= \frac{1}{N_1} \sum_{i=1}^{n} K_{i,n} \sum_{j \in S(i)} (\varepsilon_j - \sigma^2(X_j)) \\
    &= \frac{1}{N_1} \sum_{i=1}^{n} K_{i,n} \sum_{j \in S(i)} [\sigma^2(X_j) - \sigma^2(X_i)] + \frac{1}{N_1} \sum_{i=1}^{n} K_{i,n} (K_{i,n} + 1) \sigma^2(X_i) \\
    &= o_p(1) \text{ by Chebyshev’s inequality + Lemma 4} \\
    &= \frac{1}{N_1} \sum_{i=1}^{n} K_{i,n} (K_{i,n} + 1) \sigma^2(X_i) + o_p(1)
\end{align*}
\]

Meanwhile, the term \( ii \) is a “remainder term” which tends to zero in probability. This is because the terms \( \varepsilon_j \varepsilon_k \) are mean zero and conditionally uncorrelated given \( D \) so Chebyshev’s inequality allows us to write:

\[
\text{Var}(ii \mid D) = \frac{1}{N_1^2} \sum_{K_{i,n} > 0, j \in S(i), j \neq k} \mathbb{E}[\varepsilon_j \varepsilon_k | D]
\]

\[
\begin{align*}
    &= \frac{1}{N_1^2} \sum_{K_{i,n} > 0, j \in S(i), j \neq k} \sigma^2(X_i) \sigma^2(X_j) \\
    &\leq \frac{1}{N_1^2} \sum_{K_{i,n} > 0} \left( \sum_{j \in S(i)} \sigma^2(X_j) \right)^2 \\
    &= \frac{1}{N_1^2} \sum_{i=1}^{n} (K_{i,n} + 1)^2 \left( \frac{1}{K_{i,n} + 1} \sum_{j \in S(i)} \sigma^2(X_j) \right)^2 \\
    &\leq \frac{1}{N_1^2} \max(K_{i,n} + 1) \frac{1}{N_1} \sum_{i=1}^{n} \sigma^4(X_i) \\
    &= o_p(1) \times O_p(1)
\end{align*}
\]

where the last line is by Lemma 5 and the assumption that \( Y \) has more than four moments. By Chebyshev’s inequality, we may conclude that \( ii = o_p(1) \).

Combining \( i \) and \( ii \) shows that \( N_1 \text{Var}(\frac{1}{N_1} \sum_{j=1}^{n} \{\varepsilon_j - \varepsilon_{m^*(i)}\} | D) = np\sigma^2 + o_p(1) = N_1\sigma^2 + o_p(1) \). This implies the conclusion since Corollary 1 proves \( 1/(N_1\sigma^2) = O_p(1) \). \( \square \)
**Proposition 7.** (The randomization critical value)

Assume the conditions of Theorem 1. Let \( D = \{(X_i, Y_i, Z_i, U_i)\}_{i \leq n} \). Let \( \hat{q}_{1 - \alpha} = \inf \{ t \in \mathbb{R} : P(\{\hat{\tau}_s \leq t | D\}) \geq 1 - \alpha \} \) be the critical value from the stratified randomization test. Then \( \hat{q}_{1 - \alpha} = z_{1 - \alpha/2} \sigma + o_P(1/\sqrt{n}) \) where \( \sigma \) is defined in Corollary 1.

Proof. We begin by showing that \( \hat{\tau}_s / \sigma \overset{\sim}{\rightarrow} S N(0,1) \). Lemmas 26 and 27 allow us to write:

\[
\hat{\tau}_s / \sigma = \frac{1}{N_1} \sum_{i=1}^{N_1} \{ \varepsilon_i - \varepsilon_{m(i)} \} / \sigma, \quad \therefore \quad \sigma \end{align*}

Therefore, by the randomization Slutsky theorem (Lemma 3), it suffices to show that \( \frac{1}{N_1} \sum_{i=1}^{N_1} \{ \varepsilon_i - \varepsilon_{m(i)} \} / \sigma \overset{\sim}{\rightarrow} S N(0,1) \). Write the following expression:

\[
\frac{1}{N_1} \sum_{i=1}^{N_1} \{ \varepsilon_i - \varepsilon_{m(i)} \} = \sum_{K_i,n > 0} \frac{1}{N_1} \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \Rightarrow \xi_i.
\]

Conditional on \( D \), the terms \( \xi_i \) are mean zero and independent across \( i \). Thus, applying the Berry-Esseen theorem (Lemma 2) conditionally on \( D \) gives:

\[
\sup_{t \in \mathbb{R}} \left| P \left( \frac{1}{N_1} \sum_{i=1}^{N_1} \{ \varepsilon_i - \varepsilon_{m(i)} \} / \sigma \leq t \mid D \right) - \Phi(t) \right| \leq C \frac{K_{i,n} > 0}{\sqrt{N_1} \sigma^3} \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j \in S(i)} \left( \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \right)^3 \mid D \right]
\]

\[
\leq C \frac{K_{i,n} > 0}{\sqrt{N_1} \sigma^3} \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j \in S(i)} \left( \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \right)^3 \mid D \right]
\]

\[
\leq C \frac{K_{i,n} > 0}{\sqrt{N_1} \sigma^3} \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j \in S(i)} \left( \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \right)^3 \mid D \right]
\]

\[
\leq C \frac{K_{i,n} > 0}{\sqrt{N_1} \sigma^3} \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j \in S(i)} \left( \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \right)^3 \mid D \right]
\]

\[
\leq C \max_{i \leq n} (K_{i,n} + 1)^5 \frac{1}{\sqrt{N_1}} \frac{1}{(\sqrt{N_1} \sigma)^3} \frac{1}{N_1} \sum_{i=1}^{n} \left( \sum_{j \in S(i)} (Z_i^* - (1 - Z_i^*)K_{i,n}) \varepsilon_j \right)^3 \mid D \right]
\]

\[
= o_P(1) \quad \text{(by Lemma 5)}
\]

\[
= o_P(1) \quad \text{(by Lemma 27 + Corollary 1)}
\]

\[
= o_P(1). \]

Hence, \( \hat{\tau}_s / \sigma \overset{\sim}{\rightarrow} S N(0,1) \). The continuous mapping theorem for weak convergence in probability implies \( |\hat{\tau}_s / \sigma| \overset{\sim}{\rightarrow} |H| \). Since the distribution function of \( |H| \) is increasing at its \( (1 - \alpha) \)-th quantile, we conclude:

\[
\inf \{ t \in \mathbb{R} : P(|\hat{\tau}_s / \sigma| \leq t \mid D) \geq 1 - \alpha \} \overset{P}{\rightarrow} \inf \{ t \in \mathbb{R} : P(|H| \leq t) \geq 1 - \alpha \}
\]

\[
= z_{1 - \alpha/2}
\]

\[
= \bar{q}_{1 - \alpha} = \sigma \inf \{ t \in \mathbb{R} : P(|\hat{\tau}_s / \sigma| \leq t \mid D) \geq 1 - \alpha \}
\]

\[
= \sigma (z_{1 - \alpha/2} + o_P(1))
\]

\[
= \sigma z_{1 - \alpha/2} + o_P(1/\sqrt{n}) .
\]

\[ \square \]
A.9.4 Proof of the Theorem

Proof. It suffices to prove locally uniform asymptotic validity, since that certainly implies pointwise validity at \( h \equiv 0 \). Let \( \{ P_h \} \subset H_0 \) be any smooth parametric model with \( P_0 = P \). Let \( C < \infty \) arbitrary. Assume for the sake of contradiction that the claimed uniform convergence does not hold:

\[
\sup_{||h|| \leq C/\sqrt{n}} P_h(p < \alpha) \not\to \alpha.
\]

Then either \( \lim \sup \sup_{||h|| \leq C/\sqrt{n}} P_h(p < \alpha) > \alpha \) or \( \lim \inf \sup_{||h|| \leq C/\sqrt{n}} P_h(p < \alpha) < \alpha \). We shall assume that it is the former. The proof in the latter case is symmetric.

If \( \lim \sup \sup_{||h|| \leq C/\sqrt{n}} P_h(p < \alpha) = \beta > \alpha \), then along some subsequence \( \{ n_k \} \), we may find \( \{ h_{nk} \} \) with \( ||h_{nk}|| \leq C/\sqrt{n_k} \) for each \( k \) and \( \lim P_{h_{nk}}(p < \alpha) = \beta \). Since \( \{ \sqrt{n_k}||h_{nk}|| \} \) is a bounded sequence, we may find a further sub-subsequence \( \{ n_{kj} \} \) along which \( \sqrt{n_k}h_{nk_j} \to h \) for some \( h \). By our locally uniform asymptotic normality result (Proposition 5), we have:

\[
\hat{\tau}/\sigma \to^d N(0,1).
\]

Moreover, by Proposition 7 and the contiguity of \( \{ P_{h_{nk_j}} \} \) and \( \{ P_{0_{nk_j}} \} \) (Lemma 17), we have

\[
\hat{q}_{1-\alpha} = z_{1-\alpha}/2 + o_{P_{h_{nk_j}}}(1/\sqrt{n_k})
\]

Thus, by Slutsky’s theorem, we have:

\[
P_{h_{nk_j}}(p < \alpha) = P_{h_{nk_j}}(|\hat{\tau}| > \hat{q}_{1-\alpha}) = P_{h_{nk_j}}(|\hat{\tau}/\sigma| > \hat{q}_{1-\alpha}/\sigma) = P_{h_{nk_j}}(|\hat{\tau}/\sigma| > z_{1-\alpha}/2 + o_{P_{h_{nk_j}}}(1)) \to \alpha.
\]

This contradicts the result \( P_{h_{nk_j}}(p < \alpha) \to \beta > \alpha \). Thus, we must have \( \sup_{||h|| \leq C/\sqrt{n}} P_h(p < \alpha) \to \alpha \). \( \square \)

B Additional proofs

B.1 Proof of Lemma 4

In this section, we prove Lemma 4, which shows that \( K_{i,n} \) (the number of times the \( i \)-th observation is used as an untreated match) has uniformly bounded moments of all orders. Let \( D = \{(X_i, Z_i, U_i)\}_{i \leq n} \) be the dataset used in forming matches, where \( U_i \sim \text{Uniform}(0,1) \) (independently of everything else) is independent randomness used in the tie-breaking scheme of [13], which works as follows: among untreated observations \( j \) with minimal value of \( d_M(X_i, X_j) \), we choose the one which minimizes \( |U_i - U_j| \) as the match.

We use the following notation, which is slightly more descriptive than what was used in the main text:

\[
m(i, D) = \arg\min_{Z_j=0} d_M(X_i, X_j)
\]

\[
L_{i,n}(D) = \sum_{j=1}^{n} \mathbb{I}\{m(j, D) = i\}
\]

\[
K_{i,n}(D) = \sum_{j=1}^{n} Z_j \mathbb{I}\{m(j, D) = i\}
\]

In words, \( m(i, D) \) is the nearest untreated neighbor of observation \( i \). When there are no observations with \( Z_j = 0 \) in the dataset \( D \), we arbitrarily set \( m(i, D) = 0 \). \( L_{i,n}(D) \) counts the number of times observation \( i \) is the nearest untreated neighbor of another observation, and \( K_{i,n}(D) \) counts the number of times observation \( i \) is used as an untreated match. We also use the abbreviation “NN” for “nearest neighbors”.  

47
B.1.1 Preparation

**Lemma 28.** Let \( \gamma_d \) be the minimal number of cones centered at the origin of angle \( \pi/6 \) that cover \( \mathbb{R}^d \). Then for any distinct points \( x_1, \ldots, x_n \in \mathbb{R}^d \), we have:

\[
\sum_{j=1}^{n} \mathbb{I}\{x_1 \text{ is among the } k\text{-NN of } x_j \text{ in } \{x_i\}_{i \in [n] \setminus \{j\}}\} \leq k \gamma_d
\]

*Proof.* If matches are formed according to Euclidean distance, this is Corollary 11.1 in [14]. If matches are formed according to Mahalanobis distance, then replace \( x_i \) by \( (\Sigma^{-1})^{1/2} x_i \) (where we recall the convention \( \Sigma^{-1} = I_{d \times d} \) if \( \Sigma \) is singular) and apply the result in the Euclidean case. \( \square \)

**Lemma 29.** Independently of \( D \), let \( Z_2, \ldots, Z_n \overset{iid}{\sim} \text{Bernoulli}(1 - \delta) \). Let \( D' \) be the dataset \( D \) except \( Z_i \) is replaced by \( Z'_i \) for all \( i \geq 2 \). Then we have \( \mathbb{E}[L_{1,n}(D)^q] \leq \mathbb{E}[L_{1,n}(D')^q] \) for all \( q > 0 \).

*Proof.* We prove this by coupling the distributions of \( D \) and \( D' \). On some probability space, construct \( n \) independent random vectors \( (X_i, U_i, V_i) \sim P_{XY} \times \text{Uniform}(0,1) \times \text{Uniform}(0,1) \). Define \( Z_i = \mathbb{I}\{V_i > P(Z = 0|X)\} \) and \( Z'_i = \mathbb{I}\{V_i \geq \delta\} \). Since \( P(Z = 0|X) \geq \delta \) almost surely, \( Z'_i \geq Z_i \) with probability one. Set \( D = \{(X_i, Y_i, Z_i, U_i)\}_{i \leq n} \) and \( D' = \{(X_i, Y_i, Z'_i, U_i)\}_{2 \leq i \leq n} \cup \{(X_i, Y_i, Z'_i, U_i)\}_{1 \leq i \leq n} \).

With this coupling, it is easy to see that \( L_{1,n}(D) \leq L_{1,n}(D') \) almost surely. This is because changing some “untreated” units to “treated” only decreases the number of “competitors” of observation 1. Hence, \( L_{1,n}(D)^q \leq L_{1,n}(D')^q \) for any \( q > 0 \) and the conclusion follows. \( \square \)

**Lemma 30.** Let \( D' \) be as in Lemma 29. Then for all \( k \geq 2 \), we have:

\[
\mathbb{P}(m(2, D') = \ldots = m(k, D') = 1 \mid Z_1 = 0, n_0 = n_0) \leq \left( \frac{k \gamma_{d+1}}{n_0} \right)^{k-1}.
\]

*Proof.* Throughout this proof, we simply write \( m(i) \) in place of \( m(i, D') \).

Assume first that \( X_1, \ldots, X_n \) are almost surely distinct. In that case, \( m(r) = 1 \) can only happen if \( Z_r = 1 \); otherwise, \( X_r \) would be its own nearest untreated neighbor. Therefore, the pigeonhole principle implies that whenever \( (k-1) + n_0 > n \), one of the \( X \)'s must have \( Z_r = 0 \) and hence \( \mathbb{P}(m(2) = \ldots = m(k) = 1 \mid Z_1 = 0, N_0 = n_0) = 0 \). Clearly (22) holds in this case.

Now we consider the more interesting case where \( (k-1) + n_0 \leq n \). In this case, we will show that the bound holds even conditionally on \( (X_1, U_1) \). Begin by conditioning on the event \( Z_2 = \ldots = Z_k = 1 \) and translating the event into the language of nearest neighbors:

\[
\mathbb{P}(m(2) = \ldots = m(k) = 1 \mid X_1, U_1, Z_1 = 0, N_0 = n_0) \\
\leq \mathbb{P}(m(2) = \ldots = m(k) = 1 \mid X_1, U_1, Z_1 = 0, N_0 = n_0, Z_2 = \ldots = Z_k = 1) \\
= \mathbb{P}(X_1 \text{ is the NN of } \{X_r\}_{Z_r = 0} \text{ for all } r \in \{2, \ldots, k\} \mid X_1, U_1, Z_1 = 0, N_0 = n_0, Z_2 = \ldots = Z_k = 1).
\]

By symmetry, the probability in the upper bound is the same no matter which \( n_0 - 1 \) observations in \( \{1, \ldots, n\} \) are the ones with \( Z_i = 0 \). Therefore, we may as well assume they are \( k + 1, \ldots, k + 1 + n_0 \), which gives the bound:

\[
\mathbb{P}(m(2) = \ldots = m(k) = 1 \mid X_1, U_1, Z_1 = 0, N_0 = n_0) \\
\leq \mathbb{P}(\forall r \in [k]\backslash\{1\}, X_1 \text{ is the NN of } X_r \text{ in } \{X_i\}_{i \in [k-1+n_0]\{2,\ldots,k\}} \mid X_1, U_1, Z_1 = 0, Z_2 = \ldots = Z_k = 1, N_0 = n_0) \\
= \mathbb{P}(\forall r \in [k]\backslash\{1\}, X_1 \text{ is the NN of } X_r \text{ in } \{X_i\}_{i \in [k-1+n_0]\{2,\ldots,k\}} \mid X_1, U_1) \\
\leq \mathbb{P}(\forall r \in [k]\backslash\{1\}, X_1 \text{ is among the } k\text{-NN of } X_r \text{ in } \{X_i\}_{i \in [k-1+n_0]\{r\}} \mid X_1, U_1).
\]

Again by symmetry, the probability in the final line of the preceding display would be the same if we replaced \([k]\backslash\{1\}\) by any other set of \( k - 1 \) distinct indices in \([k-1+n_0]\backslash\{1\}\). Combining this observation with Lemma
28 gives the desired bound with $\gamma_d$ in place of $\gamma_{d+1}$.

$$
\Pr(X_1 = X_{m(r)} \text{ for all } r \in [k] \setminus \{1\} \mid X_1, U_1, Z_1 = 0, N_0 = n_0) \\
= \sum_{1 < i_1 < \ldots < i_{k-1} \leq k + n_0} \Pr(X_1 \text{ is among the } k\text{-NN of } X_{i_r} \text{ in } \{X_i\}_{i \in [k-1+n_0]} \text{ for each } i_r \mid X_1, U_1) \\
= \mathbb{E} \left[ \frac{1}{(n_0 + k - 1) \cdots (n_0 + k - 2) \cdots (n_0)} \sum_{1 < i_1 < \ldots < i_{k-1} \leq k + n_0} \mathbb{I}\{X_1 \text{ is among the } k\text{-NN of } X_{i_r} \text{ in } \{X_i\}_{i \in [k-1+n_0]} \text{ for each } i_r \mid X_1, U_1\} \right] \\
\leq \mathbb{E} \left[ \frac{1}{(n_0 + k - 1) \cdots (n_0)} \sum_{1 < i_1 < \ldots < i_{k-1} \leq k + n_0} \mathbb{I}\{X_1 \text{ is among the } k\text{-NN of } X_{i_r} \text{ in } \{X_i\}_{i \in [k-1+n_0]} \text{ for each } i_r \mid X_1, U_1\} \right] \\
\leq \left( \frac{k \gamma_d}{n_0} \right)^{k-1}.
$$

This proves the result in the case where the $X_i$’s are almost surely distinct. We only sketch the extension to the general case, which closely follows [13]. Replace $X_i$ by $\tilde{X}_i = (X_i, r(\varepsilon) U_i)$ where $r(\varepsilon)$ is so small that with probability at least $1 - \varepsilon$, nearest-neighbor matching based on the $\tilde{X}_i$’s is the same as matching based on the $X_i$’s and then tie-breaking using the $U_i$’s (this can always be achieved by choosing $r(\varepsilon)$ small enough so that all nonzero differences $\Vert X_i - X_j \Vert$ are larger than $r(\varepsilon)$ with probability at least $1 - \varepsilon$). Then, but for an additive slippage of $\varepsilon$, the above result applies to the $\tilde{X}_i$’s but with $\gamma_{d+1}$ instead of $\gamma_d$ since $\tilde{X}_i \in \mathbb{R}^{d+1}$. Finally, take $\varepsilon \downarrow 0$.

### B.1.2 Proof of the lemma

**Proof.** Since $K^q_{1,n} \leq L^q_{1,n}$, it suffices to prove the bound for $L_{1,n}$. Moreover, Jensen’s inequality implies we only need to consider integer values of $q$.

By Lemma 29, $\mathbb{E}[L_{1,n}(D)^q] \leq \mathbb{E}[L_{1,n}(D')^q \mid Z_1 = 0]$ where $D'$ is defined in the statement of that Lemma. We control the expectation in this upper bound by first conditioning on $N_0$.

$$
\mathbb{E}[L_{1,n}(D')^q \mid Z_1 = 0, N_0] = \mathbb{E} \left[ \left( 1 + \sum_{i=2}^{n} \mathbb{I}\{m(2, D') = 1\} \right)^q \mid Z_1 = 0, N_0 \right] \\
\leq 2^{q-1} \left\{ 1 + \mathbb{E} \left[ \left( \sum_{i=2}^{n} \mathbb{I}\{m(2) = 1\} \right)^q \right] \right\} \\
= 2^{q-2} \left\{ 1 + \sum_{2 \leq i_1 \ldots i_n \leq n} \Pr(m(i_1, D') = \ldots = m(i_q, D') = 1 \mid Z_1 = 0, N_0) \right\}
$$

Exchangeability implies that $\Pr(m(i_1, D') = \ldots = m(i_q, D') = 1 \mid Z_1 = 0, N_0)$ only depends on the number of distinct indices $(i_1, \ldots, i_q)$ and not on the identity of those indices. For any $\ell \leq q$, the number of sequences $(i_1, \ldots, i_q) \in [n-1]^q$ with $\ell$ distinct indices is at most $\ell^q \binom{n-1}{\ell}$. Thus, Lemma 30 gives the further bound:

$$
\mathbb{E}[L_{1,n}(D')^q \mid Z_1 = 0, N_0] \leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} \ell^q \binom{n-1}{\ell} \Pr((m(2) = \ldots = m(\ell+1) = 1 \mid Z_1 = 0, N_0) \right\} \\
\leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} \ell^q \binom{n-1}{\ell} \left( \frac{1}{N_0} \right)^{\ell} \right\} \\
\leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} \ell^q \gamma_{d+1}(\ell+1)^\ell(n-1)^\ell \left( \frac{1}{N_0} \right)^{\ell} \right\}.
$$

49
Now, we take expectations over $N_0$ on the far left-hand side and far right-hand side of the preceding display. The conditional distribution of $N_0$ given $Z_1 = 0$ stochastically dominates that of $1 + N'_0$ where $N'_0 \sim \text{Bernoulli}(n-1, \delta)$. By standard binomial concentration, it can be shown that for any $\ell$, there exists $C(\ell) < \infty$ and $N_\ell \geq 1$ such that $\mathbb{E}[(1/1 + N'_0)_{\ell}] \leq C(\ell)/[(n-1)\delta]^\ell$ for all $n \geq N_\ell$. Therefore, we may conclude:

$$\mathbb{E}[L_{1,n}(D') q | Z_1 = 0] \leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} \ell^q (\gamma_{d+1}(\ell + 1))^{\ell} (n-1)^\ell \frac{1}{1 + N'_0} \right\}$$

$$\leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} \ell^q (\gamma_{d+1}(\ell + 1))^{\ell} (n-1)^\ell \frac{C(\ell)}{\delta(n-1)^\ell} \right\}$$

$$\leq 2^{q-1} \left\{ 1 + \sum_{\ell=1}^{q} C(\ell) \gamma_{d+1}(\ell + 1) \right\}.$$

Since this upper bound does not depend on $n$ and holds for all large $n$, we conclude that $\mathbb{E}[L_{1,n}(D') q | Z_1 = 0]$ is uniformly bounded in $n$. Since $\mathbb{E}[K_{1,n}(D) q] \leq \mathbb{E}[L_{1,n}(D') q] \leq \mathbb{E}[L_{1,n}(D') q | Z_1 = 0]$, this proves the result.