Variance estimation after matching or re-weighting

Xiang Meng, Aaron Smith, Luke Miratrix

May 14, 2025

Abstract

This paper examines the theoretical foundations and practical implementation of inference methods for matching estimators in causal analysis. We identify that existing bootstrap procedures can produce unreliable inference when there is substantial overlap in matched samples—a common scenario when the number of treated units is small relative to controls. To address this challenge, we make two main contributions: First, we analyze an alternative inference procedure that demonstrates robust performance in simulations when the wild bootstrap fails to produce valid confidence intervals under substantial overlap conditions. This variance estimator maintains validity even with extensive control unit reuse, outperforming existing approaches. Second, we develop a theoretical framework that rigorously justifies the estimator's validity by establishing its consistency and asymptotic normality.

Our contributions include a theoretical analysis of a practical variance estimator previously proposed in the literature, a generalized framework with novel conditions that relax traditional requirements for matching estimators, and extensions of our approach to other causal inference estimators such as weighting methods. Through carefully designed simulation studies, we show that our estimator maintains proper coverage while state-of-the-art alternatives fail in common scenarios with overlapping matches. Our framework provides researchers with both theoretical guarantees and practical tools for conducting valid inference in a wide range of causal inference applications.

1 Introduction

Matching and weighting estimators are fundamental tools in causal inference for estimating treatment effects. Matching methods pair treated units with similar control units based on observed covariates (Rosenbaum and Rubin, 1983; Rubin, 1973) while weighting approaches adjust for confounding by reweighting observations to achieve covariate balance (Hirano et al., 2003; Imbens, 2004). These methods have been widely adopted across diverse fields, including economics (Dehejia and Wahba, 1999; Heckman et al., 1997), epidemiology (Stuart, 2010), and policy evaluation (Smith and Todd, 2005).

Abadie and Imbens (2006) established the foundational asymptotic theory for matching estimators, revealing their nonstandard behavior due to the fixed number of matches. Unlike other nonparametric treatment effect estimators (Heckman et al., 1998; Hirano et al., 2003), matching estimators are generally not $\sqrt{N}$-consistent. This discovery led to important developments in bias correction (Abadie and Imbens, 2011) and inference methods. Notably, Abadie and Imbens (2008) demonstrated that the standard bootstrap fails to provide valid inference for matching estimators, recommending instead the use of analytical standard errors or subsampling methods (Politis and Romano, 1994).

Otsu and Rai (2017) proposed a wild bootstrap procedure that is theoretically valid under certain asymptotic conditions. However, our empirical investigations reveal that this procedure can produce unreliable inference when there is substantial overlap in matched samples—a common scenario when the number of treated units is small but the number of control units is large. Here, by overlap, we means the phenomenon where matched controls for different treated units contain many common control units; that is, the same control units appear repeatedly in the matched sets of multiple treated units. This overlap creates complex dependencies that are not properly accounted for in existing inference procedures (Abadie and Imbens, 2012). This dependency structure becomes particularly problematic when constructing confidence intervals using bootstrap methods, as the resampling scheme fails to capture the true variance of the estimator.

We address this challenge through two main contributions. First, we analyze an alternative inference procedure that demonstrates robust performance in simulations when the wild bootstrap fails to produce valid confidence intervals under substantial overlap conditions. This procedure

is a Wald-type confidence interval based on a variance estimator that has been previously used in empirical work (e.g., Che et al. (2024); Keele et al. (2023)), but whose comparative advantage over state-of-the-art methods like Otsu and Rai (2017) has not been systematically established. Our comprehensive comparison demonstrates that this estimator maintains validity even under substantial overlap in matched samples, outperforming existing approaches. Second, we develop a theoretical framework that rigorously justifies the estimator's validity by establishing its consistency and asymptotic normality, providing practitioners with a sound statistical foundation for its application.

Our contributions are as follows:

- **Theoretical Analysis of a Practical Variance Estimator:** We provide a rigorous theoretical analysis of a variance estimator previously proposed in Che et al. (2024). Our analysis demonstrates that this estimator is computationally efficient, theoretically justified, and more practical than alternatives. Unlike the estimator in Abadie and Imbens (2006) which requires matching within both treatment groups, this approach only requires matching treated units to controls, making it particularly valuable for effect estimation with small treated samples. Furthermore, we find that this estimator can outperform current state-of-the-art alternatives in settings that mimic real-world matching contexts. In particular, our estimator remains robust when control units are reused as matches for multiple treated units. Through carefully designed simulation studies, we show that it maintains proper coverage while the wild bootstrap method proposed by Otsu and Rai (2017), which is the current state-of-the-art method for inference in matching, fails in these common scenarios. We also establish that the estimator possesses heteroskedasticity-consistent properties, drawing important parallels to the well-known Huber-White White (1980) robust standard errors in regression analysis. This connection creates a theoretical bridge between matching-based and regression-based inference methods, unifying seemingly disparate approaches to variance estimation in causal inference.

- **Generalized Theoretical Framework:** While Che et al. (2024) use this approach in the context of radial matching, we show that our framework applies to a variety of matching contexts including nearest neighbor matching, propensity score matching, and synthetic control weighting. We introduce two novel conditions to the matching literature that together create

a more practical framework for valid inference in matching estimators. First, our derivative control condition relaxes traditional requirements about how outcome functions can change. While previous work by Abadie and Imbens (2006) required the outcome function to change at a constant rate everywhere (Lipschitz continuity), our approach allows the rate of change to vary across the covariate space as long as it is appropriately balanced by the size of matched clusters. This innovation permits valid inference even with functions that have steep gradients in some regions. Second, our shrinking clusters assumption does not specify how quickly matches must improve with sample size, only that they eventually become arbitrarily close. This flexibility, new to the matching literature, accommodates many matching methods beyond the specific M-NN approach in earlier work, including radius matching and propensity score techniques. Together, these conditions significantly expand the applicability of matching methods while ensuring theoretical guarantees, allowing researchers to apply these techniques across a wider range of real-world research scenarios with greater confidence in their statistical properties.

- **Extended Applications to Other Causal Inference Estimators:** We demonstrate that the variance estimation framework extends beyond matching to other causal inference methods, particularly weighting estimators Zubizarreta (2015); Wang and Zubizarreta (2019). Through additional simulations with challenging datasets, we show that our variance estimation approach maintains proper coverage when applied to stable balancing weights, suggesting potential for creating a more coherent inference framework for both matching and weighting approaches in causal inference.

2 Problem Setup

2.1 Model

We consider a setting with n observations, each representing a unit in our study population. The sample consists of n_T treated units and n_C control units, with $n = n_T + n_C$.

For each unit i, we observe a tuple $\{Z_i, Y_i, \mathbf{X}_i\}$ where:

- $Z_i \in \{0, 1\}$ denotes its binary treatment status.

- $Y_i \in \mathbb{R}$ denotes its observed real-valued outcome.
- $\mathbf{X}_i \equiv \{X_{1i}, \ldots, X_{ki}\}^T \in \mathbb{R}^k$ denotes its k-dimensional real-valued covariate vector.

We adopt the potential outcomes framework where each unit has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. Here, $Y_i(1)$ represents the outcome if unit i receives treatment, and $Y_i(0)$ represents the outcome if unit i does not receive treatment. The fundamental problem of causal inference is that we only observe one of these potential outcomes for each unit. Specifically, the observed outcome for unit i is $Y_i \equiv (1-Z_i)Y_i(0)+Z_iY_i(1)$ under the stable unit treatment value assumption (SUTVA).

We assume the data generating process follows independent and identically distributed (i.i.d.) sampling of the potential outcome tuples $\{Y_i(0), Y_i(1), Z_i, \mathbf{X}_i\}_{i=1}^n$. For each unit i, the generic random variables $(Y(0), Y(1), Z, \mathbf{X})$ represent the population distribution from which the observed data are drawn. Throughout the paper, indexed variables (e.g., $\mathbf{X}_i$) refer to specific observations, while non-indexed variables (e.g., $\mathbf{X}$) refer to the generic random variables representing the population distribution.

To proceed with estimation, we make the following assumptions:

Assumption 1 (Compact support)**.** *The covariate vector* $\mathbf{X}$ *is a* k*-dimensional random vector with components that are continuous random variables, distributed on* $\mathbb{R}^k$ *with compact support* $\mathbb{X}$*. The density of* X *is bounded and bounded away from zero on its support.*

The compact support assumption helps ensure that the covariate space is well-behaved, which facilitates consistent estimation and rules out pathological cases where the distribution of covariates becomes too sparse or unbounded.

Assumption 2 (Unconfoundedness and overlap (Rubin, 1974))**.** *For almost every* $x \in \mathbb{X}$*:*

1. $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid \mathbf{X}$
2. $\Pr(Z = 1 \mid X = x) < 1 - \eta$ *for some* $\eta > 0$

This assumption states that, conditional on the observed covariates, treatment assignment is independent of the potential outcomes, and that both treated and control units are sufficiently represented across the covariate space.

Importantly, the treatment indicators Z_i are randomly drawn according to the treatment assignment mechanism, which implies that the number of treated units n_T is a random quantity even when the total sample size n is fixed. This leads to our next assumption:

Assumption 3 (Sampling)**.** *Conditional on $Z_i = z$, the sample consists of independent draws from the distribution of $(Y, X|Z = z)$ for $z \in \{0, 1\}$. As the sample size $n \to \infty$, we have $n_T^r/n_C \to \theta$ for some $r \geq 1$ and $0 < \theta < \infty$.*

We further assume a model where potential outcomes are generated as:

$$Y_i(0) = f_0(\mathbf{X}_i) + \epsilon_{0,i}$$
$$Y_i(1) = f_1(\mathbf{X}_i) + \epsilon_{1,i}$$

Here, $f_0(\mathbf{X}) \equiv E[Y(0)|\mathbf{X}]$ and $f_1(\mathbf{X}) \equiv E[Y(1)|\mathbf{X}]$ are the true conditional expectation functions of the potential outcomes under control and treatment, respectively (often referred to as "response surfaces" in the causal inference literature (Hahn et al., 2020; Hill, 2011)). The error terms $\epsilon_{0,i}$ and $\epsilon_{1,i}$ represent the deviations of the individual potential outcomes from their respective conditional expectations, with conditional variances $\sigma^2_{0,i}$ and $\sigma^2_{1,i}$ respectively. Further distributional assumptions about these error terms are detailed in Section 3.2.

2.2 Estimand and the Estimator

We define $\tau(\mathbf{X}_i) = f_1(\mathbf{X}_i) - f_0(\mathbf{X}_i)$ as the systematic treatment effect, which represents the systematic component of the treatment effect for units with covariates $\mathbf{X}_i$. Note that the individual treatment effect $Y_i(1) - Y_i(0) = \tau(\mathbf{X}_i) + (\epsilon_{1,i} - \epsilon_{0,i})$ includes both this systematic component and an idiosyncratic component. We consider the estimand to be the population average treatment effect on the treated (ATT)

$$\tau \;=\; E\big[f_1(X_i) - f_0(X_i) \mid Z_i = 1\big],$$

We write the set of all treated units' indices as $\mathcal{T} = \{i : Z_i = 1\}$, the set of all control units' indices as $\mathcal{C} = \{i : Z_i = 0\}$, and $t \in \mathcal{T}$, $j \in \mathcal{C}$ as individual treated and control units respectively. We denote the set of indices of control units matched to a treated unit $t \in \mathcal{T}$ as $\mathcal{C}_t = \{j \in \mathcal{C} :$ unit j is matched to unit $t\}$, which is determined by a matching procedure that maps the observed

data to these match sets. Finally, we denote the size of a set $\mathcal{S}$ as $|\mathcal{S}|$.

A matching procedure pairs each treated unit with one or more control units that have similar covariate values, thus approximating the counterfactual outcome for the treated unit. Given such a matching procedure, we define the matching estimators for the ATT:

$$\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j\big). \tag{1}$$

where $w_{jt} \in [0, 1]$ is the weight assigned to the matched control unit j for treated unit t, with $\sum_{j \in \mathcal{C}_t} w_{jt} = 1$ for each $t \in \mathcal{T}$. For example, in M-nearest neighbor (M-NN) matching (Rubin, 1973; Abadie and Imbens, 2006; Stuart, 2010), $\mathcal{C}_t$ consists of the closest M neighbors to unit t based on covariate distance, and each neighbor receives equal weight $w_{jt} = 1/M$. Another example is the synthetic control approach in Che et al. (2024), which first obtains $\mathcal{C}_t$ local radius matching, and then determines the weights w_{jt} by solving a convex optimization problem that minimizes the distance between the treated unit's covariates $\mathbf{X}_t$ and the weighted average of control units' covariates $\sum_{j \in \mathcal{C}_t} w_{jt} \mathbf{X}_j$.

Define the matching radius for a treated unit t with covariate value $\mathbf{X}_t$ as:

$$r\left(\mathcal{C}_t\right) = \sup_{j \in \mathcal{C}_t} \left\|\mathbf{X}_t - \mathbf{X}_j\right\|.$$

This radius represents the maximum distance between a treated unit and any of its matched controls. The probabilistic properties of this radius will be crucial for establishing our theoretical results.

Assumption 4 (Exponential Tail Condition)**.** *The matching radius satisfies:*

$$P\left(n_C^{1/k} r\left(\mathcal{C}_t\right) > u\right) \leq C_1 \exp\left(-C_2 u^k\right),$$

where C_1, C_2 are positive constants, k is the dimension of the covariate space, and M is the number of matches.

This assumption requires that the probability of having a large (scaled) matching radius decays exponentially, which ensures that the matches become increasingly accurate as the sample size

grows. This is a more precise characterization than simply requiring that clusters shrink asymptotically, as it specifies the rate at which the tail probability diminishes.

Several common matching methods satisfy the exponential tail condition under appropriate implementation, including M-NN matching and radius matching. When using fixed M-NN matching, the exponential tail condition is satisfied provided that the density of covariates is bounded and has overlapped support, as established by Abadie and Imbens (2006). In this approach, each treated unit is matched to its M closest control units based on covariate distance, with each neighbor receiving equal weight $w_{jt} = 1/M$.

Another approach is radius matching, where we set the matching radius as $D(n_C) = cn_C^{-1/k}$. To provide an intuitive understanding of this choice, we can use the following heuristic argument. This choice ensures two important properties. First, the probability of obtaining at least one match for a treated unit approaches 1 as $n_C \to \infty$. Heuristically, the expected number of controls in a ball of radius $D(n_C)$ around a treated unit with covariates $\mathbf{X}_t$ is approximately

$$n_C \cdot f(\mathbf{X}_t)\, v_k\, [c\, n_C^{-1/k}]^k \approx f(\mathbf{X}_t)\, v_k\, c^k,$$

where v_k is the volume of the unit ball in $\mathbb{R}^k$ and $f(\mathbf{X}_t)$ is the density at $\mathbf{X}_t$.

Second, this matching scheme has an exponential tail for the scaled discrepancy $n_C^{1/k}\|\mathbf{X}_j - \mathbf{X}_t\|$. By analyzing the order statistics of nearest neighbor distances and applying large-deviation bounds, we can show that

$$\Pr\big(n_C^{1/k}\|\mathbf{X}_j - \mathbf{X}_t\| > u\big) \le C_1 \exp(-C_2 u^k)$$

for some constants $C_1, C_2 > 0$, thus satisfying the exponential tail condition.

In the next section, we address the inference problem, focusing on the asymptotic normality of the matching estimator and the decomposition of its variance components. This provides the foundation for constructing valid confidence intervals. Following this, we turn our attention to the crucial challenge of variance estimation. We introduce a consistent estimator that accounts for both homoskedastic and heteroskedastic error structures, refining previous approaches to improve efficiency and robustness.

3 The Inference Problem

To construct valid confidence intervals for our matching estimator $\hat{\tau}$, we require asymptotic normality of the form:

$$\frac{\sqrt{n_T}\,(\hat{\tau}-\tau)}{V^{-1/2}} \xrightarrow{d} N(0,1).$$

The difference between the matching estimator $\hat{\tau}$ and the estimand τ can be decomposed into three components:

$$\hat{\tau}-\tau = \hat{\tau}-\tau_{\text{SATT}}+\tau_{\text{SATT}}-\tau = B_n + E_n + P_n \tag{2}$$

where

$$B_n = \frac{1}{n_T}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{C}_t} w_{jt}\big(f_0(X_t)-f_0(X_j)\big)$$

represents bias from imperfect covariate matching.

$$\begin{aligned} E_n &= \frac{1}{n_T}\sum_{t\in\mathcal{T}}\Big(\epsilon_t - \sum_{j\in\mathcal{C}_t} w_{jt}\epsilon_j\Big) \\ &= \frac{1}{n_T}\sum_{t\in\mathcal{T}}\epsilon_t - \frac{1}{n_T}\sum_{j\in\mathcal{C}} w_j\epsilon_j \end{aligned}$$

captures measurement error from random variation in unobserved factors.

$$P_n = \tau_{\text{SATT}} - \tau$$

measures population error between sample and population treatment effects.

where τ_{SATT} is the sample average treatment effect on the treated (SATT):

$$\tau_{\text{SATT}} = \frac{1}{n_T}\sum_{t\in\mathcal{T}}\big(f_1(X_t)-f_0(X_t)\big).$$

We now analyze the key components of our inference framework in detail. Section 3.1 examines the bias term and its asymptotic behavior under different matching schemes. Section 3.2 introduces critical assumptions about error variance that underpin our theoretical results. Section 3.3 devel-

ops the variance decomposition, separating contributions from measurement error and population heterogeneity. Together, these elements establish the foundation for our central limit theorem.

3.1 The Bias Term and Its Convergence Rate

A crucial challenge in establishing the asymptotic normality of matching estimators is the slow convergence rate of the bias term. Following Abadie and Imbens (2006), under regularity conditions on data distribution introduced at Section 2.1, this bias term converges at a rate of $O_p(n_T^{-1/k})$, where k is the dimension of the covariate space. This rate is typically slower than the $n_T^{-1/2}$ rate required for standard asymptotic normality results.

Proposition 3.1 (Bias convergence rate)**.** *Under Assumptions 1, 2, and 3, if $f_0(x)$ is Lipschitz continuous on $\mathbb{X}$, then*

$$Bias = O_p(n_T^{-1/k})$$

This slow convergence rate of the bias term necessitates explicit bias correction for valid inference. While various approaches to bias correction exist, including the method proposed by Abadie and Imbens (2011), our focus in this paper will be on variance estimation conditional on a bias correction procedure.

3.2 Error Variance Assumptions

To analyze the large-sample behavior of our variance estimator, we place structure on the conditional variance of the potential outcomes. This section introduces the regularity conditions we require on the variance functions $\sigma_0^2(x)$ and $\sigma_1^2(x)$.

Let us denote the conditional variances of the potential outcomes as:

$$\begin{aligned} \sigma_{0,i}^2 &= E\big[\big(Y_i(0) - f_0(\mathbf{X}_i)\big)^2 \,\big|\, \mathbf{X}_i\big] = E[\epsilon_{0,i}^2 \,|\, \mathbf{X}_i], \\ \sigma_{1,i}^2 &= E\big[\big(Y_i(1) - f_1(\mathbf{X}_i)\big)^2 \,\big|\, \mathbf{X}_i\big] = E[\epsilon_{1,i}^2 \,|\, \mathbf{X}_i]. \end{aligned} \tag{3}$$

We now define a class of variance functions with properties that enable consistent estimation in the matched setting.

Definition 3.1 (Regular variance function)**.** *A function $\sigma^2 : \mathcal{X} \to \mathbb{R}_+$ is said to be a regular variance function if it satisfies the following:*

- ***Uniform continuity.*** *$\sigma^2(\cdot)$ is uniformly continuous (or Lipschitz) on the support $\mathcal{X} \subset \mathbb{R}^d$ of X.*
- ***Boundedness.*** *There exist constants $0 < \sigma^2_{\min} < \sigma^2_{\max} < \infty$ such that*

$$\sigma^2_{\min} \leq \sigma^2(x) \leq \sigma^2_{\max} \quad \textit{for all } x \in \mathcal{X}.$$

- ***Higher-order moment bound.*** *There exists a constant $C < \infty$ and an exponent $\delta > 0$ such that*

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big[\big|\epsilon_i\big|^{2+\delta} \;\big|\; X_i = x\big] \;\leq\; C.$$

The first condition ensures that matched units have similar variances. Specifically, for any matching scheme with $\|X_{tj} - X_t\| \to 0$ (as guaranteed by Assumption 4), we have $\sigma^2(X_{tj}) \to \sigma^2(X_t)$. Hence, $\sigma^2_j \approx \sigma^2_t$ for $j \in \mathcal{C}_t$ whenever $\mathcal{C}_t$ is constructed by matching on X. In particular,

$$\max_{j \in \mathcal{C}_t} \big|\sigma^2(X_{tj}) \;-\; \sigma^2(X_t)\big| \;\to 0,$$

provided that $\max_{j \in \mathcal{C}_t} \|X_{tj} - X_t\| \to 0$. *Remark:* This generalizes Assumption 4.1 in Abadie and Imbens (2006), which assumes Lipschitz continuity.

The second condition ensures that the conditional variance is bounded away from both zero and infinity. The lower bound prevents degeneracy in the asymptotic distribution and ensures that confidence intervals have positive width. While it is theoretically possible for the variance to approach zero, this would imply that all outcome variability is explained by covariates and there is no residual noise — i.e., $\sigma^2(x) \to 0$ for all x. In this case, the contribution of the error term to the sampling variability of $\hat{\tau}$ would vanish. The resulting inference problem becomes degenerate: estimation is still possible, but all uncertainty would stem entirely from treatment effect heterogeneity, not from residual noise. This setting is simpler but often unrealistic, as in practice we typically expect some irreducible measurement error in outcomes. The upper bound limits the influence of outliers, which is needed to establish convergence rates.

The third condition imposes a uniform bound on a higher-order conditional moment of the errors. This assumption is standard in high-dimensional estimation and facilitates the use of maximal inequalities and uniform convergence tools.

We now formally state the assumption we make on the conditional variances of the potential outcomes:

Assumption 5 (Regular error variances)**.** *We assume that both $\sigma_0^2(x)$ and $\sigma_1^2(x)$ are regular variance functions.*

Together, these three properties ensure that both the level (expected magnitude of the errors) and the variability (how much the errors fluctuate around their means) of the error process are well-behaved across the full range of covariates. This structure plays a key role in enabling consistent variance estimation under matching, as we will see in the following sections.

3.3 The Form of the Asymptotic Variance

We now analyze the asymptotic variance of the matching estimator $\hat{\tau}$. Recall from Equation (2) that the estimation error decomposes as:

$$\hat{\tau} - \tau = B_n + E_n + P_n,$$

where B_n captures bias from covariate mismatch, E_n captures sampling error due to residual outcome noise, and P_n captures the discrepancy between the sample and population average treatment effect on the treated (ATT). As discussed earlier, under appropriate regularity conditions and bias correction, the asymptotic distribution of $\hat{\tau}$ is primarily governed by the variability in the two stochastic terms: E_n and P_n.

We are therefore interested in the asymptotic variance of $\hat{\tau}$ as determined by:

$$\mathrm{Var}(\hat{\tau} - \tau) \approx \frac{V_E}{n_T} + V_P,$$

where V_E reflects the contribution from measurement noise (variance of E_n), V_P reflects the population-level heterogeneity in treatment effects among treated units (variance of P_n).

Measurement Error Component V_E. We first consider the component due to residual outcome noise. Conditional on the covariates $\mathbf{X}$ and treatment assignment vector $\mathbf{Z}$, the variance of E_n is given by:

$$\begin{aligned} V_E :=& \mathbb{E}[E_n^2 \mid \mathbf{X}, \mathbf{Z}] \\ =& \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma_{1,t}^2 + \sum_{j \in \mathcal{C}} (w_j)^2 \sigma_{0,j}^2 \right), \end{aligned} \tag{4}$$

where $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all matched treated units.

This decomposition reflects how residual variance enters the estimator: treated units contribute through their own variances, and controls contribute via squared weight accumulation. Reused controls (with large w_j) disproportionately affect the overall variance. Prior work including Kallus (2020) and Che et al. (2024) use this variance structure to study the bias-variance tradeoff in matching estimators—highlighting that tighter matches (which reduce bias) can increase variance due to heavy reuse of control units.

Importantly, in the central limit theorem (CLT), the contribution of V_E is scaled by n_T. Since $E_n = O_p(n_T^{-1/2})$, we have:

$$\sqrt{n_T}(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, V_E + V_P),$$

as formally stated in Theorem 3.2.

Population Heterogeneity Component V_P. The second term $P_n = \tau_{\text{SATT}} - \tau$ captures how the realized sample of treated units may differ from the target population of treated units. That is, even if outcomes were observed without error, the sample ATT may deviate from the population ATT due to treatment effect heterogeneity.

To clarify this, define:

$$V_P' = \text{Var}(\tau_{\text{SATT}} \mid \mathbf{Z}) = \frac{1}{n_T} \mathbb{E}\left[(\tau(X_i) - \tau)^2 \mid Z_i = 1 \right],$$

so that

$$V_P = n_T \cdot V_P'$$

captures the contribution of this sampling variation to the CLT variance.

Putting both pieces together, the asymptotic variance of $\hat{\tau}$ is given by:

$$\text{Var}(\hat{\tau}) \approx \frac{V_E}{n_T} + V_P,$$

which corresponds directly to the decomposition into E_n and P_n discussed above. The variance term V_E arises from residual outcome variation, while V_P reflects how treatment effect heterogeneity among the treated translates into sampling variability. In the next section, we show how to consistently estimate each component from observed data.

3.4 The Central Limit Theorem

We now present our main asymptotic normality result, which forms the basis for valid inference.

Theorem 3.2 (Central Limit Theorem)**.** *Under Assumptions 1, 2, 3, 4 and 5, as $n_T \to \infty$:*

$$\frac{\sqrt{n_T}\left(\hat{\tau} - B_n - \tau\right)}{V^{-1/2}} \xrightarrow{d} N(0,1),$$

where

$$V = V_E \cdot n_T + V_P.$$

In the special case where the dimension of the covariate space satisfies $k \leq 2$, the bias term B_n becomes negligible at a faster rate, yielding:

$$\frac{\sqrt{n_T}\left(\hat{\tau} - \tau\right)}{V^{-1/2}} \xrightarrow{d} N(0,1).$$

This theorem generalizes the seminal results of Abadie and Imbens (2006), which were limited to M-NN matching with uniform weights ($w_{jt} = 1/M$). Our framework makes two significant advances: (a) it accommodates arbitrary matching procedures including radius matching, caliper matching, and optimal matching; and (b) it allows for flexible weighting schemes such as kernel weights, bias-corrected weights, and synthetic control weights. This increased flexibility enables practitioners to choose matching methods that better balance bias reduction and variance minimization for their specific applications.

For practical implementation of inference procedures, we develop a consistent estimator $\hat{V}$ for

the asymptotic variance V in the subsequent section. By Slutsky's theorem, this will yield:

$$\frac{\sqrt{n_T}\left(\hat{\tau} - B_n - \tau\right)}{\hat{V}^{-1/2}} \xrightarrow{d} N(0,1).$$

This result provides the foundation for constructing asymptotically valid confidence intervals for the treatment effect. The next section will focus on the consistent estimation of the variance components required for implementation.

4 The Standard Error Estimator

To establish valid inference for matching estimators, we analyze a standard error estimation strategy that accommodates both homogeneous and heterogeneous error structures. This approach, previously used in establishing local radius matching in Che et al. (2024) lacks thorough theoretical justification. It is different from existing methods by relaxing traditional assumptions while maintaining consistency under general matching procedures. Our theoretical analysis provides the missing foundation for this estimator's widespread application.

The organization of this section is as follows: In Section 4.1, we introduce the derivative control condition that generalizes previous assumptions; in Section 4.2, we present the formal variance estimator and establish its consistency; and in Section 4.3, we compare our approach with previous estimators in the literature, highlighting its practical advantages.

4.1 Derivative Control Condition

Before introducing the variance estimator, we first establish a key theoretical condition that enables our analysis. The derivative control condition presented below is more general than the Lipschitz continuity assumption on f used in Abadie and Imbens (2006), and allows for broader applicability in settings where f' is not uniformly bounded.

Assumption 6 (Derivative control)**.** *Let f be differentiable on the support of X, and denote its derivative by $f' : \mathcal{X} \to \mathbb{R}$. There exists a constant $M < \infty$ (possibly depending on n) such that, for*

all $t \in \mathcal{T}$*,*

$$\sup_{x \in \mathcal{C}_t} \left| f'(x) \right| \cdot r(\mathcal{C}_t) \leq M,$$

or equivalently,

$$\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} \left| f'(x) \right| \cdot r(\mathcal{C}_t) \right] < \infty.$$

This condition ensures that in regions where the derivative $f'(x)$ is large, the matching clusters $\mathcal{C}_t$ are sufficiently tight—so that the product of local slope magnitude and cluster size remains uniformly bounded. In contrast to the Lipschitz condition, which imposes a global bound on f', this condition accommodates functions with steep regions, as long as tighter matches are used locally.

To illustrate the practical advantage, consider $f(x) = x^2$ on $[0, 100]$, where $|f'(x)| = 2|x|$ grows with x, and a global Lipschitz constant would be $L = 200$. Such a large constant makes inference difficult in finite samples. Our condition instead allows for looser matches in flatter regions and tighter matches in steeper regions—offering better practical guidance for match design.

Together with the shrinking cluster condition in Assumption 4, this implies:

$$\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} \left| f'(x) \right| \cdot r(\mathcal{C}_t) \right] \xrightarrow[n \to \infty]{} 0,$$

under the mild requirement that f' is continuous and the support of X remains in a compact region. Since f' is fixed (i.e., does not grow with n), and $r(\mathcal{C}_t) \to 0$ uniformly in t by Assumption 4, this convergence follows directly. This vanishing bound ensures that local linear approximations to f within each matched cluster incur asymptotically negligible error.

With this theoretical foundation established, we now turn to the variance estimator itself and analyze its consistency properties.

4.2 Proposed Variance Estimator

In this section, we introduce our standard error estimator for matching estimators. We begin by presenting the underlying modeling assumptions, then develop the formula for our proposed estimator, and finally establish the consistency results.

Assumption 7 (Homoskedasticity and Regular Variance)**.** *We assume that each unit has the same conditional variance under both treatment and control:*

$$\sigma_{0,i}^2 = \sigma_{1,i}^2 = \sigma_i^2.$$

Furthermore, we assume σ_i^2 is regular in the sense of Definition 3.1.

This homoskedasticity assumption simplifies the derivation and allows for tractable plug-in variance formulas. While the assumption may seem restrictive—since in practice the variance may differ across potential outcomes—it serves as a useful approximation, especially when matching quality is high and clusters are tight. By assuming a single variance function $\sigma^2(x)$ governs both outcomes, we avoid needing to estimate two separate variance surfaces.

The variance V consists of two components: the measurement error variance V_E and the population heterogeneity variance V_P. We begin by developing an estimator for V_E, which presents more technical challenges, before extending our approach to estimate the full variance V. Our methodology for V_E establishes key techniques that will later be applied to the full variance estimator. In both cases, our primary theoretical contribution is proving consistency of these estimators under general conditions.

4.2.1 A Consistent Estimator for V_E

To build intuition for our approach, let us first consider the special case where the variance function is constant across all covariate values, i.e., $\sigma^2(x) \equiv \sigma^2$. Under this homoskedasticity assumption, the measurement error variance simplifies to:

$$\begin{aligned} V_E &= \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma^2 + \sum_{j\in\mathcal{C}} (w_j)^2\,\sigma^2\right) \\ &= \sigma^2\left(\frac{1}{n_T} + \frac{1}{\mathrm{ESS}(\mathcal{C})}\right), \end{aligned} \tag{5}$$

where $\mathrm{ESS}(\mathcal{C})$ is the effective sample size of the weighted control sample:

$$\mathrm{ESS}(\mathcal{C}) = \frac{\left(\sum_{i\in\mathcal{C}} w_i\right)^2}{\sum_{i\in\mathcal{C}} w_i^2}. \tag{6}$$

This metric quantifies the number of independent observations that would provide equivalent precision under equal weighting (Potthoff et al., 2024), and reflects efficiency loss from reusing controls with varying weights.

Based on this formula, our proposed plug-in estimator for V_E is:

$$\hat{V}_E = S^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right), \tag{7}$$

where S^2 is a pooled variance estimator for σ^2 defined across non-singleton matched clusters. Specifically:

$$S^2 = \frac{1}{N_C} \sum_{t \in \mathcal{T}_+} |\mathcal{C}_t| s_t^2 \quad \text{with} \quad N_C = \sum_{t \in \mathcal{T}_+} |\mathcal{C}_t|, \tag{8}$$

where $\mathcal{T}_+ = \{t \in \mathcal{T} : |\mathcal{C}_t| > 1\}$ excludes singleton clusters.

For each cluster, the residual variance is computed as:

$$s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left(Y_j - \bar{Y}_t \right)^2, \quad \text{where} \quad \bar{Y}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} Y_j. \tag{9}$$

We now establish that this estimator is consistent.

Lemma 4.1 (Consistency of the Pooled Variance Estimator)**.** *Let $\{\mathcal{C}_t, t \in \mathcal{T}\}$ be a collection of matched control sets. Assume Assumptions 4 (Shrinking Clusters), 5 (Regular Variance), and 6 (Derivative Control). Then, as $n_T \to \infty$:*

$$\left| S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right| \xrightarrow{a.s.} 0. \tag{10}$$

Proof: See Appendix B.

This result shows that even if variances are not constant across units, the pooled estimator S^2 consistently estimates the average variance across treated units. This aligns with the spirit of heteroskedasticity-robust variance estimation in White (1980), where consistency is achieved via aggregation even under variance heterogeneity.

Lemma 4.2 (Asymptotic Equivalence to Error Variance)**.** *Under the same assumptions as Lemma 4.1,*

define:

$$\hat{V}_{E,lim} := \left(\frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2\right)\left(\frac{1}{n_T} + \frac{1}{ESS(\mathcal{C})}\right).$$

Then:

$$\left|\hat{V}_{E,lim} - V_E\right| \xrightarrow{p} 0 \quad as\ n_T \to \infty.$$

Proof: See Appendix C.

Although Equation (7) was motivated under homoskedasticity, Lemma 4.2 shows that the same formula consistently estimates V_E even when variances are heterogeneous. This is because local variance estimates from matched clusters are close to the true $\sigma^2(X_t)$ due to shrinking clusters and regularity of the variance function. Aggregating over many clusters smooths out local errors—mirroring the robustness of White's heteroskedasticity-consistent variance estimator.

Theorem 4.3 (Consistency of the Variance Estimator)**.** *Under Assumptions 4, 5, and 6, the proposed estimator $\hat{V}_E$ is consistent:*

$$\left|\hat{V}_E - V_E\right| \xrightarrow{p} 0 \quad as\ n_T \to \infty.$$

Proof. From Lemma 4.1, we have $\left|S^2 - \frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2\right| \xrightarrow{a.s.} 0$. Substituting into our estimator formula:

$$\begin{aligned}
\hat{V}_E &= S^2\left(\frac{1}{n_T} + \frac{1}{\mathrm{ESS}(\mathcal{C})}\right) \\
&= \left(\frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2 + o_p(1)\right)\left(\frac{1}{n_T} + \frac{1}{\mathrm{ESS}(\mathcal{C})}\right) \\
&= \hat{V}_{E,\mathrm{lim}} + o_p(1)
\end{aligned}$$

By Lemma 4.2, we have $\left|\hat{V}_{E,\mathrm{lim}} - V_E\right| \xrightarrow{p} 0$. Therefore:

$$\begin{aligned}
\left|\hat{V}_E - V_E\right| &= \left|\hat{V}_{E,\mathrm{lim}} + o_p(1) - V_E\right| \\
&\le \left|\hat{V}_{E,\mathrm{lim}} - V_E\right| + |o_p(1)| \\
&\xrightarrow{p} 0
\end{aligned}$$

□

This theorem provides the key theoretical guarantee of our method: a plug-in variance estimator motivated by homoskedasticity remains consistent even under general heteroskedastic error structures, as long as the regularity conditions are met. Our estimator offers practical advantages in high-overlap settings, where reuse of control units inflates variance—an effect directly captured by the ESS term.

Our non-parametric approach differs from Theorem 1 of White (1980), which uses a regression-based (semi-parametric) method. While our matching procedure is governed by hyperparameters such as the number of neighbors or the maximum allowed radius, these parameters are not estimated from the data.Consequently, we require Assumption 5 (Regular Variance), especially the continuity condition in Definition 3.1, whereas White (1980) does not need such an assumption. While both proofs share the same overall strategy, but the specific technical details differ: White (1980)'s argument relies on compactness of the parameter space to bound the difference between the estimator and the truth, whereas we rely on Assumptions 4 (Shrinking Clusters) and 6 (Derivative Control). Further details on this comparison can be found in Appendix F.

4.2.2 A Consistent Estimator for V

Building on our analysis of the measurement error variance component V_E, we now develop a consistent estimator for the total variance V. While V_E captures the variance due to residual outcome noise, the complete variance V must also account for treatment effect heterogeneity among the treated units.

We start by exploring the relationship between the squared deviations of individual treatment effects and the components of the total variance:

$$\begin{aligned}
&E\left[\left(Y_t(1) - \hat{Y}_t(0) - \tau\right)^2\right] \\
\approx &E\left[(\tau(x) - \tau)^2\right] + E\left[\varepsilon_t^2 + \sum_{j \in C_t} w_{jt}^2 \varepsilon_j^2\right] \\
\approx &V_P + \frac{1}{n_T}\left[\sum_{t \in \mathcal{T}} \sigma_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2\right) \sigma_j^2\right]
\end{aligned}$$

This expectation can also be approximated empirically as:

$$E\left[\left(Y_t(1) - \hat{Y}_t(0) - \tau\right)^2\right] \approx \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2$$

By equating these expressions and rearranging terms, we can derive an estimator for V_P:

$$\begin{aligned}\hat{V}_P \approx &\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 \\ &- \frac{1}{n_T}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\hat{\sigma}_j^2\right]\end{aligned}$$

Combining this with our estimator for V_E, we obtain:

$$\begin{aligned}\hat{V} =& n_T\hat{V}_E + \hat{V}_P \\ =& \frac{1}{n_T}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2\hat{\sigma}_j^2\right] \\ &+ \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 \\ &- \frac{1}{n_T}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\hat{\sigma}_j^2\right]\end{aligned}$$

Through algebraic simplification, this expression reduces to:

$$\begin{aligned}\hat{V} =& \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 \\ &+ \hat{\sigma}^2\frac{1}{n_T}\left[\sum_{j\in\mathcal{C}}\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]\right]\end{aligned}$$

where $\hat{\sigma}^2$ is the pooled variance defined in Equation 8. This estimator effectively combines the empirical squared deviations with a correction term that accounts for the matching structure.

Theorem 4.4 (Consistency of the Total Variance Estimator)**.** *Under Assumptions 4, 5, and 6, the proposed estimator $\hat{V}$ is consistent:*

$$\left|\hat{V} - V\right| \xrightarrow{p} 0 \quad \text{as } n_T \to \infty.$$

The proof follows similar steps to those used in establishing the consistency of $\hat{V}_E$ in Theorem 4.3. The key insight is that both the empirical squared deviations and the correction term converge to their respective population counterparts in probability, leveraging the properties of shrinking clusters, regular error variance functions, and our derivative control condition.

This consistency result ensures that confidence intervals constructed using $\hat{V}$ will have asymptotically correct coverage, providing practitioners with reliable inference tools for matching estimators across a wide range of applications.

4.3 Comparison with Abadie and Imbens (2006)Estimator

To position our work within the existing literature and highlight its advantages, we now compare our variance estimator with that proposed by Abadie and Imbens (2006). This comparison is particularly relevant as their work established the foundational theory for matching estimators, and our analysis builds upon and extends their approach for practical applications in modern causal inference settings.

Adapting their estimator to our notation:

$$\widehat{V}_{AI06} = \frac{1}{n_T^2} \sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \frac{1}{n_T^2} \sum_{j \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} w_{jt} \right)^2 \hat{\sigma}_j^2, \tag{11}$$

where $w_{jt} = 1/M$ if unit j is among the M closest controls to unit t, and $w_{jt} = 0$ otherwise, and $\hat{\sigma}_i^2$ is an estimate of the conditional outcome variance for unit i, defined as:

$$\hat{\sigma}_i^2 = \frac{M}{M+1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{m(i)} \right)^2.$$

Here, $Y_{m(i)}$ denotes the outcome of the m-th closest unit to unit i among units with the same treatment status, and M is a fixed small number (typically set to match the number of matches used in the estimator).

The fundamental difference is in variance estimation: while Abadie and Imbens (2006) estimates variance by comparing each unit to its nearest same-treatment neighbors individually, our approach calculates variance within entire matched clusters. This cluster-based approach leverages

information from all control units matched to a treated unit simultaneously, resulting in more stable variance estimates, especially when match quality is heterogeneous.

To clarify the key methodological differences, let's compare the variance components directly. In Abadie and Imbens' approach, $\hat{\sigma}_i^2 = \frac{M}{M+1}\left(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{m(i)}\right)^2$ estimates variance by comparing each unit to its nearest same-treatment neighbors individually. By contrast, our estimator uses $s_t^2 = \frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left(Y_j - \bar{Y}_t\right)^2$, which calculates variance within entire matched clusters. Our approach effectively pools information across all controls matched to a treated unit, leveraging their collective variation to estimate outcome variability more robustly, particularly in settings with variable match quality.

This use of all controls in variance estimation provides several key advantages. First, our estimator requires matching only for treated units, whereas Abadie and Imbens (2006) requires matching for both treated and control units—significantly reducing computational burden when the control group is large. Second, Abadie and Imbens (2006)'s approach necessitates matching treated units with other treated units, which becomes problematic when the treated group is small or highly heterogeneous, as is common in many applications. Our approach is tailored specifically for ATT estimation in these realistic scenarios. Third, our framework naturally accommodates flexible weighting schemes, including kernel weights, caliper matching weights, and optimal transportation weights, whereas Abadie and Imbens (2006)'s approach was primarily designed for fixed-number nearest neighbor matching with equal weights.

While our estimator does not utilize within-treated-group variation for variance estimation (i.e., our approach does not use the observed outcomes Y_t of treated units when estimating variance), this potential efficiency loss is typically minor in ATT applications where the treated group is small relative to the control group. Furthermore, within-treated-group variation becomes unreliable when the number of treated units is small, making our approach more robust in such common scenarios. The limitation of not using treated outcomes for variance estimation is typically minor in ATT applications where the treated group is small relative to the control group—a common scenario in practice. Indeed, many influential ATT applications feature relatively small treated samples, including job training program evaluations (LaLonde, 1986), educational interventions (Abadie et al., 2002), and health policy assessments (Keele et al., 2023). Imbens (2004) notes that ATT estimation is often preferred precisely in contexts where treatment is relatively rare or targeted,

resulting in small treated groups compared to the potential control pool.

5 Simulation

In this section, we conduct simulation studies to validate the two main theoretical results established in earlier sections: Theorem 3.2 (Central Limit Theorem) and the consistency of our variance estimator. The primary focus is threefold: first, to verify the asymptotic normality of our estimator, second, to assess whether confidence intervals constructed using our variance estimator achieve near-nominal coverage, thereby confirming the accuracy of inference procedures built from it, and third, to compare the performance of our variance estimator to that of existing methods, showing how our approach outperforms the bootstrap variance estimator proposed by Otsu and Rai (2017). These simulations provide empirical insights into the reliability and robustness of our methods under different data-generating scenarios and matching conditions.

5.1 Simulation to Verify the Theoretical Results

In this subsection, we conduct a simulation study to verify the Central Limit Theorem (CLT) result (Theorem 3.2) and demonstrate the consistency of our variance estimator, as proper inference depends critically on accurate variance estimation. Following the simulation setup from Che et al. (2024), we generate treated and control units from bivariate distributions with varying degrees of overlap and measure the coverage of the resulting confidence intervals. The complete details of the data-generation process, including specific distribution parameters and matching procedures, are provided in Appendix G.1.

We first compute the coverage of a 95% confidence interval using the theoretically correct variance V by constucting $(\hat{\tau} - B_n) \pm \sqrt{\frac{V}{n_T}}$ over 500 replications. The results are presented in Table 1. The coverage rates for varying degrees of overlap range from 94% to 95.2%, demonstrating that the nominal 95% level is closely attained. We also plot the distribution of $\frac{\sqrt{n_T}\left(\hat{\tau} - B_n - \tau\right)}{V^{-1/2}}$ against a standard normal distribution in Figure 1. We observed a close alignment between these distributions. This provides intuitive evidence that our estimator's distribution converges to the theoretical limit.

We then repeat the coverage analysis using an estimated variance $\widehat{V}$ rather than the true variance

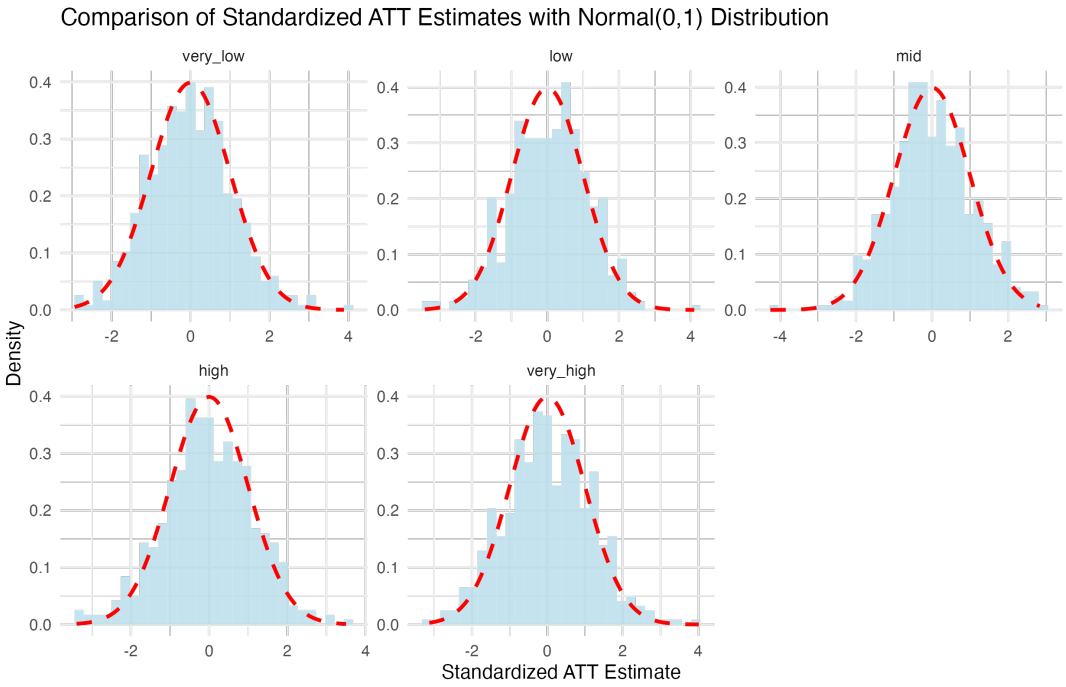

Figure 1: Empirical distribution of $\frac{\sqrt{n_T}\left(\hat{\tau}-B_n-\tau\right)}{V^{-1/2}}$ versus the standard normal.

V. The results remain consistently close to 95% across all overlap levels, demonstrating that our variance estimation method is accurate even under varying degrees of overlap. This is an indication of robustness of the variance estimator.

Table 1: Coverage of 95% Confidence Intervals Using True Variance V and Estimated Variance $\widehat{V}$

Degree of Overlap	**Coverage (%) with V**	**Coverage (%) with $\widehat{V}$**
very low	94.0	95.0
low	95.0	94.0
mid	95.2	96.0
high	94.0	95.0
very high	95.0	94.0

These findings confirm that the CLT is well-supported empirically across various overlap scenarios. Moreover, our proposed variance estimator demonstrates reliable coverage properties in practice. The consistently strong performance shown in Table 1 is particularly noteworthy given the results that follow in the next subsection. In the following comparison with alternative methods, we will demonstrate that while our variance estimator maintains reliable coverage across different degrees of overlap, the bootstrap method struggles substantially in certain scenarios, highlighting

the practical superiority of our approach in realistic matching conditions.

5.2 A comparison to the bootstrap variance estimator

After validating the CLT, we now compare our variance estimator with the bootstrap variance estimator. To evaluate our proposed methods, we replicate a specific case from Otsu and Rai (2017), focusing on a two-dimensional setting with a complex nonlinear outcome function. The data generating process consists of a treatment assignment mechanism governed by parameters $\gamma_1 = 0.15$ and $\gamma_2 = 0.7$, with treatment probability determined by $P(X_i) = \gamma_1 + \gamma_2 \|X_i\|$. The outcome follows a nonlinear pattern defined by $m(z) = 0.4 + 0.25\sin(8z-5) + 0.4\exp(-16(4z-2.5)^2)$, where potential outcomes are generated as $Y_i(1) = \tau + m(\|X_i\|) + \epsilon_i$ and $Y_i(0) = m(\|X_i\|) + \epsilon_i$, with $\epsilon_i \sim N(0, 0.2^2)$ and true treatment effect $\tau = 0$. Full simulation setting is provided in Appendix G.2.

Our analysis employs 8-nearest neighbor matching with uniform weights of 1/8 assigned to each matched control unit. As reported in the paper, the true 95% confidence interval (CI) coverage is 0.9473, with an average CI length of 0.2381. Through a comparison of our pooled variance estimator against the wild bootstrap method (see Appendix G.2 for details on the wild bootstrap implementation) proposed by Otsu and Rai (2017) across 100 replications, we find differences in performance and present in Table 2. The wild bootstrap method achieves only 61% coverage with an average confidence interval length of 0.1561, which is substantially shorter than the true interval length of 0.2381. In contrast, our method maintains coverage closer to the nominal rate.

Method	Coverage (%)	Average CI Length	Difference from True CI Length
Wild Bootstrap	61.00	0.1561	-0.0820
Our Method	97.00	0.3127	+0.0746

Table 2: Comparison of confidence interval performance across methods. The Wild Bootstrap method has a lower coverage and shorter intervals compared to the true values, while the A-E Bootstrap method achieves coverage closer to the true value but at the cost of longer intervals.

The poor performance of the wild bootstrap can be attributed to the substantial overlap in the matching structure. With an average of approximately 50 units in each of the treatment and control groups, we observe controls being shared across an average of 92 treated-control pairs out of approximately 400 total pairs (roughly 25%), indicating extensive reuse of control units in our two-dimensional matching setting. Our analysis shows consistent patterns across different inference

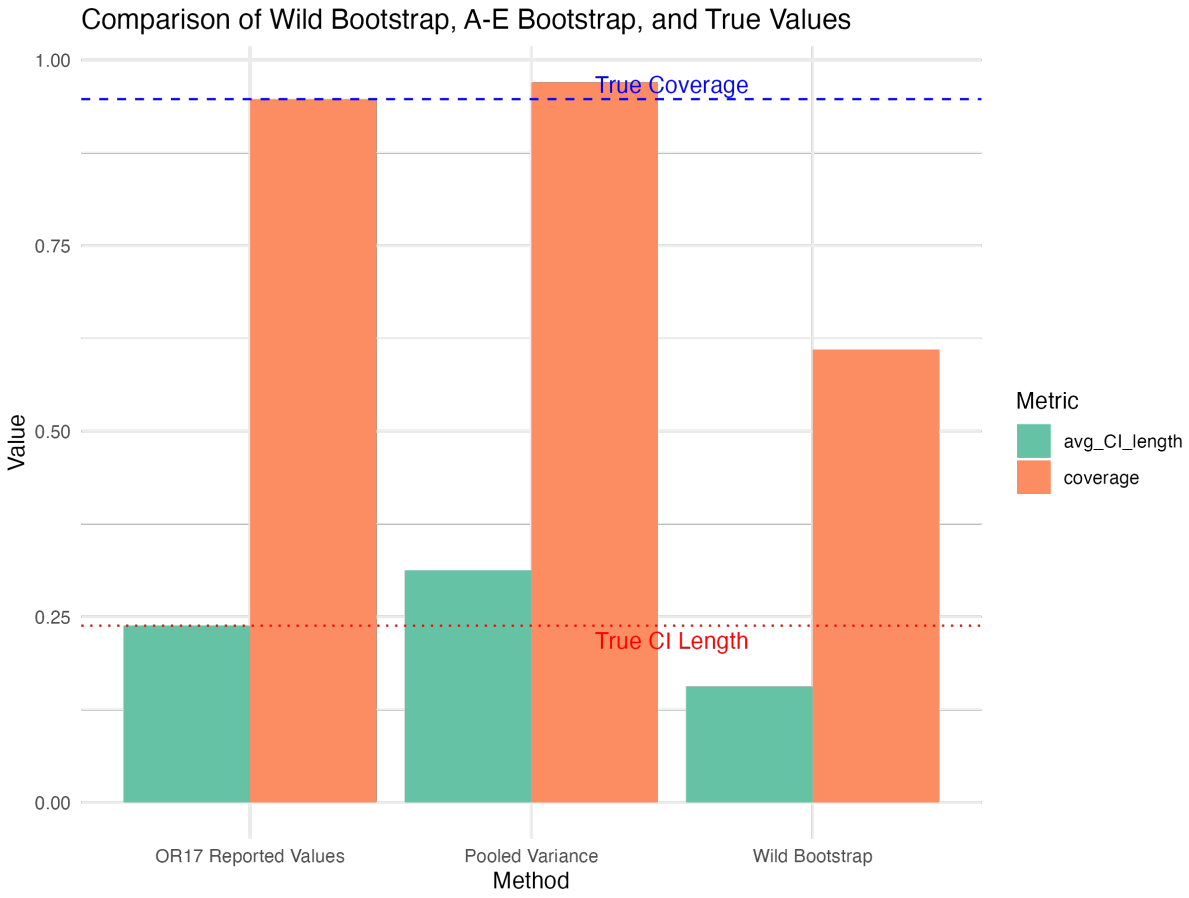

Figure 2: Comparison of the Wild Bootstrap, A-E Bootstrap, and true values for confidence interval (CI) coverage and average CI length. The Wild Bootstrap method achieves shorter CIs but at the cost of lower coverage, while the A-E Bootstrap method closely approximates the true coverage but results in longer intervals. Horizontal dashed and dotted lines indicate the true coverage and CI length, respectively, for reference.

methods, with an average of 23 shared controls per treated unit (specifically, 23.17 and 23.45 for wild bootstrap and our method respectively), while the average number of shared treated units per control is approximately 92 (92.63 and 92.23 respectively). This high degree of overlap is not unexpected given that the numbers of treated (n_T) and control (n_C) units are of similar magnitude. To investigate this relationship further, we conducted an additional experiment where we reduced the overlap by setting $N_T = 25$ and $N_C = 1000$. Under these conditions, the wild bootstrap recovers its nominal coverage, while our pooled variance estimator continues to perform well. These findings support our hypothesis that the wild bootstrap procedure is asymptotically valid primarily in settings with minimal overlap.

6 Application of the Variance Estimator to Other Estimators

While our variance estimator was developed in the context of matching methods, its utility extends to other classes of estimators that share similar properties. In this section, we demonstrate how our approach can be applied to weighting estimators, specifically the stable balancing weights method proposed by Zubizarreta (2015).

6.1 Stable Balancing Weights Estimator

The stable balancing weights approach of Zubizarreta (2015) finds weights that minimize the variance of the weighted estimator while satisfying covariate balance constraints. Using our notation, the stable balancing weights estimator for the ATT can be expressed as:

$$\hat{\tau}_{SBW} = \frac{1}{n_T} \sum_{t \in \mathcal{T}} Y_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j Y_j \tag{12}$$

where w_j are weights assigned to control units that minimize $\sum_{j \in \mathcal{C}} w_j^2$ subject to balance constraints of the form $\left| \sum_{t \in \mathcal{T}} \frac{1}{n_T} X_t - \sum_{j \in \mathcal{C}} w_j X_j \right| \leq \delta$ for some small tolerance δ.

The variance estimator for the stable balancing weights approach directly extends our framework. We can apply our estimator from Equation 7 with only one modification: while the weights

w_j are determined through quadratic optimization rather than matching, the construction of the heteroskedastic variance component S^2 still requires forming local neighborhoods through matching. This hybrid approach leverages the computational advantages of both techniques—optimal weights from the balancing procedure and accurate variance estimation from local matching—resulting in valid inference for the weighting estimator.

6.2 Simulation Study: Kang and Schafer (2007) DGP

To evaluate the performance of our variance estimator when applied to the stable balancing weights estimator, we conducted a simulation study using the data generating process (DGP) proposed by Kang and Schafer (2007). This DGP is widely used in the causal inference literature as a challenging benchmark due to its non-linear relationships between covariates, treatment, and outcomes.

The Kang and Schafer DGP generates four standard normal covariates (X_1, X_2, X_3, X_4) and then creates non-linear transformations to produce observed covariates. The treatment assignment is a function of these covariates, and the outcome model includes interactions between treatment and covariates, creating a complex setting where many estimators struggle to achieve proper coverage. Full mathematical details of this DGP are provided in Appendix G.3.

6.3 Results and Discussion

We applied the stable balancing weights estimator with our variance estimation approach to the Kang and Schafer DGP over 100 independent replications. The 95% confidence intervals constructed using our variance estimator achieved a coverage rate of 98%. This slightly conservative coverage indicates that our variance estimator remains valid even when applied to weighting estimators in challenging settings.

The simulation results demonstrate that our variance estimation framework has broader applicability beyond matching estimators. The slightly higher-than-nominal coverage (98% versus the target 95%) can be attributed to the steep gradient in the Kang and Schafer outcome model, which creates larger effective bias terms that are not fully accounted for in the first-order approximation.

This slight over-coverage suggests directions for future research, particularly in developing refined variance estimators that can better account for steep derivatives in the outcome model. Po-

tential approaches could include higher-order bias corrections or adaptive methods that estimate the curvature of the outcome model.

Nevertheless, the strong performance of our variance estimator when applied to the stable balancing weights approach demonstrates its flexibility and robustness. This extension opens possibilities for creating a unified framework for inference across various classes of weighting and matching estimators in causal inference.

7 Conclusion

This paper develops new methods for statistical inference in matching estimators, addressing key challenges in both theoretical foundations and practical implementation. We propose a novel variance estimator that remains valid under extensive control unit reuse and is computationally more efficient than existing approaches. Our theoretical framework generalizes the standard Lipschitz continuity assumption, allowing for broader applicability in settings where the derivative grows moderately but not uniformly. Additionally, we resolve the ongoing debate about bootstrap validity for matching estimators by carefully analyzing the conditions under which different bootstrap procedures succeed or fail.

Through extensive simulation studies, we demonstrate that our variance estimator achieves more reliable inference than existing methods do, particularly in settings with substantial overlap between matched samples. Our findings show that the widely used Wild bootstrap approach underperforms in such scenarios, leading to underestimated standard errors and invalid confidence intervals. In contrast, our proposed method maintains accurate coverage even in high-overlap settings, reinforcing its robustness and practical applicability.

Empirical applications suggest that our approach provides a practical and theoretically grounded solution for researchers conducting inference in matched designs, with potential applications in economics, epidemiology, and policy evaluation. Future work will explore extensions to high-dimensional covariate spaces, alternative bias correction techniques, and adaptations for time-varying treatment effects in longitudinal matching designs. We also aim to examine alternative resampling techniques that better account for matched-set dependencies while maintaining computational efficiency.

Our contributions advance the state of the art in matching-based inference by providing a computationally efficient, theoretically justified, and empirically validated approach to variance estimation. These results offer important guidance for practitioners using matching methods and highlight the need for careful consideration of variance estimation strategies in applied settings.

7.1 Future Work

Our research opens several promising avenues for further investigation. First, while our simulation evidence demonstrates the superiority of our approach's finite-sample performance to that of existing methods, particularly in settings with extensive control unit reuse, a more rigorous theoretical comparison would enhance our understanding of when and why different inference procedures outperform others. Such theoretical analysis could formally characterize the efficiency gains of our approach relative to alternatives under various data-generating processes and matching configurations.

Second, it is worth exploring improvements to bootstrap procedures specifically designed for matching estimators. For instance, adapting moving block bootstrap procedures like those described in Lahiri (2003) could better handle the complex dependency structure induced by overlapping matches. This approach might preserve the intuitive appeal of bootstrap methods while addressing their current limitations in the matching context. By explicitly modeling the correlation structure among matched units, such modified bootstrap procedures could potentially achieve both valid inference and computational efficiency.

References

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

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

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

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

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, 70(1):91–117, 2002.

Jonathan Che, Xiang Meng, and Luke Miratrix. Caliper synthetic matching: Generalized radius matching with local synthetic controls. arXiv preprint arXiv:2411.05246, 2024.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448): 1053–1062, 1999.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis, 15(3):965–1056, 2020.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies, 64(4):605–654, 1997.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. The Review of Economic Studies, 65(2):261–294, 1998.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86(1):4–29, 2004.

Nathan Kallus. Generalized optimal matching methods for causal inference. J. Mach. Learn. Res., 21:62–1, 2020.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. 2007.

Luke J. Keele, Eli Ben-Michael, Avi Feller, Rachel Kelz, and Luke Miratrix. Hospital quality risk standardization via approximate balancing weights. The Annals of Applied Statistics, 17(2), June 2023. ISSN 1932-6157. doi: 10.1214/22-AOAS1629. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-2/Hospital-quality-risk-standardization-via-approximate-balancing-weights/10.1214/22-AOAS1629.full.

S. N. Lahiri. Resampling Methods for Dependent Data. Springer Series in Statistics. Springer New York, New York, NY, 2003. ISBN 978-1-4419-1848-2 978-1-4757-3803-2. doi: 10.1007/978-1-4757-3803-2. URL http://link.springer.com/10.1007/978-1-4757-3803-2.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. The American economic review, pages 604–620, 1986.

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.

Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. The Annals of Statistics, 22(4):2031–2050, 1994.

Richard F Potthoff, Max A Woodbury, and Kenneth G Manton. "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. 2024.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Donald B Rubin. Matching to remove bias in observational studies. Biometrics, pages 159–183, 1973.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1-2):305–353, 2005.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1):1–21, 2010.

Yixin Wang and Jose R Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. Biometrika, page asz050, October 2019. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asz050. URL https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asz050/5602475.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: journal of the Econometric Society, pages 817–838, 1980.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511):910–922, 2015.

Appendix

A Proof of Theorem 3.2

Note that

$$\sqrt{n_T}\,(\hat{\tau} - B_n - \tau) = \underbrace{\sqrt{n_T}(\tau_{SATT} - \tau)}_{\text{Population error}} + \underbrace{\sqrt{n_T}E_n}_{\text{measurement error}}$$

where $E_n = \left(\frac{1}{n_T}\sum_{t\in\mathcal{T}}\epsilon_t - \frac{1}{n_T}\sum_{j\in\mathcal{C}} w_j\epsilon_j\right)$

We focus on the population error and the measurement error separately.

First, we focus on the population error. By the Central Limit Theorem (CLT):

$$\begin{aligned}
&\sqrt{n_T}(\tau_{SATT} - \tau) \\
&= \sqrt{n_T}\left(\frac{1}{n_T}\sum_{t\in\mathcal{T}}\big(f_1(X_t) - f_0(X_t)\big) - \mathbb{E}_{X|Z=1}\big[f_1(X_i) - f_0(X_i) \mid Z_i = 1\big]\right) \\
&\xrightarrow{d} \mathcal{N}(0, V_P)
\end{aligned}$$

where

$$V_P = \mathbb{E}_{X|Z=1}\big[(f_1(X_i) - f_0(X_i) - \tau)^2 \mid Z_i = 1\big] \tag{13}$$

Second, we focus on the measurement error part. We want to establish the asymptotic normality:

$$\frac{E_n}{\sqrt{V_E}} \xrightarrow{d} N(0,1) \tag{14}$$

This is equivalent to showing that $\sqrt{n_T}E_n \xrightarrow{d} N(0, V_E)$ conditional on $\mathbf{X}, \mathbf{Z}$.

Let us denote $e_i = (Z_i - (1 - Z_i)W_i)\epsilon_i$ as the weighted error contribution of the i-th unit, where

$W_i = \sum_{t\in\mathcal{T}} w_{it}$ when $Z_i = 0$ and $W_i = 0$ otherwise. Then:

$$\frac{E_n}{\sqrt{V_E}} = \frac{1}{n_T\sqrt{V_E}} \sum_{i=1}^{n} e_i \tag{15}$$

To establish the asymptotic normality, we verify the Lindeberg condition for triangular arrays, for all $\epsilon > 0$:

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[T_{n,i}^2 \mathbf{1}(|T_{n,i}| > \epsilon s_n) \mid \mathbf{X}, \mathbf{Z}\right] \to 0 \text{ as } n \to \infty \tag{16}$$

where

$$\begin{aligned} T_{n,i} &= \frac{e_i}{n_T\sqrt{V_E}} \\ \mathbb{E}[T_{n,i}^2 \mid \mathbf{X}, \mathbf{Z}] &= \frac{\mathbb{E}[e_i^2 \mid \mathbf{X}, \mathbf{Z}]}{n_T^2 V_E} \\ &= \frac{(Z_i - (1 - Z_i)W_i)^2\sigma_i^2}{n_T^2 V_E} \end{aligned}$$

and $s_n^2 = \sum_{i=1}^{n} \mathbb{E}[T_{n,i}^2 \mid \mathbf{X}, \mathbf{Z}] = 1$ [1]

Hence,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[T_{n,i}^2 \mathbf{1}(|T_{n,i}| > \epsilon s_n) \mid \mathbf{X}, \mathbf{Z}\right] \tag{17}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{e_i}{n_T\sqrt{V_E}}\right)^2 \mathbf{1}\left(\left|\frac{e_i}{n_T\sqrt{V_E}}\right| > \epsilon\right) \,\middle|\, \mathbf{X}, \mathbf{Z}\right] \tag{18}$$

$$= \frac{1}{n_T^2 V_E} \sum_{i=1}^{n} \mathbb{E}\left[e_i^2 \cdot \mathbf{1}(|e_i| > \epsilon n_T\sqrt{V_E}) \,\middle|\, \mathbf{X}, \mathbf{Z}\right] \tag{19}$$

Focusing on the i-th summand and applying Hölder's inequality with conjugate exponents $\frac{2+\delta}{2}$

[1] Recall that $V_E = \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma_t^2 + \sum_{j\in\mathcal{C}} (w_j)^2\sigma_j^2\right)$

and $\frac{2+\delta}{\delta}$:

$$
\begin{aligned}
&\mathbb{E}\left[e_i^2 \cdot \mathbf{1}(|e_i| > \epsilon n_T \sqrt{V_E}) \;\middle|\; \mathbf{X}, \mathbf{Z}\right] \\
&\leq \mathbb{E}\left[|e_i|^{2+\delta} \;\middle|\; \mathbf{X}, \mathbf{Z}\right]^{\frac{2}{2+\delta}} \cdot \mathbb{E}\left[\mathbf{1}(|e_i| > \epsilon n_T \sqrt{V_E}) \;\middle|\; \mathbf{X}, \mathbf{Z}\right]^{\frac{\delta}{2+\delta}} \\
&= \mathbb{E}\left[|e_i|^{2+\delta} \;\middle|\; \mathbf{X}, \mathbf{Z}\right]^{\frac{2}{2+\delta}} \cdot \mathbb{P}\left[|e_i| > \epsilon n_T \sqrt{V_E} \;\middle|\; \mathbf{X}, \mathbf{Z}\right]^{\frac{\delta}{2+\delta}} \\
&\leq \mathbb{E}\left[|e_i|^{2+\delta} \,\middle|\, \mathbf{X}, \mathbf{Z}\right]^{\frac{2}{2+\delta}} \cdot \left(\frac{\mathbb{E}\left[e_i^2 \,\middle|\, \mathbf{X}, \mathbf{Z}\right]}{\epsilon^2 n_T^2 V_E}\right)^{\frac{\delta}{2+\delta}} \qquad \text{by Markov's inequality} \\
&= (Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{\mathbb{E}\left[|\epsilon_i|^{2+\delta} \;\middle|\; \mathbf{X}, \mathbf{Z}\right]^{\frac{2}{2+\delta}} \cdot \sigma_i^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (n_T^2 V_E)^{\frac{\delta}{2+\delta}}} \\
&\leq (Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (\sum_{i=1}^2 (Z_i - (1-Z_i)W_i)^2 \sigma_{min}^2)^{\frac{\delta}{2+\delta}}}
\end{aligned}
$$

In the last step, we use the bounds from our assumptions: $\sigma_{min}^2 \leq \sigma_i^2 \leq \sigma_{max}^2$ from Assumption **??**, and the bound $\mathbb{E}\left[\left|\epsilon_i\right|^{2+\delta} \;\middle|\; X_i = x\right] \leq C$ from Assumption **??**.

Hence, the Lindeberg condition in Equation 19 is upper bounded by

$$
\begin{aligned}
&\frac{1}{n_T^2 V_E} \sum_{i=1}^n \left[(Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (\sum_{i=1}^n (Z_i - (1-Z_i)W_i)^2 \sigma_{\min}^2)^{\frac{\delta}{2+\delta}}}\right] && (20) \\
\leq &\frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot \sigma_{\min}^2 \cdot (\sum_{i=1}^n (Z_i - (1-Z_i)W_i)^2)^{\frac{\delta}{2+\delta}}} \sum_{i=1}^n \left[(Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}}\right] && (21) \\
= &\frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \sigma_{\min}^{\frac{4+6\delta}{2+\delta}}} \cdot \frac{\sum_{i=1}^n \left[(Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}}\right]}{(\sum_{i=1}^n (Z_i - (1-Z_i)W_i)^2)^{\frac{2+3\delta}{2+\delta}}} && (22) \\
= &\left(\frac{1}{n}\right)^{\frac{\delta}{2+\delta}} \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \sigma_{\min}^{\frac{4+6\delta}{2+\delta}}} \cdot \frac{\frac{1}{n}\sum_{i=1}^n \left[(Z_i - (1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}}\right]}{\left(\frac{1}{n}\sum_{i=1}^n (Z_i - (1-Z_i)W_i)^2\right)^{\frac{2+3\delta}{2+\delta}}} && (23)
\end{aligned}
$$

The term $\frac{\frac{1}{n}\sum_{i=1}^n \left[(Z_i-(1-Z_i)W_i)^{2+\frac{2\delta}{2+\delta}}\right]}{\left(\frac{1}{n}\sum_{i=1}^n (Z_i-(1-Z_i)W_i)^2\right)^{\frac{2+3\delta}{2+\delta}}}$ is bounded in probability by Markov's inequality and because all moments of W_i are bounded according to Lemma 3(i) in Abadie and Imbens (2006). Therefore, Equation 23 converges to 0 as $n \to \infty$ since $\left(\frac{1}{n}\right)^{\frac{\delta}{2+\delta}} \to 0$. Thus, the Lindeberg condition is satisfied, establishing asymptotic normality.

Finally, $\sqrt{n_T}P_n = \sqrt{n_T}(\tau_{SATT} - \tau)$ and $\sqrt{n_T}E_n$ are asymptotically independent, as the central limit theorem for $\sqrt{n_T}E_n$ holds conditional on the covariates $\mathbf{X}$ and treatment assignment $\mathbf{Z}$. Since both terms converge to normal distributions and given that V_E is bounded and bounded away from zero by Assumption **??**, while V_P remains bounded by the properties of the treatment effect function under Assumption **??**, we can conclude that

$$\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\sqrt{V_E + V_P}} \xrightarrow{d} N(0,1) \tag{24}$$

This establishes the asymptotic normality of our estimator after accounting for the bias term.

B Proof of Lemma 4.1

Proof. Let us decompose the difference between our variance estimator and the true average variance:

$$\begin{aligned}
S^2 - \frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2 &= \frac{1}{N_C}\sum_{t\in\mathcal{T}}|\mathcal{C}_t|s_t^2 - \frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2 \\
&= \sum_{t\in\mathcal{T}} u_t s_t^2 - \frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2 \\
&= \sum_{t\in\mathcal{T}}(u_t s_t^2 - \frac{1}{n_T}\sigma_t^2) \\
&= \sum_{t\in\mathcal{T}}\underbrace{(u_t s_t^2 - u_t\sigma_t^2)}_{\text{Term A}} + \sum_{t\in\mathcal{T}}\underbrace{(u_t\sigma_t^2 - \frac{1}{n_T}\sigma_t^2)}_{\text{Term B}}
\end{aligned}$$

where $u_t = \frac{|\mathcal{C}_t|}{N_C}$ represents the weight of cluster t in the pooled estimator. Note that

$$N_C = \sum_{t\in\mathcal{T}}|\mathcal{C}_t| \tag{25}$$

is the total number of matches[2].

We first analyze Term A, which measures the difference between the estimated and true variance

[2]If a control unit is matched to multiple treated units, it contributes to N_C multiple times. For example, if a control unit is matched to three treated units, it adds 3 to N_C rather than 1.

within each cluster. For a fixed treatment t, for each individual matched control j in $\mathcal{C}_t$, we focus on the summand in $s_t^2 = \frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}(Y_j - \bar{Y}_t)^2$ (introduced in Equation 9) and expand the squared deviation:

$$\begin{aligned}
(Y_j - \bar{Y}_t)^2 &= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k) + \epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right)^2 \\
&= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)^2 \\
&\quad + 2\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)\left(\epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right) \\
&\quad + \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right)^2
\end{aligned}$$

Therefore, the difference between the sample variance and the true variance can be written as:

$$\begin{aligned}
s_t^2 - \sigma_t^2 &= \frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}(Y_j - \bar{Y}_t)^2 - \sigma_t^2 \\
&= \underbrace{\left(\frac{1}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}\epsilon_j^2 - \sigma_t^2\right)}_{\text{Sampling error}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[-2\epsilon_j\left(\frac{1}{|\mathcal{C}_t|}\sum_{\substack{k\in\mathcal{C}_t \\ k\neq j}}\epsilon_k\right)\right]}_{\text{Cross-product of errors}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[2\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)\left(\epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right)\right]}_{\text{Interaction between function and errors}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)^2\right]}_{\text{Systematic differences within cluster}}
\end{aligned}$$

□

Now, Term A becomes the following decomposition:

$$
\begin{aligned}
\text{Term A} &= \sum_{t=1}^{n_T}(u_t s_t^2 - u_t\sigma_t^2) \quad \text{where } u_t = \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T}|\mathcal{C}_t|} \\
&= \underbrace{\sum_{t=1}^{n_T}\frac{u_t}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}(\varepsilon_j^2 - \sigma_t^2)}_{\text{Sampling error}} \\
&+ \underbrace{\sum_{t=1}^{n_T}\frac{u_t}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}\left[-2\varepsilon_j\left(\frac{1}{|\mathcal{C}_t|}\sum_{\substack{k\in\mathcal{C}_t\\k\neq j}}\varepsilon_k\right)\right]}_{\text{Cross-product of errors}} \\
&+ \underbrace{\sum_{t=1}^{n_T}\frac{u_t}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[-2\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}f_0(X_k)\right)\left(\varepsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right)\right]}_{\text{Interaction between function and errors}} \\
&+ \underbrace{\sum_{t=1}^{n_T}\frac{u_t}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}f_0(X_k)\right)^2\right]}_{\text{Systematic differences within cluster}}
\end{aligned}
$$

Let's focus on the first component of Term A, the sampling error:

$$
\begin{aligned}
&\sum_{t=1}^{n_1}\frac{u_t}{|C_t|}\sum_{j\in Ct}\left(\varepsilon_j^2 - \sigma_t^2\right) \\
=&\sum_{c=1}^{n_c}\sum_{t\in T_c}\frac{1}{\sum_{c=1}^{n_c}K(c)}\left(\varepsilon_c^2 - \sigma_t^2\right) \\
=&\sum_{c=1}^{n_c}\sum_{t\in T_c}\frac{1}{\sum_{c=1}^{n_c}K(c)}\left(\varepsilon_c^2 - \sigma_c^2 + \sigma_c^2 - \sigma_t^2\right) \\
=&\frac{1}{\sum_{c=1}^{n_c}K(c)}\sum_{c=1}^{n_c}K(c)\left(\varepsilon_c^2 - \sigma_c^2\right) + \frac{1}{\sum_{c=1}^{n_c}K(c)}\sum_{c=1}^{n_c}\sum_{t\in T_c}\left(\sigma_c^2 - \sigma_t^2\right)
\end{aligned}
$$

where $K(c)$ represents the number of times control unit c is used across all matches. Note that $\sum_{c=1}^{n_C}K(c) = \sum_{t=1}^{n_T}|\mathcal{C}_t| = N_C$ is the total number of matches (Equation 25).

For the first term, we can apply Hölder's inequality and the law of large numbers to show it converges to zero in probability:

Applying Hölder's inequality with conjugate exponents p and q where $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$:

$$\begin{aligned}
\left| \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\varepsilon_c^2 - \sigma_c^2) \right| &\leq \frac{1}{\sum_{c=1}^{n_c} K(c)} \left(\sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \\
&= \frac{n_c^{1/p+1/q}}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \\
&= \frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}}
\end{aligned}$$

Now, we analyze each component:

1. For the ratio $\frac{n_c}{\sum_{c=1}^{n_c} K(c)}$: By the Law of Large Numbers (LLN), $\frac{\sum_{c=1}^{n_c} K(c)}{n_c} \xrightarrow{p} E[K(c)]$ as $n_c \to \infty$. Since $K(c) \geq 1$ for all c (each control unit is matched at least once), we have $E[K(c)] \geq 1$. Therefore: $\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \xrightarrow{p} \frac{1}{E[K(c)]} < \infty$

2. For the term $\left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}}$: By Lemma 3 of Abadie and Imbens (2006), $\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \xrightarrow{p} E[K(c)^p] < \infty$ for $p > 1$. The lemma applies because the matching process ensures that $K(c)$ has finite moments of all orders under standard matching schemes.

3. For the term $\left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}}$: We need to verify that $E[|\varepsilon_c^2 - \sigma_c^2|^q] < \infty$ to apply the LLN. By Assumption **??**, $E[|\varepsilon_c|^{2+\delta} | X_c] \leq C$ for some $\delta > 0$ and constant C. For $q \leq 1 + \frac{\delta}{2}$, we can show that $E[|\varepsilon_c^2 - \sigma_c^2|^q] < \infty$ using Jensen's inequality and the fact that $\sigma_c^2 = E[\varepsilon_c^2 | X_c]$. By the LLN, $\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \xrightarrow{p} E[|\varepsilon_c^2 - \sigma_c^2|^q]$. Since $\varepsilon_c^2 - \sigma_c^2$ is a mean-zero random variable (by definition of σ_c^2), with $q > 1$, the expectation $E[|\varepsilon_c^2 - \sigma_c^2|^q]$ is strictly positive but finite.

Combining these results:

$$\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \xrightarrow{p} \frac{1}{E[K(c)]} \cdot E[K(c)^p]^{1/p} \cdot E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \tag{26}$$

By selecting q sufficiently close to 1 (and thus p sufficiently large), and applying Jensen's in-

equality:

$$E[K(c)^p]^{1/p} \leq (E[K(c)]^p)^{1/p} \cdot O(1) \tag{27}$$

$$= E[K(c)] \cdot O(1) \tag{28}$$

Therefore, as $n_c \to \infty$:

$$\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \xrightarrow{p} O(1) \cdot E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \tag{29}$$

As $n_c \to \infty$, $E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \to 0$ for properly chosen q. This establishes that $\frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\varepsilon_c^2 - \sigma_c^2) \xrightarrow{p} 0$.

For the second term:

$$\frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} \sum_{t \in T_c} (\sigma_c^2 - \sigma_t^2) = \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \tag{30}$$

where $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t \in T_c} \sigma_t^2$ is the average variance of the treated units matched to control unit c, and $K(c) = |T_c|$ represents the number of treated units to which control unit c is matched.

We can bound this term as follows:

$$\left| \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \right| \leq \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c) \cdot \max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \tag{31}$$

$$= \max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{\text{a.s.}} 0 \text{ as } n_c, n_T \to \infty \tag{32}$$

where the last convergence follows from Lemma D.2, which establishes the uniform convergence of variance differences across all control units.

For the second component of Term A (cross-product of errors):

$$\begin{aligned}
&\sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right) \right] \\
&= \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right] \\
&= \sum_{t=1}^{n_T} \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \frac{1}{|\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C}_t \\ j \neq k}} (-4\varepsilon_j \varepsilon_k) \\
&\leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} -4 \cdot \frac{K(j,k)}{2} \varepsilon_j \varepsilon_k \\
&\leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} -2 \cdot K(j,k) \varepsilon_j \varepsilon_k
\end{aligned}$$

where $K(j,k)$ represents the number of times control units j and k appear together in the same matched cluster. Since $|\mathcal{C}_t| \geq 2$ for all clusters (as we exclude singleton clusters), we have $\frac{1}{|\mathcal{C}_t|} \leq \frac{1}{2}$, which gives us the inequality in the last step.

To establish that $(A2) \xrightarrow{p} 0$, we apply the same Hölder's inequality approach as used for the sampling error term. Specifically, we can bound:

$$\left| \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} K(j,k) \varepsilon_j \varepsilon_k \right| \leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \left(\sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} K(j,k)^p \right)^{\frac{1}{p}} \left(\sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} |\varepsilon_j \varepsilon_k|^q \right)^{\frac{1}{q}}$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$. Under our matching schemes, $K(j,k)$ is asymptotically sparse—the probability that two specific units j and k are repeatedly matched together diminishes as n_T increases. Using the technique developed in our analysis of the sampling error term and the properties of $K(j,k)$ (which is bounded by $\min\{K(j), K(k)\}$), we can show that the first factor converges to zero in probability.

Additionally, while $\mathbb{E}[\varepsilon_j \varepsilon_k] = 0$ due to independence and zero mean of the errors, we need to be careful about the higher moments in the Hölder's inequality. When $q > 1$, $\mathbb{E}[|\varepsilon_j \varepsilon_k|^q] = \mathbb{E}[|\varepsilon_j|^q] \cdot \mathbb{E}[|\varepsilon_k|^q]$ is bounded but generally non-zero. However, $(A2) \xrightarrow{p} 0$ primarily because $\mathbb{E}[K(j,k)^p] \to 0$ as $n_T \to \infty$ under our asymptotically sparse matching scheme, while $\mathbb{E}[|\varepsilon_j|^q \cdot |\varepsilon_k|^q]$ remains bounded by Assumption **??**. Therefore, $(A2) \xrightarrow{p} 0$ as $n_T \to \infty$.

For the third component of Term A (interaction between function values and errors):

$$(A3) = \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right) \right]$$

By the Mean Value Theorem and Assumption 6, we can bound the first factor:

$$\begin{aligned} \left| f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right| &\leq \max_{k \in \mathcal{C}_t} |f_0(X_j) - f_0(X_k)| \\ &\leq \sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot \max_{j,k \in \mathcal{C}_t} \|X_j - X_k\| \\ &\leq \sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \end{aligned}$$

Therefore:

$$\begin{aligned} |(A3)| &\leq \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} 2 \cdot \sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \cdot \left| \varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \\ &\leq 2 \cdot \sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right] \cdot \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left| \varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \\ &= 2 \cdot \sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right] \cdot \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{c=1}^{n_C} K(c) \left| \varepsilon_c - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \end{aligned}$$

By Assumption 6, the first term $\sup_{t \in \mathcal{T}} [\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t)]$ is bounded by a constant M.

For the remaining term, we can apply a similar Hölder's inequality argument as developed for the sampling error term earlier. The structure involves products of $K(c)$ with error differences, which have the same statistical properties (independence, mean zero) as the $\varepsilon_c^2 - \sigma_c^2$ terms analyzed

above. Following the same steps—applying Hölder's inequality with conjugate exponents, leveraging Lemma 3 of Abadie and Imbens (2006) for the moments of $K(c)$, and using the moment bounds from Assumption **??**—we can establish that this term converges to zero in probability as $n_T \to \infty$.

Specifically, the weighted error differences satisfy:

$$\frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{c=1}^{n_C} K(c) \left| \varepsilon_c - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \xrightarrow{p} 0$$

as $n_T \to \infty$, by the convergence properties established in our analysis of the sampling error term. Therefore, $(A3) \xrightarrow{p} 0$ as $n_T \to \infty$.

For the fourth and final component of Term A (systematic differences within cluster):

$$(A4) = \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \right]$$

Similar to our analysis of term (A3), we can apply the Mean Value Theorem to bound each squared difference:

$$\begin{aligned} \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 &\leq \left(\max_{k \in \mathcal{C}_t} |f_0(X_j) - f_0(X_k)| \right)^2 \\ &\leq \left(\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot \max_{j,k \in \mathcal{C}_t} \|X_j - X_k\| \right)^2 \\ &\leq \left(\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right)^2 \end{aligned}$$

Thus:

$$
\begin{aligned}
|(A4)| &\leq \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left(\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right)^2 \\
&= \sum_{t=1}^{n_T} \frac{u_t \cdot |\mathcal{C}_t|}{|\mathcal{C}_t| - 1} \left(\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right)^2 \\
&\leq 2 \cdot \sum_{t=1}^{n_T} u_t \left(\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right)^2 \\
&\leq 2 \cdot \left(\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \right] \right)^2
\end{aligned}
$$

By Assumption 6, $\sup_{x \in \mathcal{C}_t} |f_0'(x)| \cdot r(\mathcal{C}_t) \leq M$ for all t. Furthermore, by Assumption 4 (Shrinking Clusters), we have $\lim_{n \to \infty} \sup_{t \in \mathcal{T}} r(\mathcal{C}_t) = 0$. Therefore, even if $\sup_{x \in \mathcal{C}_t} |f_0'(x)|$ is unbounded as $n \to \infty$, their product with $r(\mathcal{C}_t)$ remains bounded by M, and the entire term $(A4) \xrightarrow{p} 0$ as $n_T \to \infty$.

For Term B, which involves the difference between the weighted and unweighted average of true variances:

$$
\begin{aligned}
\text{Term B} &= \sum_{t=1}^{n_T} \left(u_t \sigma_t^2 - \frac{1}{n_T} \sigma_t^2 \right) \\
&= \sum_{t=1}^{n_T} \left(\frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T} \right) \sigma_t^2
\end{aligned}
$$

By Assumption **??**, we know that $\sigma_{\min}^2 \leq \sigma_t^2 \leq \sigma_{\max}^2$ for all t. Therefore, by the triangle inequality:

$$
\begin{aligned}
|\text{Term B}| &\leq \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T} \right| \sigma_{\max}^2 \\
&= \sigma_{\max}^2 \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T} \right|
\end{aligned}
$$

Let $|\bar{\mathcal{C}}| = \frac{1}{n_T} \sum_{t=1}^{n_T} |\mathcal{C}_t|$ be the average cluster size. Then:

$$
\begin{aligned}
|\text{Term B}| &\leq \sigma^2_{\max} \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{n_T \cdot |\bar{\mathcal{C}}|} - \frac{1}{n_T} \right| \\
&= \frac{\sigma^2_{\max}}{n_T} \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{|\bar{\mathcal{C}}|} - 1 \right| \\
&= \frac{\sigma^2_{\max}}{n_T \cdot |\bar{\mathcal{C}}|} \sum_{t=1}^{n_T} \left| |\mathcal{C}_t| - |\bar{\mathcal{C}}| \right|
\end{aligned}
$$

By the Law of Large Numbers, as $n_T \to \infty$, the variability in cluster sizes relative to their mean diminishes. Specifically, under our assumptions: The Shrinking Clusters Assumption 4 ensures that all clusters become increasingly homogeneous. For typical matching procedures like M-nearest neighbor matching, $|\mathcal{C}_t| = M$ for all t, making this term exactly zero. For other matching procedures, the variance of $|\mathcal{C}_t|$ relative to n_T approaches zero as n_T increases.

This means that $\frac{1}{n_T} \sum_{t=1}^{n_T} \left| |\mathcal{C}_t| - |\bar{\mathcal{C}}| \right| \xrightarrow{p} 0$ as $n_T \to \infty$. Therefore, Term B converges to zero in probability as $n_T \to \infty$, completing our proof that $S^2 \xrightarrow{p} \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2$.

C Proof of Lemma 4.2

Proof. We begin by expanding the pooled variance estimator:

$$
\begin{aligned}
\hat{V}_{E,\text{lim}} &:= \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \\
&= \frac{1}{n_T^2} \sum_{t=1}^{n_T} \sigma_t^2 + \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \left(\frac{\sum_{j \in \mathcal{C}} w_j^2}{n_T^2} \right) \\
&= \frac{1}{n_T^2} \sum_{t=1}^{n_T} \sigma_t^2 + \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} \left[w_j^2 \cdot \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \right]
\end{aligned}
$$

The error variance V_E can be similarly expanded:

$$V_E = \frac{1}{n_T^2}\left(\sum_{t=1}^{n_T}\sigma_t^2 + \sum_{j=n_T+1}^{n_T+n_C} w_j^2\sigma_j^2\right)$$
$$= \frac{1}{n_T^2}\sum_{t=1}^{n_T}\sigma_t^2 + \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j^2\sigma_j^2$$

To establish asymptotic equivalence, we analyze the difference:

$$\hat{V}_{E,\text{lim}} - V_E = \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C}\left[w_j^2 \cdot \left(\frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2\right)\right] - \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j^2\sigma_j^2$$
$$= \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j^2 \cdot \left[\left(\frac{1}{n_T}\sum_{t=1}^{n_T}\sigma_t^2\right) - \sigma_j^2\right]$$

Using the definition of $\text{ESS}(\mathcal{C})$, we can rewrite:

$$\hat{V}_{E,\text{lim}} - V_E = \frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\sigma_t^2 - \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j^2\sigma_j^2$$

We decompose this difference into two terms:

$$\hat{V}_{E,\text{lim}} - V_E = \underbrace{\left(\frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\sigma_t^2 - \frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\overline{\sigma_t^2}\right)}_{(I)}$$
$$+ \underbrace{\left(\frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\overline{\sigma_t^2} - \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j^2\sigma_j^2\right)}_{(II)}$$

where $\overline{\sigma_t^2} = \sum_{j\in\mathcal{C}_t} w_{jt}\sigma_j^2$ represents the weighted average of variances for control units matched to treated unit t.

For Term (I), we have:

$$(I) = \frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\left(\sigma_t^2 - \overline{\sigma_t^2}\right) \xrightarrow{p} 0$$

This convergence follows from Assumption **??** (Continuity/Lipschitz Variance). As the matching quality improves under Assumption 4 (Shrinking Clusters), the difference between σ_t^2 and $\overline{\sigma_t^2}$ dimin-

ishes because matched control units have variance values increasingly similar to their corresponding treated units.

For Term (II), we analyze:

$$\begin{aligned}\frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\overline{\sigma_t^2} &= \frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{t=1}^{n_T}\sum_{j\in\mathcal{C}_t} w_{jt}\sigma_j^2 \\ &= \frac{1}{n_T}\frac{1}{\text{ESS}(\mathcal{C})}\sum_{j=n_T+1}^{n_T+n_C} w_j\sigma_j^2 \quad \text{where } w_j = \sum_{t=1}^{n_T} w_{jt} \\ &= \frac{1}{n_T}\frac{\sum_{j\in\mathcal{C}} w_j^2}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C} w_j\sigma_j^2\end{aligned}$$

Substituting this into Term (II):

$$\begin{aligned}(II) &= \frac{1}{n_T^2}\sum_{j=n_T+1}^{n_T+n_C}\left(\frac{\sum_{j'\in\mathcal{C}} w_{j'}^2}{n_T}w_j - w_j^2\right)\sigma_j^2 \\ &= \frac{1}{n_T}\frac{1}{n_T}\sum_{j=n_T+1}^{n_T+n_C}\left(\frac{\sum_{j'\in\mathcal{C}} w_{j'}^2}{n_T}w_j - w_j^2\right)\sigma_j^2\end{aligned}$$

Under standard matching procedures, the expression $\left(\frac{\sum_{j'\in\mathcal{C}} w_{j'}^2}{n_T}w_j - w_j^2\right)$ is bounded. This follows from the properties of weights in matching estimators.

Thus, Term (II) can be bounded:

$$\begin{aligned}(II) &\leq \frac{1}{n_T}\frac{1}{n_T}n_C\left(C\cdot\bar{\sigma}^2\right) \\ &\xrightarrow{p} 0 \quad \text{as } n_T\to\infty\end{aligned}$$

where C is a constant and $\bar{\sigma}^2$ is bounded by Assumption **??**. This convergence holds because $\frac{n_C}{n_T^2}\to 0$ as $n_T\to\infty$ under standard sampling conditions (Assumption 3).

Since both Term (I) and Term (II) converge to zero in probability, we conclude:

$$\left|\hat{V}_{E,\text{lim}} - V_E\right| \xrightarrow{p} 0 \quad \text{as } n_T\to\infty$$

□

D Other useful Lemmas

D.1 Lemma and Proof of Shrinking Cluster Distance under Compact Support

Lemma D.1 (Compact Support $\implies$ Vanishing Matching Discrepancy)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be the support of the covariates X, and assume $\mathcal{X}$ is compact. Assume further that the distribution of X has a density $f_X(x)$ that is bounded above and below on $\mathcal{X}$ (i.e., $0 < f_{\min} \leq f_X(x) \leq f_{\max} < \infty$ for all $x \in \mathcal{X}$). Consider any matching procedure that pairs each observation with at least one other "nearest" neighbor (for example, one-to-M nearest neighbor matching or radius matching with a fixed caliper). Then as the sample size $N \to \infty$, the maximum distance between any matched units goes to zero. In particular:*

$$\max_{i=1,\ldots,N} \min_{j \neq i} \|X_i - X_j\| \xrightarrow{p} 0.$$

That is, the distance between each observation and its closest match converges to zero in probability (and in fact, almost surely).

Proof Sketch. This result is a direct consequence of the compactness of $\mathcal{X}$ and the bounded-positive density assumption. The argument follows the intuition of Lemma 1 in **?**.

Because $\mathcal{X}$ is compact, for any given radius $\varepsilon > 0$ we can **cover $\mathcal{X}$ by finitely many small balls** (or other simple sets) of diameter at most ε. Concretely, by the Heine–Borel covering theorem, there exists a finite collection of sets $\{B_1, \ldots, B_R\}$ such that $\mathcal{X} \subseteq \bigcup_{r=1}^{R} B_r$ and each B_r has $\operatorname{diam}(B_r) < \varepsilon$. For example, one can take B_r to be balls (in $\|\cdot\|$ norm) of radius $\varepsilon/2$, or cubes of side-length ε, etc., partitioning the space. By construction, for any x, x' in the same B_r, we have $\|x - x'\| < \varepsilon$.

Next, because $f_X(x) \geq f_{\min} > 0$ on $\mathcal{X}$, *every region of $\mathcal{X}$ has some probability mass*. In particular, each B_r has $\Pr(X \in B_r) > 0$. When we draw N i.i.d. observations $\{X_i\}_{i=1}^{N}$, the expected number of samples falling in B_r is $N \Pr(X \in B_r)$, which grows linearly with N. By the law of large numbers, for large N it is overwhelmingly likely that each B_r contains at least one sample point (indeed, at least $\approx N \Pr(X \in B_r)$ points). More strongly, since $N \Pr(X \in B_r) \to \infty$, the probability that any given B_r contains **fewer than 2 points** goes to zero as $N \to \infty$. In

fact, one can apply a union bound or a Poisson approximation to show:

$$P\Big(\exists\, r:\ B_r \text{ contains 0 or 1 points}\Big)\ \to\ 0, \quad \text{as } N \to \infty.$$

(Informally: with infinitely many draws, every subset B_r will eventually have multiple points due to the density's support.)

Now consider any observation i and let B_r be one of the covering sets that contains X_i. By the above argument, for large N that B_r will contain at least one **other** observation $j \neq i$. Thus, i has at least one neighbor j with $X_j \in B_r$ alongside X_i. By the diameter property of B_r, the distance between i and this neighbor j is bounded by $\|X_i - X_j\| < \varepsilon$. Since i was arbitrary, we have shown that **for every i there exists some match j with $\|X_i - X_j\| < \varepsilon$** (with probability $\to 1$ as N large).

Because $\varepsilon > 0$ was arbitrary, it follows that the maximum matching distance in the sample is $< \varepsilon$ w.p. $\to 1$ for any ε. In probabilistic terms:

$$\Pr\Big\{\max_{1\le i\le N}\min_{j\neq i}\|X_i - X_j\| < \varepsilon\Big\}\ \to\ 1 \quad \forall\, \varepsilon > 0,$$

which is equivalent to $\max_i \min_j \|X_i - X_j\| \xrightarrow{p} 0$. (In fact, one can show almost sure convergence to 0 by invoking the Borel–Cantelli lemma, since the event that a given B_r is empty eventually occurs at most finitely many times.)

This proves that the largest distance within any matched pair (or cluster) converges to zero as N increases, under the stated assumptions. □ □

Remark: This lemma provides the theoretical justification for Assumption 4 in our paper. It confirms that under compact support (and overlap), **common matching methods produce asymptotically exact matches**. Notably, nearest-neighbor matching on $\mathbf{X}$ yields $\|\widehat{X}_i - X_i\| = O_p(N^{-1/d})$, so the discrepancy vanishes as $N \to \infty$. This is analogous to the overlap condition in propensity score matching, where a bounded propensity support guarantees treated units find control units with arbitrarily close propensity scores. Conversely, if covariate support were unbounded or the density went to zero in some region, one could not guarantee such shrinking distances – there would always be a chance of an isolated observation with no close neighbor (e.g. an outlier), resulting in

a non-vanishing maximum discrepancy.

D.2 Lemma for Proof of Theorem 4.1

Lemma D.2 (Uniform convergence of variances)**.** *Under Assumptions 4 (Shrinking Clusters) and* **??** *(Continuity/Lipschitz Variance), we have:*

$$\max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{a.s.} 0 \quad \text{as } n_c, n_T \to \infty \tag{33}$$

where $\sigma_c^2 = \sigma^2(X_c)$ *and* $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t \in T_c} \sigma^2(X_t)$ *with* T_c *being the set of treated units matched to control unit* c*.*

Proof. Let us establish a framework for proving the uniform convergence. For a given sample size $n = n_c + n_T$, define d_n as the maximum matching distance such that $T_c = \{t \in \mathcal{T} : \|X_c - X_t\| \leq d_n\}$ for each control unit c. Under Assumption 4 (Shrinking Clusters), we have $d_n \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

By Assumption **??** (Continuity/Lipschitz Variance), there exists a Lipschitz constant L such that:

$$|\sigma^2(x) - \sigma^2(y)| \leq L \cdot \|x - y\| \tag{34}$$

for all $x, y \in \mathcal{X}$.

For any control unit c, we have:

$$\begin{aligned}
|\sigma_c^2 - \bar{\sigma}_c^2| &= \left| \sigma^2(X_c) - \frac{1}{K(c)} \sum_{t \in T_c} \sigma^2(X_t) \right| \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} |\sigma^2(X_c) - \sigma^2(X_t)| \quad \text{(by triangle inequality)} \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} L \cdot \|X_c - X_t\| \quad \text{(by Lipschitz condition)}
\end{aligned}$$

Since all $t \in T_c$ satisfy $\|X_c - X_t\| \leq d_n$ by construction, we have:

$$\begin{aligned}
|\sigma_c^2 - \bar{\sigma}_c^2| &\leq \frac{1}{K(c)} \sum_{t \in T_c} |\sigma^2(X_c) - \sigma^2(X_t)| \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} L \cdot \|X_c - X_t\| \\
&\leq \frac{L}{K(c)} \sum_{t \in T_c} d_n \\
&= L \cdot d_n
\end{aligned}$$

This inequality holds uniformly for every control unit c, as the Lipschitz constant L applies across all matches and d_n represents the maximum matching distance. Therefore, the maximum deviation across all control units is bounded by:

$$\max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \leq L \cdot d_n$$

By Assumption 4, $d_n \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Since L is a finite constant, we conclude:

$$\max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n_c, n_T \to \infty \tag{35}$$

This establishes the uniform convergence of variance differences across all control units.

This establishes the uniform convergence of $\sigma_c^2 - \bar{\sigma}_c^2$ across all control units simultaneously, not merely pointwise convergence for each fixed c. □

E Compare the Lipschitz Condition to that in the Existing Litarature

In the existing literature, the function $f(x)$ is often assumed to be locally Lipschitz on any compact set $\mathcal{X} \subset \mathbb{R}$. This implies that for any compact set $\mathcal{X} = [a, b]$, there exists a constant $L_{\mathcal{X}} < \infty$ such that:

$$|f(x) - f(y)| \leq L_{\mathcal{X}} |x - y|, \quad \forall x, y \in \mathcal{X}.$$

For example, consider $f(x) = x^2$, where the derivative $f'(x) = 2x$. On $\mathcal{X} = [0, 100]$, the Lipschitz constant is:

$$L_{\mathcal{X}} = 2 \cdot \max_{x \in \mathcal{X}} |x| = 200.$$

This large constant makes the bound impractical in matching-based inference, where overly conservative bounds can restrict the formation of matched sets.

In contrast, our Derivative Control condition improves on the Lipschitz assumption by explicitly tying the slope of $f(x)$ to the size of the matched set. Specifically, it requires:

$$\sup_{x \in \mathcal{C}_t} \left| f'(x) \right| \cdot \text{radius}(\mathcal{C}_t) \leq M,$$

where:

- $\mathcal{C}_t$ is the matched set for a given t,
- $\text{radius}(\mathcal{C}_t)$ is the diameter of the matched set in x-space,
- M is a universal constant independent of the matched set size.

This condition offers several practical advantages:

1. Localized Control: Instead of requiring a single large Lipschitz constant $L_{\mathcal{X}}$ over a wide range, our condition focuses on smaller, localized matched sets.
2. Adaptive Bounds: When the derivative $f'(x)$ is large, our condition naturally enforces smaller matched set radii to maintain practical bounds. For instance:

$$\text{If } f'(x) = 100 \text{ (as for } x = 50), \text{ then } \text{radius}(\mathcal{C}_t) \leq \frac{M}{100}.$$

3. Real-World Applicability: In real-world matching problems, matched sets are typically small, and our condition aligns with this reality by providing sharper, more practical bounds than the overly conservative Lipschitz constant.

To summarize, while the Lipschitz assumption is valid on compact sets, the associated constants $L_{\mathcal{X}}$ can become impractically large for functions like $f(x) = x^2$ over wide intervals. By explicitly

accounting for both the derivative and the size of matched sets, our condition provides a more precise and practical framework for matching-based inference.

F Comparison with Theorem 1 of White (1980)

Our theorem, stated as Theorem **??**, differs from Theorem 1 of White (1980) in several key aspects. While both results address consistency in variance estimation under heteroskedasticity, the differences lie in the frameworks, assumptions, and proof strategies.

F.1 Parametric vs. Nonparametric Framework

White's Theorem 1 is based on a regression model $Y_i = X_i\beta_0 + \varepsilon_i$, where ε_i represents independent but non-identically distributed (i.n.i.d.) errors. The parametric form $X_i\beta_0$ is central, and β_0 is estimated via ordinary least squares (OLS). Heteroskedasticity arises through $\mathrm{Var}(\varepsilon_i \mid X_i) = g(X_i)$, where $g(X_i)$ is a known (possibly parametric) function. In contrast, our theorem relies on a nonparametric matching estimator for treatment effects, without assuming a parametric form for $f(X_i)$. Matching is governed by hyperparameters like the number of neighbors or the maximum matching radius, but these are not estimated from the data in the regression sense. Heteroskedasticity arises through $\sigma^2(X_i)$, where $\sigma^2(\cdot)$ is a uniformly continuous function.

F.2 White's Setup: Estimating $\mathrm{Var}(\hat{\beta})$ vs. Cluster-Based Variance Estimation

White's Theorem 1 focuses on the heteroskedasticity-consistent (HC) covariance matrix estimator for $\hat{\beta}$. It defines the matrix

$$\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 X_i' X_i, \quad \text{where} \quad \hat{\varepsilon}_i = Y_i - X_i\hat{\beta}.$$

White proves $\hat{V}_n \xrightarrow{\text{a.s.}} \bar{V}_n$, where $\bar{V}_n$ is the asymptotic covariance matrix of the regressors. Our theorem, on the other hand, defines cluster-level residual variance estimators s_t^2 for each treated

unit $t \in \mathcal{T}$, given its matched controls $\mathcal{C}_t$. The overall variance estimator is

$$S^2 = \frac{1}{n_T} \sum_{t=1}^{n_T} s_t^2, \quad \text{where} \quad s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} e_{tj}^2.$$

We prove $|S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2| \xrightarrow{\text{a.s.}} 0$, showing consistency for the average cluster variance.

F.3 Homoskedasticity in Matched Clusters vs. General Heteroskedasticity

White's Theorem 1 allows general heteroskedasticity: $\text{Var}(\varepsilon_i \mid X_i) = g(X_i)$, where $g(\cdot)$ can vary arbitrarily across observations. Errors are independent but not identically distributed (i.n.i.d.). Our theorem also allows heteroskedasticity: $\sigma^2(X_i)$ varies with X_i. However, within each matched cluster $\{t\} \cup \mathcal{C}_t$, we assume $\sigma_j^2 \approx \sigma_t^2$ for $j \in \mathcal{C}_t$, based on a uniform continuity (or Lipschitz) assumption on $\sigma^2(\cdot)$.

F.4 Proof Strategy and Key Assumptions

White's proof strategy relies on expanding $\hat{V}_n - \bar{V}_n$ and showing that

$$\hat{V}_n - \bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i^2 X_i' X_i - E[\varepsilon_i^2 X_i' X_i] \right) \xrightarrow{\text{a.s.}} 0.$$

White uses assumptions on finite moments of ε_i and X_i (Assumptions 2–4 in White (1980)) and uniform integrability conditions. Our proof, in contrast, relies on showing that for matched clusters $\{t\} \cup \mathcal{C}_t$, the residual variance s_t^2 converges to the true variance σ_t^2. We leverage uniform continuity of $\sigma^2(\cdot)$ to argue that $\sigma_j^2 \to \sigma_t^2$ as $\|X_{tj} - X_t\| \to 0$. We then apply a version of the Law of Large Numbers (LLN) for matched clusters.

F.5 Summary of Differences

The key differences between White's theorem and our theorem can be summarized as follows. First, White's theorem is regression-based, while our theorem is matching-based. Second, White

assumes a parametric model $Y_i = X_i\beta_0 + \varepsilon_i$, whereas our model assumes a nonparametric $f_1(X)$, $f_0(X)$. Third, White's focus is on a robust covariance estimator for $\hat{\beta}$, while ours is on residual variance from matched clusters. Fourth, White allows fully general $g(X_i)$, whereas our clusters assume approximate homoskedasticity ($\sigma_j^2 \approx \sigma_t^2$). Finally, White's framework has no matching hyperparameters, while ours depends on predefined criteria for matching (e.g., number of neighbors or radius).

G Full Simulation Settings

G.1 Simulation Settings for Caliper Matching

G.1.1 Data-Generating Process

Covariates. We simulate:

- Fifty treated units, each drawn from multivariate normal distributions centered at $(0.25, 0.25)$ and $(0.75, 0.75)$, with covariance matrices

$$\begin{bmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{bmatrix}.$$

- Two hundred and twenty-five control units, each from multivariate normal distributions centered at $(0.25, 0.75)$ and $(0.75, 0.25)$, also with covariance matrices

$$\begin{bmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{bmatrix}.$$

- One hundred additional control units distributed uniformly on the unit square $[0, 1] \times [0, 1]$ to ensure an adequate overlap region.

Outcomes. For each unit with covariates (x_1, x_2), we generate an outcome via:

$$Y \;=\; f_0(x_1, x_2) \;+\; Z \cdot \tau(x_1, x_2) \;+\; \epsilon,$$

where:

- f_0 is the density function for a multivariate normal distribution centered at $(0.5, 0.5)$ with covariance matrix
$$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$
- $\tau(x_1, x_2) = 3x_1 + 3x_2$ is the true treatment effect.
- $\epsilon \sim N(0, 0.5^2)$ is homoskedastic measurement error.

Overlap Variation. We vary the degree of overlap by increasing the proportion of control units drawn uniformly from the unit square and reducing those centered at $(0.25, 0.75)$ and $(0.75, 0.25)$. For each overlap configuration, we run 500 Monte Carlo trials.

G.1.2 Matching and Estimation

Matching is performed using a synthetic-control-like optimization within local neighborhoods (calipers). We use an adaptive strategy for the caliper size: shrinking it to select no more than five units in dense regions, and expanding it to include at least one unit in sparse regions. With larger calipers, the optimization often selects only a few control units, which can be distant from the treated point. We refer readers to Che et al. (2024) for details.

G.2 Simulation Settings for Bootstrap Variance Estimation

G.2.1 Data-Generating Process

The Data Generating Process (DGP) is defined as:

$$
\begin{aligned}
&\{Y_i, Z_i, X_i\}_{i=1}^N, \\
&Y_i(1) = \tau + m\left(\|X_i\|\right) + \epsilon_i, \quad Y_i(0) = m\left(\|X_i\|\right) + \epsilon_i, \\
&Z_i = \mathbb{I}\left\{P\left(X_i\right) \geq v_i\right\}, \quad v_i \sim U[0,1], \\
&P\left(X_i\right) = \gamma_1 + \gamma_2 \left\|X_i\right\|, \quad X_i = \left(X_{1i}, \ldots, X_{ki}\right)', \\
&X_{ji} = \xi_i \left|\zeta_{ji}\right| / \left\|\zeta_i\right\| \quad \text{for } j = 1, \ldots, k, \\
&\xi_i \sim U[0,1], \quad \zeta_i \sim N\left(0, I_k\right), \quad \epsilon_i \sim N\left(0, 0.2^2\right),
\end{aligned}
$$

The function $m(z)$ is defined as:

$$
m(z) = 0.4 + 0.25\sin(8z - 5) + 0.4\exp\left(-16(4z - 2.5)^2\right)
$$

[3]

Additional parameters include: Treatment Model Parameters: $\gamma_1 = 0.15$, $\gamma_2 = 0.7$; True Effect: $\tau = 0$; Number of Replicates: 100; Estimand: Average Treatment Effect on the Treated (ATT).

G.2.2 Matching Method and Point Estimator

We implement the paper's matching procedure using 8 nearest neighbors (8-NN) with uniform weighting, assigning a weight of 1/8 to each matched control unit.

Given a matching procedure, we define the matching estimator for the ATT as:

$$
\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j\right) \tag{36}
$$

where $\mathcal{C}_t$ is the set of matched controls, and w_{jt} represents the weight assigned to control unit j

[3] This is curve 6 in Otsu and Rai (2017).

when matched to treated unit t.

G.2.3 Debiasing Method

A debiasing model estimates the conditional mean function $\mu(z, x) = E[Y \mid Z = z, X = x]$. It is used to offset the bias to achieve valid inference (see Section 3.4 for discussion of the issue). The debiased estimator is defined as:

$$\tilde{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{\mu}(0, X_t) - \sum_{j \in \mathcal{C}_t} w_{jt} (Y_j - \hat{\mu}(0, X_j)) \right) \tag{37}$$

Additional implementation details include:

- **Model Choice**: Linear model
- **Training Data**: Control data only
- **Cross-fitting**: Implemented by dividing the control data into two halves

G.2.4 Variance Estimators

Bootstrap Variance Estimator.

- Step 1: Use data with $Z_i = 0$ to construct $\hat{\mu}(0, x) = \hat{E}[Y|Z = 0, X = x]$.
- Step 2: Construct debiased estimate for each treated unit $t \in \mathcal{T}$:

$$\tilde{\tau}_t = (Y_t - \hat{\mu}(0, X_t)) - \sum_{j \in \mathcal{C}_t} w_{jt} (Y_j - \hat{\mu}(0, X_j))$$

- Step 3: Construct the debiased estimator: $\tilde{\tau} = \frac{1}{n_t} \sum_{t \in \mathcal{T}} \tilde{\tau}_t$
- Step 4: Construct the debiased residuals $R_t = \tilde{\tau}_t - \tilde{\tau}$
- Step 5: Perform Wild bootstrap on $\{R_t\}$ with special sampling weights
- Step 6: Construct confidence interval from bootstrap distribution

Pooled Variance Estimator.

- Step 1: Obtain the debiased estimator $\tilde{\tau}_t$
- Step 2: Estimate the variance using:

$$\hat{V} = S^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \tag{38}$$

  where S^2 is a pooled variance estimator across clusters of treated and matched controls

- Step 3: Construct the 95% confidence interval by $\tilde{\tau} \pm 1.96 * \sqrt{\hat{V}}$

G.3 Kang and Schafer Simulation Settings

G.3.1 Data-Generating Process

The Kang and Schafer data generating process is structured as follows:

1. Generate latent covariates $Z_1, Z_2, Z_3, Z_4 \sim \mathcal{N}(0, I_4)$ where I_4 is the 4×4 identity matrix.
2. Calculate propensity scores:

$$p(Z) = \frac{1}{1 + \exp(Z_1 - 0.5Z_2 + 0.25Z_3 + 0.1Z_4)} \tag{39}$$

3. Generate treatment assignment as $T \sim \text{Bernoulli}(p(Z))$.
4. Define the outcome model:

$$f_0(Z) = \frac{210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4}{50} \tag{40}$$

5. Generate potential outcomes with treatment effect τ:

$$Y(0) = f_0(Z) + \epsilon \tag{41}$$

$$Y(1) = f_0(Z) + \tau + \epsilon \tag{42}$$

   where $\epsilon \sim \mathcal{N}(0, 1)$.

6. The observed outcome is:

$$Y = Y(0)(1-T) + Y(1)T = f_0(Z) + \tau \cdot T + \epsilon \tag{43}$$

7. Transform the latent covariates Z to create the observed covariates X:

$$X_1 = \exp(Z_1/2) \tag{44}$$

$$X_2 = \frac{Z_2}{1+\exp(Z_1)} + 10 \tag{45}$$

$$X_3 = \left(\frac{Z_1 \cdot Z_3}{25} + 0.6\right)^3 \tag{46}$$

$$X_4 = (Z_2 + Z_4 + 20)^2 \tag{47}$$

In our simulations, we use $n = 500$ observations and set the true treatment effect $\tau = 0$.