Asymptotics of Caliper Matching Estimators for Average Treatment Effects

Máté Kormos[1], Stéphanie van der Pas[2], and Aad van der Vaart[1]

[1]Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, Delft, The Netherlands

[2]Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, The Netherlands and Amsterdam Public Health Methodology, Amsterdam, The Netherlands

Abstract

Caliper matching is used to estimate causal effects of a binary treatment from observational data by comparing matched treated and control units. Units are matched when their propensity scores, the conditional probability of receiving treatment given pretreatment covariates, are within a certain distance called caliper. So far, theoretical results on caliper matching are lacking, leaving practitioners with ad-hoc caliper choices and inference procedures. We bridge this gap by proposing a caliper that balances the quality and the number of matches. We prove that the resulting estimator of the average treatment effect, and average treatment effect on the treated, is asymptotically unbiased and normal at parametric rate. We describe the conditions under which semiparametric efficiency is obtainable, and show that when the parametric propensity score is estimated, the variance is increased for both estimands. Finally, we construct asymptotic confidence intervals for the two estimands.

Máté Kormos thanks Hong Deng for her help and Lisa Voois for suggestions (both of them are from Tinbergen Institute and Erasmus School of Economics, Erasmus University Rotterdam). Stéphanie van der Pas acknowledges that the research leading to these results has received funding from the Dutch Research Council (NWO) under grant VI.Veni.192.087. This research was also partially funded by a Spinoza grant from NWO.

1. Introduction

Matching is applied in empirical studies to estimate the causal effect of a binary treatment from observational data. The estimate is the mean difference in the outcome of interest of matched treated and control units. Matches may be formed in various ways. We consider matching on the propensity score, the conditional probability of receiving treatment given the observed pretreatment covariates (Rosenbaum and Rubin, 1983). Specifically, we consider caliper matching, where a treated and a control unit are matched if their propensity scores are within a certain distance called *caliper* (Cochran and Rubin, 1973; Dehejia and Wahba, 1998). Caliper matching is applied in empirical research such as labour (Dehejia and Wahba, 2002; Huber et al., 2015b) and health economics (Erhardt, 2017; Salmasi and Pieroni, 2015; Keng and Sheu, 2013), policy evaluation (Bannor et al., 2020; Patel-Campillo and García, 2022), business and finance (Shen and Chang, 2009; Heese et al., 2017) as well as healthcare (Capogrossi and You, 2017; Cho, 2018; Vecchio et al., 2018; Izudi et al., 2019; Wang et al., 2020; Brenna, 2021; Krishnamoorthy and Rehman, 2022). Nonetheless, no rigorous results have been established on the choice of the caliper and the limiting distribution of the estimator.

Our contribution is a theory driven caliper choice, the derivation of the asymptotic distribution of the caliper matching estimator based on propensity scores, and the construction of asymptotic confidence intervals. We consider the estimation of the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT). We show that when the order of the caliper decreases at the right speed as the sample size n increases, the estimators of both estimands are asymptotically unbiased and normal at $\sqrt{n}$-rate, even when the parametric propensity score is estimated. In the rest of this section, we situate our contribution in the literature.

Matching has attracted much attention in the literature, with the idea of comparing similar units dating back to at least Densen et al. (1952); see Cochran (1953). Cochran and Rubin (1973) review then-available matching methods applicable to observational studies. The reader is referred to Rubin (2006) for a collection of historical results and to Stuart (2010) for a comprehensive survey. Abadie and Imbens (2006) present a key result closely related to ours. They study nearest neighbor matching, where the M closest units in terms of covariates are matched to a given unit. They show that nearest neighbor matching on covariates is asymptotically normal, but unbiased only when we match on a scalar variable, such as the propensity score. Providing the identification results for unbiasedness, the foundations of propensity score matching is laid down by Rosenbaum and Rubin (1983). Abadie and Imbens (2016) derive some asymptotic properties of nearest neighbor matching on the estimated parametric propensity score. They discretise the maximum likelihood estimator of the propensity score parameter and show that the resulting matching estimator converges to a normal distribution as, first, the sample size increases and, *then*, the discretisation gets finer. Since their approach changes the estimator, this asymptotic result is not equivalent to the asymptotic normality of nearest neighbor matching on the estimated parametric propensity score. In contrast, we do not change the estimator, nor do we appeal to discretisation arguments and double limits. Employing sample-splitting to estimate the propensity score, we establish the asymptotic normality of caliper matching on the estimated parametric propensity score as the sample size increases. Consequently, we are able to construct confidence intervals for ATE and ATT, centered at the caliper matching estimator based on the estimated propensity scores, which get more reliable as the sample size increases.

The first mention of caliper matching appears to be in Cochran and Rubin (1973). Therein, it is analysed for a few specific models and is compared with other matching methods, such as nearest neighbor. The caliper is chosen based on the variances of the outcome in the treatment and control group. Rosenbaum and Rubin (1985) seem to be the first to consider caliper matching involving the propensity score as well as the covariates. They assume a logistic model for the propensity score, and match on the logit of the propensity score, that is, a linear function of the covariates. They choose the caliper based on the variances of the logit in the treatment and control group. The caliper choices of Cochran and Rubin (1973) and Rosenbaum and Rubin (1985) may lead to a large enough number of matches to reduce the variance of the caliper matching estimator. However, they do not make the bias of the caliper matching estimator converge to zero — unless the caliper is used in combination with nearest neighbor matching; see next paragraph —, for that the caliper needs to shrink with the sample size as we show in our present work.

Some authors, including Rosenbaum and Rubin (1985), use the term caliper matching to refer to nearest neighbor matching with a caliper restriction: the M nearest units are to be matched, but only if they are within the caliper. Others, for instance Dehejia and Wahba (1998), use the term to mean that all units within the caliper are matched, even though they may be differently weighted.[1] We adopt the latter approach with uni-

[1] The two interpretations coincide when M is taken to be, for example, n in the caliper restriction case. As M is usually set to a constant independent of n, it is reasonable to distinguish the two interpretations.

form weights, sometimes also called radius matching (Huber et al., 2015a), because of its simplicity. Caliper matching can then be regarded as a kernel matching method with rectangular kernel and the bandwidth equal to the caliper. As such, the seminal work of Heckman et al. (1998), establishing the asymptotic normality of the kernel matching estimator of ATT even for nonparametrically estimated propensity score — with bandwidth choice further investigated by Frölich (2005) —, is closely related to our work. However, their results do not apply to caliper matching because they require the kernel to be Lipschitz continuous. The rectangular kernel fails to be so, prohibiting the asymptotic linear expansion of the kernel matching estimator, which is key to their argument. The work of Lee (2018) is similar in spirit. It extends Heckman et al. (1998) to a richer set of estimands beyond average effects using kernel matching methods, but also assuming a smooth kernel, excluding the rectangular one of caliper matching.

We overcome the nonsmoothness of the rectangular kernel by employing empirical process theory in Alexander (1987) and Van der Vaart and Wellner (1996). Writing the number of matches in terms of empirical measures enables us to characterise the asymptotic behaviour of caliper matching using ratio and tail bounds for empirical measures and processes. Furthermore, we can establish the efficiency properties of caliper matching. More efficient estimators have smaller variance and thus yield narrower confidence intervals. The efficiency of caliper matching depends on the estimand, the observed sample, the regression of the outcome on the covariates, and the knowledge of the propensity score.

First, we consider the case when the propensity score is known. We prove that if we only observe the propensity scores in our sample but not the covariates, or the regression of the outcome on the covariates only depends on the covariates through the propensity score, then the limiting variance of the caliper matching estimator of (i) ATE reaches the semiparametric lower bound; (ii) ATT reaches the semiparametric lower bound for *unknown* propensity score (Hahn, 1998). The latter is not the best possible result as the lower bound for ATT, unlike ATE, is smaller when the propensity score is known (Hahn, 1998). Yet, we show that caliper matching is more efficient than nearest neighbor matching on the propensity scores studied by Abadie and Imbens (2006, 2016), yielding narrower confidence intervals for ATE as well as ATT — regardless of whether we observe the covariates in the sample or whether the outcome regression depends on the covariates or the propensity scores.

Second, if the propensity score is unknown, but we assume and estimate a parametric specification such as the logit or probit model, then the limiting variance of the caliper matching estimator of both estimands is in general larger compared to when the propensity score is known. Consequently, it remains unclear whether the caliper or the nearest neighbor matching (Abadie and Imbens, 2016) on the estimated propensity scores is more efficient.

Our assumptions include the usual common support for the propensity score, and smoothness conditions for the conditional moments of the outcome and for (the density of) the propensity score. We verify our assumptions for a logit or probit model for the propensity score and for smooth, potentially nonlinear and heteroskedastic, regression of the outcome on the covariates with a well-behaved density on a compact support.

The rest of the paper is organised as follows. In Section 2, we introduce the conceptual framework and the caliper matching estimator. Section 3 contains our contributions, the caliper choice and the asymptotic properties of the estimator. Section 4 concludes.

2. Preliminaries

2.1. Framework

We adopt the potential outcome framework of Neyman (1924) and Rubin (1974) with no interference between the units (stable unit-treatment value assumption, Rosenbaum and Rubin (1983)). Let D be the treatment indicator with value one corresponding to treatment and zero to control. The real-valued Y^1, Y^0 are the potential outcomes under treatment and control, respectively. We observe exactly one of Y^1 and Y^0, depending on D, so that the observed outcome is $Y = DY^1 + (1-D)Y^0$. The estimands of interest, ATE and ATT, are defined respectively as

$$\tau \coloneqq \mathbb{E}\left[Y^1 - Y^0\right], \quad \tau_{\mathrm{t}} \coloneqq \mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right].$$

To identify ATE and ATT from observational data, we assume that the observed pretreatment covariates X, taking values in $\mathcal{X} \subset \mathbb{R}^K$, account for all the systematic differences between treated and control units. Formally, the potential outcomes are assumed to be independent of the treatment participation given the covariates, which is a standard assumption of causal inference (Rubin, 1974).

Assumption 1 (Unconfoundedness). *$Y^0 \perp\!\!\!\perp D \mid X$ and $Y^1 \perp\!\!\!\perp D \mid X$.*

Let $\pi(x) \coloneqq \mathbb{P}\left(D = 1 \mid X = x\right)$ be the propensity score with conditional distribution function $F_d(p) \coloneqq \mathbb{P}\left(\pi(X) \leq p \mid D = d\right)$. The F_d are assumed to satisfy Assumption 2.

Assumption 2 (Propensity Score Distribution)**.** *(i)* F_0, F_1 *admit densities* f_0, f_1*, respectively. (ii)* f_0, f_1 *have the same compact support* $[\underline{p}, \bar{p}]$*,* $0 < \underline{p} < \bar{p} < 1$*. (iii)* f_0, f_1 *are strictly positive on their support. (iv)* f_0, f_1 *are continuous on their support.*

Assumption 2 imposes the same requirements on the propensity score distribution as Abadie and Imbens (2016), except that it also requires the densities f_0, f_1 to be strictly positive. This requirement ensures that the quantile functions F_d^{-1} have bounded derivatives, which we use for the caliper choice. It also plays a role in the proof of the asymptotic normality by ensuring that ratio bounds for empirical processes apply.[2]

Assumption 2 implies that if there is a unit with propensity score in some region of $[0, 1]$, then there is a positive probability of finding a unit from the opposite treatment group therein. This ensures that treated and control units can be compared in terms of their propensity scores. In combination with Assumption 1, this yields the identification of the estimands from observed variables, by comparing treated and control units with the same propensity scores (Rosenbaum and Rubin, 1983):

$$\tau = \mathbb{E}\left[\mathbb{E}\left[Y \mid D = 1, \pi(X)\right] - \mathbb{E}\left[Y \mid D = 0, \pi(X)\right]\right], \tag{1}$$

$$\tau_{\mathrm{t}} = \mathbb{E}\left[\mathbb{E}\left[Y \mid D = 1, \pi(X)\right] - \mathbb{E}\left[Y \mid D = 0, \pi(X)\right] \mid D = 1\right]. \tag{2}$$

2.2. Caliper Matching Estimator

We wish to construct estimators based on identification formulae (1) and (2) from an independently and identically distributed (i.i.d.) sample from the distribution of (Y, D, X), denoted by $((Y_i, D_i, X_i))_{i\in[n]}$, where $[n] \coloneqq \{1, 2, \ldots, n\}$. This would necessitate finding sample units with the same value of the propensity score, which is infeasible for continuously distributed propensity scores. Rather, matching estimators look for units with *similar* propensity scores. The caliper matching estimator explicitly controls the extent of similarity with the caliper δ, whose choice is discussed later on in Section 3.

Suppose for now that the propensity score is known. Given $\delta > 0$, the caliper matching estimator constructs the match set $\mathcal{J}(i) \coloneqq \{j \in [n] : D_j \neq D_i, |\pi(X_j) - \pi(X_i)| \leq \delta\}$ of unit $i \in [n]$. Next, it estimates the missing potential outcome of the unit with the mean outcome of units in the match set. Averaging out the difference between the (estimated) potential outcomes then gives the estimate of the causal effect. Let $M_i \coloneqq |\mathcal{J}(i)|$ be the

[2] The strict positivity of f_d implies that $\inf_{p\in[\underline{p},\bar{p}]} \int_{p-\delta}^{p+\delta} f_d(\tilde{p})\mathrm{d}\tilde{p} \gtrsim \delta > 0$, so that the denominator in the ratios of empirical to true measures is bounded away from zero, keeping the ratios finite.

number of matches of unit $i \in [n]$, and write $N_0 := \sum_{i\in[n]}(1 - D_i)$, $N_1 := \sum_{i\in[n]} D_i$ for the number of control and treated units, respectively. The estimators of ATE and ATT are defined respectively as

$$\hat{\tau}_\pi := \frac{1}{n}\sum_{i\in[n]}\left[D_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right) + (1 - D_i)\left(\frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j - Y_i\right)\right]\mathbb{1}_{M_i>0},$$

$$\hat{\tau}_{\mathrm{t},\pi} := \frac{1}{N_1}\sum_{i\in[n]} D_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right)\mathbb{1}_{M_i>0}.$$

The indicator $\mathbb{1}_{M_i>0}$, being one if unit i has matches and zero if not, ensures that only units that have matches are included in the estimate.

In practice, the propensity score is usually unknown. Often, it is assumed to follow a smooth parametric model, such as logit or probit. Following Abadie and Imbens (2016), we also make this assumption.

Assumption 3 (Smooth Parametric Propensity Score)**.** *(i) The propensity score is* $\mathbb{P}(D = 1 \mid X) = \pi(X, \theta_0)$ *for a parametric model* $\{\pi(\cdot, \theta) : \theta \in \Theta \subset \mathbb{R}^K\}$ *with* θ_0 *in the interior of* Θ*. (ii)* $\theta \mapsto \pi(x, \theta)$ *is differentiable in the neighbourhood of* θ_0 *for all* $x \in \mathcal{X}$*. (iii) The derivative in Assumption 3(ii) is bounded uniformly in* $x \in \mathcal{X}$ *in the neighbourhood of* θ_0*.*

The caliper matching estimator is then defined by a plug-in rule. Let $\mathcal{J}_\theta(i) := \{j \in [n] : D_j \neq D_i, |\pi(X_j, \theta) - \pi(X_i, \theta)| \leq \delta\}$ be the match set and $M_i(\theta) := |\mathcal{J}_\theta(i)|$ its cardinality for some $\theta \in \Theta$. For an estimator $\hat{\theta}$ of θ_0, the matching estimators of ATE and ATT are, respectively,

$$\hat{\tau}_{\hat{\pi}} := \frac{1}{n}\sum_{i\in[n]}\left[D_i\left(Y_i - \frac{1}{M_i(\hat{\theta})}\sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j\right) + (1 - D_i)\left(\frac{1}{M_i(\hat{\theta})}\sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j - Y_i\right)\right]\mathbb{1}_{M_i(\hat{\theta})>0},$$

$$\hat{\tau}_{\mathrm{t},\hat{\pi}} := \frac{1}{N_1}\sum_{i\in[n]} D_i\left(Y_i - \frac{1}{M_i(\hat{\theta})}\sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j\right)\mathbb{1}_{M_i(\hat{\theta})>0}.$$

3. Asymptotics

In this section, we state our main results: the caliper choice (Section 3.1), the asymptotic normality of the caliper matching estimators of ATE and ATT for known (Section 3.2) and estimated (Section 3.2) propensity scores, and the variance estimation (Section 3.4).

3.1. Caliper Choice

A smaller caliper means that the propensity scores of matched treated and control units are closer, so the match quality is better. At the same time, a smaller caliper leads to fewer matches. Hence, the caliper controls directly the quality and, indirectly, the number of matches, which, in turn, govern the properties of the matching estimator. The match quality determines the bias: comparing dissimilar units threatens the identification of estimands in (1) and (2). The number of matches determines the bias — by excluding units with no matches — as well as the variance of the estimator: since the estimator involves averages over the match set, a small match set gives large variance.

Thus, the right caliper choice must balance the quality and the number of matches. As the sample size increases, we expect that under Assumption 2, we can find both treated and control units in every region of $[\underline{p}, \bar{p}]$ with increasing probability. It is then reasonable to aim for finding matches for each unit in the large sample limit. If we were to set the caliper to $\overline{\underline{\Delta}}_n := \max_{i\in[n]} \min_{j\in[n]:D_j\neq D_i} |\pi(X_i) - \pi(X_j)|$, the largest closest distance between treated and control units, we would have at least one match for each unit. The order of $\mathbb{E}\overline{\underline{\Delta}}_n$ can be concisely described in terms of the sample size, relying on the results of Shorack and Wellner (2009) on spacings (all proofs are presented in Appendix B and in the Supplement).

Proposition 1 (Order of Expected Largest Closest Distance)**.** *Under Assumption 2, there exist constants* $0 < n_0, c < \infty$ *such that* $\mathbb{E}\overline{\underline{\Delta}}_n \leq c\frac{\log n}{n}$ *for all* $n \geq n_0$.

This suggests that the caliper choices, for $n \geq 2$,

$$\delta := \delta_n := s\frac{\log n}{n} \quad \text{or} \quad \delta := \delta_n := \overline{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1} \tag{3}$$

for any fixed constant $s > 0$ are asymptotically of the same order and large enough to guarantee matches for each unit, although the data-dependent choice $\delta_n = \overline{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ can better accommodate smaller samples thus it is generally preferred. Indeed, Proposition 2 shows that, in fact, the implied number of matches is of the order $\log n$.

Proposition 2 (Number of Matches)**.** *Let the caliper satisfy* (3)*. If Assumption 2 holds, then there exist constants* $0 < c_l, c_u < \infty$ *such that*

$$c_l(1 + o_P(1)) \log n \leq \min_{i\in[n]} M_i \leq \max_{i\in[n]} M_i \leq c_u(1 + o_P(1)) \log n$$

as $n \to \infty$*. Thus,* $\mathbb{P}\left(\min_{i\in[n]} M_i \geq 1\right) \to 1$ *as* $n \to \infty$.

3.2. Known Propensity Score

Assume for now that the propensity score $x \mapsto \pi(x)$ is known. We derive the asymptotic distribution of caliper matching in this setting, and show that the caliper choice (3) not only leads to a number of matches increasing in the sample size, but also to the asymptotic unbiasedness of the matching estimator.

In the following, we make a series of assumptions amounting to the asymptotic normality of caliper matching, and we prove that, for instance, the models of Example 1 satisfy these assumptions. Popular models, including the logit and probit for the propensity score and smooth heteroskedastic outcome regressions, are all covered by Example 1 as long as the covariates admit a well-behaved density.[3] The condition of having $K \geq 2$ continuously distributed covariates with nonzero propensity score parameters is not restrictive; for if we had only one, then matching on the propensity score and matching on the covariate would be akin.[4] Regarding other conditions of Example 1, $\nu_d \perp\!\!\!\perp D \mid X$ implies Assumption 1, while differentiability of $x \mapsto \mathbb{E}[\nu_d^2 \mid X = x]$ allows for smooth heteroskedastic models.

Example 1 (Admissible Models)**.** *Let $g : \mathbb{R} \to [0, 1]$ be a strictly increasing function that is twice continuously differentiable on $\mathbb{R}$, with first derivative g' satisfying $\sup_{t \in \mathbb{R}} g'(t) < \infty$. The $(K \geq 2)$-dimensional covariates have density Ψ, which is strictly positive on the compact support $\mathcal{X}$ and continuously differentiable. The propensity score and the potential outcomes satisfy*

$$\begin{aligned} \pi(x) &= g(\theta_0^\intercal x) \\ Y^d &= m_d(X) + \nu_d, \quad \mathbb{E}[\nu_d \mid X] = 0, \quad d \in \{0, 1\}, \end{aligned}$$

where θ_0 is in the interior of $\Theta \subset \mathbb{R}^K$, and it has at least two nonzero coordinates, Θ is bounded, and the m_d are continuously differentiable. For all $d \in \{0, 1\}$, $\nu_d \perp\!\!\!\perp D \mid X$, the $x \mapsto \mathbb{E}[\nu_d^r \mid X = x]$, $r \in \{2, 4\}$, are continuously differentiable on $\mathcal{X}$, and $\inf_{x \in \mathcal{X}} \mathbb{E}[\nu_d^2 \mid X = x] > 0$.

[3] For simplicity of exposition, we assume throughout the paper that X does not include an intercept. The intercept can be accommodated by redefining the distributional assumptions on X to refer to the nonintercept coordinates of X.

[4] Replacing the propensity score with the scalar covariate in Assumptions 2, 4 and 5 would yield a version of Propositions 1 and 2 and Theorems 1 and 2 with the propensity score replaced with the covariate.

We can rewrite $\hat{\tau}_\pi, \hat{\tau}_{t,\pi}$ as weighted averages of the outcome variable Y as follows:

$$\hat{\tau}_\pi = \frac{1}{n}\sum_{i\in[n]}(2D_i - 1)(\mathbb{1}_{M_i>0} + w_i)Y_i, \quad \hat{\tau}_{t,\pi} = \frac{1}{N_1}\sum_{i\in[n]}(\mathbb{1}_{M_i>0}D_i - (1-D_i)w_i)Y_i,$$
$$w_i \coloneqq \sum_{j\in\mathcal{J}(i)}\frac{1}{M_j},$$

where $M_j = 0$ only if $\mathcal{J}(i)$ is empty, in which case the sum in w_i is taken to be zero.[5]

Let $\mu^d(p) \coloneqq \mathbb{E}[Y \mid D = d, \pi(X) = p]$ be the regression function and $\varepsilon \coloneqq Y - \mu^D(\pi(X))$ be the corresponding disturbance term with conditional variance

$$\sigma_d^2(p) \coloneqq \mathbb{V}[\varepsilon \mid D = d, \pi(X) = p] = \mathbb{V}[Y \mid D = d, \pi(X) = p], \quad d \in \{0, 1\}.$$

When we apply caliper matching to imitate (1) and (2), we make two approximations. First, we compare the outcome Y, rather than the regression $\mu^D(\pi(X))$, of the units. The error we make in doing so is ε. Second, we compare units with similar, rather than the same, propensity scores. Therefore, some assumptions must be imposed on the magnitude of ε and the smoothness of μ^d. The magnitude of ε cannot be too large, but also, for convenience, not too small either to avoid degenerate limits. Assumptions 4 and 5 are the same as Assumption 4 in Abadie and Imbens (2006), adapted to matching on the propensity score $\pi(X)$, rather than on the covariates X.

Assumption 4 (Disturbance Term)**.** *(i) The* σ_d^2 *satisfy* $\inf_{d\in\{0,1\},p\in[\underline{p},\bar{p}]}\sigma_d^2(p) > 0$ *and* $\sup_{d\in\{0,1\},p\in[\underline{p},\bar{p}]}\sigma_d^2(p) < \infty$. *(ii)* $\sup_{d\in\{0,1\},p\in[\underline{p},\bar{p}]}\mathbb{E}[\varepsilon^4 \mid D = d, \pi(X) = p] < \infty$.

Assumption 5 (Lipschitz Regression Functions)**.** *The* μ^d *are Lipschitz continuous: there exists a constant* $0 < L_\mu < \infty$ *such that* $|\mu^d(p) - \mu^d(p')| \le L_\mu|p - p'|$ *for all* $p, p' \in [\underline{p}, \bar{p}]$ *for all* $d \in \{0, 1\}$.

Lipschitz continuity guarantees that when the propensity scores $\pi(X_i)$ and $\pi(X_j)$ are close, which we control with δ_n, then so are $\mu^d(\pi(X_i))$ and $\mu^d(\pi(X_j))$. This is in agreement with identification formulae (1) and (2), leading to asymptotic unbiasedness. Similarly to

[5]This follows from the symmetry of caliper matching: $j \in \mathcal{J}(i)$ if and only if $i \in \mathcal{J}(j)$.

Abadie and Imbens (2006), write the ATE estimator as

$$\hat{\tau}_\pi = \overline{\tau(\pi(X))} + E + B, \tag{4}$$

$$\overline{\tau(\pi(X))} := \frac{1}{n}\sum_{i\in[n]} \tau(\pi(X_i)), \quad \tau(\pi(X_i)) := \mu^1(\pi(X_i)) - \mu^0(\pi(X_i)), \tag{5}$$

$$E := \frac{1}{n}\sum_{i\in[n]} E_i, \quad E_i := (2D_i - 1)(\mathbb{1}_{M_i>0} + w_i)\varepsilon_i, \tag{6}$$

$$B := \frac{1}{n}\sum_{i\in[n]} B_i, \tag{7}$$

$$\begin{aligned} B_i := (2D_i - 1)\frac{\mathbb{1}_{M_i>0}}{M_i}\sum_{j\in\mathcal{J}(i)} (\mu^{1-D_i}(\pi(X_i)) - \mu^{1-D_i}(\pi(X_j))) \\ + (2D_i - 1)(\mathbb{1}_{M_i>0} - 1)(\mu^{1-D_i}(\pi(X_i)) - \mu^{D_i}(\pi(X_i))). \end{aligned} \tag{8}$$

The first term $\overline{\tau(\pi(X))}$ has mean τ and the second term E has mean zero. After centering at τ, the first two terms shall be shown to be asymptotically jointly normal and independent at $\sqrt{n}$-rate. The third term B has two sources of bias. The first term in (8) is the bias stemming from imperfect matches. If matches were exact, this term would be zero. By Assumption 5, the magnitude of this term is δ_n, hence it tends to zero even when multiplied with $\sqrt{n}$. The second term in (8) is due to discarding unmatched units, which may happen for the caliper choice $\delta_n = s\frac{\log n}{n}$, unlike for the data-dependent choice $\delta_n = \underline{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$. This leads to a bias because we introduce an artificial sample selection based on δ_n. If every unit had at least one match, as is the case for the data-dependent caliper choice, this term would be zero. But, as shown in Proposition 2, this happens in the large sample limit, giving the asymptotic normality of the ATE estimator $\hat{\tau}_\pi$.

Theorem 1 (Asymptotic Normality for Known Propensity Score (ATE))**.** *Suppose that $x \mapsto \pi(x)$ is known and the caliper δ_n satisfies* (3)*. If Assumptions 1, 2, 4 and 5 all hold, then*

$$\sqrt{n}(\hat{\tau}_\pi - \tau) \rightsquigarrow \mathcal{N}(0, V) \qquad \text{as } n \to \infty,$$

where $V := V_\tau + V_{\sigma,\pi}$ *with* $V_\tau := \mathbb{E}\left[(\tau(\pi(X)) - \tau)^2\right]$ *and* $V_{\sigma,\pi} := \mathbb{E}\left[\frac{\sigma_0^2(\pi(X))}{1-\pi(X)} + \frac{\sigma_1^2(\pi(X))}{\pi(X)}\right]$.

Abadie and Imbens (2006) prove that nearest neighbor matching is asymptotically unbiased only when we match on a scalar covariate. Caliper matching is very much alike. If we were to match on the K-dimensional covariates, similar arguments show that, under regularity conditions, the bias of $\sqrt{n}(\hat{\tau}_\pi - \tau)$ would be of the order $\sqrt{n}(\delta_n + \mathbb{1}_{\{\exists i\in[n]:M_i=0\}})$ and the number of matches would be of the order $n\delta_n^K$. It would then be impossible to have sufficiently good match quality and enough matches at the same time for $K \geq 2$, so the bias B would not vanish. Therefore, it is crucial that we match on the scalar propensity score. When we do so, the ATT estimator $\hat{\tau}_{t,\pi}$ is also asymptotically normal.

Theorem 2 (Asymptotic Normality for Known Propensity Score (ATT)). *Suppose that $x \mapsto \pi(x)$ is known and the caliper δ_n satisfies (3). Let $p_1 := \mathbb{E}\pi(X)$. If Assumptions 1, 2, 4 and 5 all hold, then*

$$\sqrt{n}(\hat{\tau}_{t,\pi} - \tau_t) \rightsquigarrow \mathcal{N}(0, V_t) \qquad \text{as } n \to \infty,$$

where $V_t := V_{\tau_t} + V_{t,\sigma,\pi}$ with $V_{\tau_t} := \frac{1}{p_1^2}\mathbb{E}\left[\pi(X)(\tau(\pi(X)) - \tau_t)^2\right]$ and

$$V_{t,\sigma,\pi} := \frac{1}{p_1^2}\mathbb{E}\left[\frac{\pi(X)^2\sigma_0^2(\pi(X))}{1-\pi(X)} + \pi(X)\sigma_1^2(\pi(X))\right].$$

To examine the efficiency of $\hat{\tau}_\pi$ and $\hat{\tau}_{t,\pi}$, let

$$\mu_{\mathcal{X}}^d(x) := \mathbb{E}\left[Y \mid D = d, X = x\right] \text{ and } \sigma_{\mathcal{X},d}^2(x) := \mathbb{V}\left[Y \mid D = d, X = x\right]$$

for $d \in \{0, 1\}$. The semiparametric efficiency bound of ATE is

$$V_{\text{eff}} := \mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau)^2 + \frac{\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \frac{\sigma_{\mathcal{X},1}^2(X)}{\pi(X)}\right], \tag{9}$$

irrespective of whether or not the propensity scores are known (Hahn (1998, Theorems 1 and 2)). The semiparametric efficiency bound of ATT is

$$V_{t,\text{eff},\pi} := \frac{1}{p_1^2}\mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau_t)^2\pi(X)^2 + \frac{\pi(X)^2\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \pi(X)\sigma_{\mathcal{X},1}^2(X)\right]$$

if the propensity scores are known, and

$$V_{t,\text{eff}} := \frac{1}{p_1^2}\mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau_t)^2\pi(X) + \frac{\pi(X)^2\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \pi(X)\sigma_{\mathcal{X},1}^2(X)\right]$$

if the propensity scores are unknown (Hahn (1998, Theorems 1 and 2)). The limiting variance V of $\hat{\tau}_\pi$ resembles the efficiency bound V_{eff}, except that V involves moments of the outcome conditional on the propensity score $\pi(X)$, rather than on the covariates X as in V_{eff}. Hence, if we were to observe only $\pi(X)$ in our sample, instead of X, $\hat{\tau}_\pi$ would be semiparametrically efficient, reaching V_{eff}. It is also immediate from Theorem 1 and (9), that if we had $\mu_{\mathcal{X}}^d(X) = \mu^d(\pi(X))$ and $\sigma_{\mathcal{X},d}^2(X) = \sigma_d^2(\pi(X))$ for all $d \in \{0, 1\}$ — so that the conditional moments of the outcome given the covariates only depended on the propensity score —, then too, the ATE estimator $\hat{\tau}_\pi$ would be semiparametrically efficient. In truth, a more precise result in Proposition 3 holds.

Proposition 3 (Semiparametric Efficiency). *Suppose that Assumption 1 holds. Then $V_{\mathrm{eff}} \le V$ and $V_{\mathrm{t,eff}} \le V_{\mathrm{t}}$ with equality in both cases if and only if*

$$\mu_{\mathcal{X}}^D(X) = \mu^D(\pi(X)) \quad \textit{almost surely.} \tag{10}$$

Suppose that (10) in Proposition 3 holds. Even then, in contrast to the ATE estimator $\hat{\tau}_\pi$, the ATT estimator $\hat{\tau}_{\mathrm{t},\pi}$ only reaches $V_{\mathrm{t,eff}}$, the semiparametric efficiency bound for *unknown* propensity scores, which is larger than the bound $V_{\mathrm{t,eff},\pi}$ for known propensity scores. The difference between them, under (10), is

$$V_{\mathrm{t,eff}} - V_{\mathrm{t,eff},\pi} = \frac{1}{p_1^2}\mathbb{E}\left[\pi(X)(1-\pi(X))(\tau(\pi(X)) - \tau_{\mathrm{t}})^2\right] \ge 0. \tag{11}$$

As $\pi(X)(1-\pi(X)) \le 1/2$, the difference is bounded by $\frac{1}{2p_1^2}\mathbb{E}\left[(\tau(\pi(X)) - \tau_{\mathrm{t}})^2\right]$. Thus, the more homogeneous the treatment effects are across $\pi(X)$ (equivalently, under (10), across X) and the treatment groups D, the smaller the difference is.

The efficiency loss (11) is not specific to caliper matching. In fact, the limiting variance of the ATT estimator in Theorem 2 is lower than that of the nearest neighbor matching estimator in Abadie and Imbens (2016, Proposition 1). The difference is

$$\frac{1}{2Mp_1^2}\mathbb{E}\left[\sigma_0^2(\pi(X))\pi(X)\left(2 + \frac{\pi(X)}{1-\pi(X)}\right)\right] \ge 0, \tag{12}$$

where the *constant* M is the number of nearest neighbors to match. This shows that the efficiency gain (12) of caliper matching is smaller for larger M. However, there is no proof that letting M to infinity closes the gap as the results of Abadie and Imbens (2016) are contingent on a fixed M. In contrast, with the caliper choice of Theorem 2, the number of matches for caliper matching goes to infinity by Proposition 2, thereby cutting variance. Unless $p \mapsto \sigma_0^2(p)$ decreases rapidly around one, which is ruled out by Assumption 4(i), (12) is larger when the propensity score tends to be close to one. In that case, we gain even more by using caliper instead of nearest neighbor matching, although then $V_{\mathrm{t},\sigma,\pi}$, and thus V_{t}, increases too.

We close the case for the known propensity score by verifying the assumptions of Theorems 1 and 2 for the models of Example 1.

Proposition 4 (Admissible Models (Known Propensity Score)). *The family of models described in Example 1 satisfies all Assumptions 1, 2, 4 and 5.*

3.3. Estimated Propensity Score

Suppose that the propensity score $\pi(\cdot, \theta_0)$ of Assumption 3 is estimated. A reasonable estimator of θ_0 will converge to θ_0. We then expect that if local versions of Assumptions 2, 4 and 5 hold in the neighbourhood of θ_0, then the caliper matching estimators on the estimated propensity scores will also be asymptotically normal, provided they are smooth enough in θ.

To this end, we require the conditional distribution $F_{d,\theta}(p) := \mathbb{P}_{\theta_0}(\pi(X,\theta) \leq p \mid D = d)$ to resemble that of the true propensity score, but only locally. Extending Assumption 2, we need that the densities $f_{0,\theta}, f_{1,\theta}$ are not only continuous but differentiable, and that they depend smoothly on θ. For some arbitrary fixed constant $\epsilon > 0$, let $\mathrm{Nb}(\theta_0, \epsilon) := \{\theta \in \Theta : \|\theta - \theta_0\| < \epsilon\}$ denote a neighbourhood of θ_0, and further let

$$\mathcal{S}_{\theta_0,\epsilon} := \left\{(\theta, p) : p \in [\underline{p}_\theta, \bar{p}_\theta], \theta \in \mathrm{Nb}(\theta_0, \epsilon)\right\}.$$

Assumption 6 (Distribution of the Parametric Propensity Score)**.** *(i) $F_{0,\theta}$, $F_{1,\theta}$ admit densities $f_{0,\theta}, f_{1,\theta}$, respectively, for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$.*
(ii) $f_{0,\theta}, f_{1,\theta}$ have the same support $[\underline{p}_\theta, \bar{p}_\theta]$ with $0 < \underline{p}_\theta < \bar{p}_\theta < 1$ for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$.
(iii) $f_{0,\theta}, f_{1,\theta}$ are bounded away from zero: $\inf_{\theta \in \mathrm{Nb}(\theta_0,\epsilon)} \inf_{p \in [\underline{p}_\theta, \bar{p}_\theta]} f_{d,\theta}(p) > 0$ for all $d \in \{0, 1\}$.
(iv) $(\theta, p) \mapsto f_{d,\theta}(p)$ is continuously differentiable on $\mathcal{S}_{\theta_0,\epsilon}$ for all $d \in \{0, 1\}$.

Next, we decompose the outcome in a way that depends on the propensity score parameter θ. Rather than the continuity of Assumption 5, we need that the regression function $\mu^d(\theta, p) := \mathbb{E}[Y \mid D = d, \pi(X, \theta) = p]$ is continuously differentiable, also in θ. In combination with Assumption 3, Assumption 7 implies that $\theta \mapsto \mu^d(\theta, \pi(x, \theta))$ can be approximated in the neighbourhood of θ_0 with an error of the order $\|\theta - \theta_0\|$. Specifically, they imply that the derivative of $\theta \mapsto \mu^d(\theta, \pi(x, \theta))$ exists for all $(\tilde{\theta}, x) \in \mathrm{Nb}(\theta_0, \epsilon) \times \mathcal{X}$ and it takes the form $\Lambda^d(\tilde{\theta}, x) := \frac{\partial \mu^d}{\partial \theta^\intercal}(\tilde{\theta}, \pi(x, \tilde{\theta})) + \frac{\partial \mu^d}{\partial p}(\tilde{\theta}, \pi(x, \tilde{\theta}))(\mathrm{D}_\theta \pi)(x, \tilde{\theta})$.

Assumption 7 (Differentiability of Regression Functions)**.** *The $(\theta, p) \mapsto \mu^d(\theta, p)$ are continuously differentiable on $\mathcal{S}_{\theta_0,\epsilon}$ with partial derivatives $\frac{\partial \mu^d}{\partial \theta} : \Theta \times [0, 1] \to \mathbb{R}^K$ and $\frac{\partial \mu^d}{\partial p} : \Theta \times [0, 1] \to \mathbb{R}$ uniformly bounded on $\mathcal{S}_{\theta_0,\epsilon}$ for all $d \in \{0, 1\}$.*

To ensure the smoothness, and to control the magnitude of the disturbance term $\varepsilon_i(\theta) := Y_i - \mu^{D_i}(\theta, \pi(X_i, \theta)), i \in [n]$, we require that the functions

$$\sigma_d^r(\theta, p) := \mathbb{E}\left[(Y - \mu^D(\theta, p))^r \mid D = d, \pi(X, \theta) = p\right], \quad r \in \{2, 4\}, d \in \{0, 1\},$$

satisfy the following conditions.

Assumption 8 (Smooth Parametric Disturbance Term)**.** *(i) The σ_d^2 satisfy the Lipschitz-condition $|\sigma_d^2(\theta, p) - \sigma_d^2(\theta', p')| \leq L_\sigma(\|\theta - \theta'\| + |p - p'|)$ for all $p \in [\underline{p}_\theta, \bar{p}_\theta]$ and $p' \in [\underline{p}_{\theta'}, \bar{p}_{\theta'}]$ for all $\theta, \theta' \in \mathrm{Nb}(\theta_0, \epsilon)$ for some constant $0 < L_\sigma < \infty$ and the lower bound $\inf_{p \in [\underline{p}_{\theta_0}, \bar{p}_{\theta_0}]} \sigma_d^2(\theta_0, p) > 0$ for all $d \in \{0, 1\}$. (ii) The σ_d^4 satisfy the condition $\sup_{\theta \in \mathrm{Nb}(\theta_0,\epsilon)} \sup_{p \in [\underline{p}_\theta, \bar{p}_\theta]} \sigma_d^4(\theta, p) < \infty$ for all $d \in \{0, 1\}$.*

Finally, we need that the estimator $\hat{\theta}$ of the propensity score parameter converges to θ_0 in an appropriate sense. For instance, if $\hat{\theta}$ is the maximum likelihood estimator, it converges appropriately under regularity conditions. We further assume that θ_0 is estimated from a sample that is independent of $((Y_i, D_i, X_i))_{i\in[n]}$. In practice, sample splitting may be applied to ensure the independence: one can halve a $2n$-large sample and use the first half to estimate θ_0, and plug the resulting estimator $\hat{\theta}$ back into the second half to compute $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{\mathrm{t},\hat{\pi}}$.

Assumption 9 (Estimator of the Propensity Score Parameter)**.** *(i) $\hat{\theta}$ is asymptotically normal with $\sqrt{n}(\hat{\theta}-\theta_0) \rightsquigarrow \mathcal{N}(0, V_{\theta_0})$ as $n \to \infty$ for a finite invertible matrix V_{θ_0}. (ii) $\hat{\theta}$ is independent of the data set from which the matching estimator is computed: $\hat{\theta} \perp\!\!\!\perp ((Y_i, X_i, D_i))_{i\in[n]}$.*

To accommodate the propensity score estimation, we introduce

$$\widehat{\overline{\underline{\Delta}}}_n := \max_{i\in[n]} \min_{j\in[n]:D_j\neq D_i} |\pi(X_i, \hat{\theta}) - \pi(X_j, \hat{\theta})|,$$

the estimated analogue of $\overline{\underline{\Delta}}_n$, and the corresponding caliper choices

$$\delta_n := s\frac{\log n}{n} \quad \text{or} \quad \delta_n := \widehat{\overline{\underline{\Delta}}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1} \tag{13}$$

for any fixed constant $s > 0$. Proposition 5 shows that the number of matches based on the estimated propensity scores and the caliper choice (13) is also of the order $\log n$ as in Proposition 2. This yields Theorems 3 and 4, establishing the asymptotic normality of caliper matching on the estimated propensity score.

Proposition 5 (Number of Matches for Estimated Propensity Score)**.** *Suppose that the caliper δ_n satisfies (13). If Assumptions 6 and 9 hold, then there exist constants $0 < \bar{c}_l, \bar{c}_u < \infty$ such that*

$$\bar{c}_l(1 + o_P(1)) \log n \leq \min_{i\in[n]} M_i(\hat{\theta}) \leq \max_{i\in[n]} M_i(\hat{\theta}) \leq \bar{c}_u(1 + o_P(1)) \log n$$

as $n \to \infty$. Thus, $\mathbb{P}\left(\min_{i\in[n]} M_i(\hat{\theta}) \geq 1\right) \to 1$ as $n \to \infty$.

Theorem 3 (Asymptotic Normality for Estimated Propensity Score (ATE))**.** *Suppose that the caliper δ_n satisfies (13). If Assumptions 1, 3 and 6 to 9 all hold, then*

$$\sqrt{n}(\hat{\tau}_{\hat{\pi}} - \tau) \rightsquigarrow \mathcal{N}(0, V_{\hat{\pi}}) \qquad \text{as } n \to \infty,$$

where $V_{\hat{\pi}} := V_\tau + V_{\sigma,\pi} + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$ for $V_\tau, V_{\sigma,\pi}$ of Theorem 1, V_{θ_0} of Assumption 9, and $q_d \in \mathbb{R}^K$ arising as the probability limit $\frac{1}{n}\sum_{i\in[n]} \Lambda^d(\hat{\theta}, X_i) \xrightarrow{P} q_d^\intercal$ as $n \to \infty$ for $d \in \{0, 1\}$.

Theorem 4 (Asymptotic Normality for Estimated Propensity Score (ATT)). *Suppose that the caliper δ_n satisfies (13). If Assumptions 1, 3 and 6 to 9 all hold, then*

$$\sqrt{n}(\hat{\tau}_{t,\hat{\pi}} - \tau_t) \rightsquigarrow \mathcal{N}(0, V_{t,\hat{\pi}}) \qquad \text{as } n \to \infty,$$

where $V_{t,\hat{\pi}} := V_{\tau_t} + V_{t,\sigma,\pi} + (1/p_1^2)(q_{t,1} - q_{t,0})^\intercal V_{\theta_0}(q_{t,1} - q_{t,0})$ for V_{τ_t}, $V_{t,\sigma,\pi}$ and p_1 of Theorem 2, V_{θ_0} of Assumption 9, and $q_{t,d} \in \mathbb{R}^K$ arising as the probability limit $\frac{1}{n}\sum_{i\in[n]} D_i \Lambda^d(\hat{\theta}, X_i) \xrightarrow{P} q_{t,d}^\intercal$ as $n \to \infty$ for $d \in \{0, 1\}$.

Compared to Theorems 1 and 2, the variances are increased by $(q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$ for the ATE estimator $\hat{\tau}_{\hat{\pi}}$, and by $(1/p_1^2)(q_{t,1} - q_{t,0})^\intercal V_{\theta_0}(q_{t,1} - q_{t,0})$ for the ATT estimator $\hat{\tau}_{t,\hat{\pi}}$, representing the uncertainty from the propensity score estimation. The more precisely we can estimate the propensity score, the smaller V_{θ_0} is, resulting in smaller differences. Alternatively, if $q_1 \approx q_0$ or $q_{t,1} \approx q_{t,0}$, then the respective increments are also small. This is the case if the derivatives Λ^1 and Λ^0 are close to each other, although it is difficult to see if and when that happens, even for simple linear regressions in Example 1. As a consequence, it remains unclear whether caliper or nearest neighbor matching (Abadie and Imbens, 2016) is more efficient when the parametric propensity score is estimated.

Remark 1 (Variance Comparison). *The asymptotic variances of $\hat{\tau}_\pi, \hat{\tau}_{t,\pi}$ and $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{t,\hat{\pi}}$ are comparable as in the preceding paragraph if and only if we use only half of a $2n$-large sample to compute $\hat{\tau}_\pi, \hat{\tau}_{t,\pi}$ and the caliper (3), because of the sample-splitting in the computation of $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{t,\hat{\pi}}$ and (13). If we use the whole $2n$-large sample to compute $\hat{\tau}_\pi, \hat{\tau}_{t,\pi}$, (3), and only n observations to evaluate $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{t,\hat{\pi}}$, (13) — with the remaining n observations reserved to estimate θ_0 —, then the standard error of $\hat{\tau}_\pi$ is $\sqrt{\frac{V_\tau + V_{\sigma,\pi}}{2n}}$, while that of $\hat{\tau}_{\hat{\pi}}$ is $\sqrt{\frac{V_\tau + V_{\sigma,\pi} + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)}{n}}$, which is an increment by a factor of $\sqrt{2}$ even without the contribution of $(q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$. The same applies to $\hat{\tau}_{t,\pi}$ and $\hat{\tau}_{t,\hat{\pi}}$.*

Abadie and Imbens (2016) account for the estimation of the propensity score by considering a shifted law of $(Y, D, X) \sim \mathbb{P}_{\theta_0}$. They assume that conditional expectations under the shifted law converge weakly to conditional expectations under the nonshifted law. We pursue a different approach. Our Assumptions 6 to 8 do not involve shifted laws. Rather, they impose smoothness of conditional expectations in θ and may be regarded as local versions of Assumptions 2, 4 and 5 in the neighbourhood of θ_0. Moreover, we verify the assumptions of Theorems 3 and 4 for the models in Example 1.

Proposition 6 (Admissible Models (Estimated Propensity Score)). *Consider the family of models described in Example 1, with the propensity score model $\{\pi(x, \theta) = g(\theta^\intercal x) : \theta \in$*

$\Theta\}$ *estimated with maximum likelihood on an independent n-large i.i.d. sample from the distribution of (D, X). Then Assumptions 1, 3 and 6 to 9 are all satisfied.*

3.4. Variance Estimation

In this section, we provide consistent estimators for the components of $V_{\hat{\pi}}$ and $V_{\mathrm{t},\hat{\pi}}$ so that we can construct asymptotically valid confidence intervals for ATE and ATT. To prove consistency, we impose some further assumptions, which are all in accordance with the models in Example 1.

Namely, we need that certain estimators are almost surely bounded, which is implied if the outcome is almost surely bounded. Furthermore, $\theta \mapsto \pi(\cdot, \theta)$ may take many forms in general, which renders Λ^d intractable. Requiring that the propensity score follows a single-index model, such as the logit or the probit, and that the covariates have a well-behaved density, alleviates these difficulties, provided the outcome regression is smooth enough. Imposing $K \geq 2$ continuously distributed covariates and certain smoothness conditions implies that Λ^d is expressible in a way suitable for showing consistency.

Assumption 10 (Outcome and Covariate Distribution)**.** *(i) The outcome is almost surely bounded: there exists a constant $0 < \bar{y} < \infty$ such that $\mathbb{P}(|Y| > \bar{y}) = 0$. (ii) The covariate vector X has at least $K \geq 2$ coordinates, and X admits a density Ψ on the compact $\mathcal{X}$; the Ψ is as specified in Example 1.*

Assumption 11 (Single-Index Propensity Score and Smooth Outcome Regression)**.** *(i) The propensity score model of Assumption 3 is $\pi(x, \theta) = g(\theta^\intercal x)$ for g as specified in Example 1. (ii) The $m(x) := \mathbb{E}[Y \mid X = x]$ is bounded, and there exist two covariates — X_1 and X_2 without loss of generality — such that $\frac{\partial m}{\partial x_1}$ and $\frac{\partial m}{\partial x_2}$ are well-defined and continuous for all $x \in \mathcal{X}$.*

The variance estimators are

$$\hat{V}_{\hat{\pi}} := \hat{V}_{\tau} + \hat{V}_{\sigma,\pi} + (\hat{q}_1 - \hat{q}_0)^\intercal \hat{V}_{\theta_0}(\hat{q}_1 - \hat{q}_0), \tag{14}$$

$$\hat{V}_{\mathrm{t},\hat{\pi}} := \hat{V}_{\tau_{\mathrm{t}}} + \hat{V}_{\mathrm{t},\sigma,\pi} + (1/\hat{p}_1^2)(\hat{q}_{\mathrm{t},1} - \hat{q}_{\mathrm{t},0})^\intercal \hat{V}_{\theta_0}(\hat{q}_{\mathrm{t},1} - \hat{q}_{\mathrm{t},0}), \tag{15}$$

where the component estimators are as follows. We assume that $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ is a consistent estimator of V_{θ_0}. In practice, $\hat{\theta}$ is usually the maximum likelihood estimator, as supported by Proposition 6, in which case, under Assumption 11,

$$\hat{V}_{\theta_0} := \left(\frac{1}{n} \sum_{i \in [n]} \frac{(g'(\hat{\theta}^\intercal X_i))^2}{g(\hat{\theta}^\intercal X_i)(1 - g(\hat{\theta}^\intercal X_i))} X_i X_i^\intercal \right)^{-1}$$

is well-known to be consistent for V_{θ_0}. The p_1 is consistently estimated with $\hat{p}_1 := \frac{1}{n}\sum_{i\in[n]} D_i$ by the law of large numbers. The nonparametric estimators of the remaining components in (14) and (15) are developed in Appendix A.

Proposition 7 (Consistent Variance Estimators). *Suppose that Assumptions 1, 3 and 6 to 11 all hold, that the caliper δ_n satisfies (13), and that $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ as $n \to \infty$. Then $\hat{V}_{\hat{\pi}} \xrightarrow{P} V_{\hat{\pi}}$ and $\hat{V}_{\mathrm{t},\hat{\pi}} \xrightarrow{P} V_{\mathrm{t},\hat{\pi}}$ as $n \to \infty$. In particular, the estimators on the right side of (14) and (15) are all consistent for their respective estimands.*

In view of Theorems 3 and 4, an immediate implication is that we can construct asymptotic confidence intervals for ATE and ATT. Let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$th quantile of the standard normal distribution for $\alpha \in (0,1)$, and let $[a \pm b]$ denote the interval $[a-b, a+b]$ for $a, b \in \mathbb{R}$, $b \geq 0$. Then the intervals $[\hat{\tau}_{\hat{\pi}} \pm z_{1-\alpha/2}(\hat{V}_{\hat{\pi}}/n)^{1/2}]$ and $[\hat{\tau}_{\mathrm{t},\hat{\pi}} \pm z_{1-\alpha/2}(\hat{V}_{\mathrm{t},\hat{\pi}}/n)^{1/2}]$ are asymptotically valid confidence intervals for ATE and ATT, respectively.

Corollary 1 (Asymptotic Confidence Intervals). *Suppose that Assumptions 1, 3 and 6 to 11 all hold, that the caliper δ_n satisfies (13), and that $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ as $n \to \infty$. Then $\mathbb{P}\left(\left[\hat{\tau}_{\hat{\pi}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}_{\hat{\pi}}}{n}}\right] \ni \tau\right) \to 1-\alpha$ and $\mathbb{P}\left(\left[\hat{\tau}_{\mathrm{t},\hat{\pi}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}_{\mathrm{t},\hat{\pi}}}{n}}\right] \ni \tau_{\mathrm{t}}\right) \to 1-\alpha$ as $n \to \infty$.*

4. Conclusion

We study the caliper matching estimator when matching is performed on the (estimated) propensity scores. We propose a caliper, and prove that the resulting estimator of the Average Treatment Effect (ATE), and of the Average Treatment Effect on the Treated (ATT), is asymptotically unbiased and normal.

When the propensity score is known, our estimator of ATE reaches the semiparametric lower bound in the restricted model where only the propensity scores and not the covariates are observed in the sample or where the outcome regression on the covariates only depend on the propensity score. In this restricted model, the estimator of ATT only reaches the larger lower bound corresponding to unknown propensity score. Even in the unrestricted model, both our estimators are more efficient than nearest neighbor matching estimators on the known propensity scores, and are, therefore, preferred over the latter method in the large sample limit, provided our assumptions hold. When the parametric propensity score is estimated, the variances of both our estimators increase, hence it remains unclear whether caliper or nearest neighbor matching will be more efficient.

We facilitate empirical application of the estimator by verifying our assumptions for a family of often employed models, and by constructing asymptotic confidence intervals for the average treatment effects. An interesting avenue for future research is to study in-sample estimation of the propensity score, and to allow for nonparametric propensity score estimators. The main challenge arising is to see how uncertainty from the propensity score estimation propagates to the matching estimator, which is more difficult to quantify for nonparametric models.

A. Variance Estimation

In this section, we define the nonparametric variance estimators of Section 3.4 for the components in (14) and (15). Let $K(u) = (2\pi)^{-1/2}e^{-u^2/2}, u \in \mathbb{R}$, be the Gaussian kernel and $K'(u)$ its derivative. Let $0 < \gamma_n \lesssim a_n$ be two arbitrary sequences $\gamma_n := \kappa_0 n^{-\beta}$, $a_n := \kappa_1 n^{-\alpha}$ for fixed finite constants $0 < \alpha < \beta < 1/4$ and $\kappa_0, \kappa_1 > 0$. We employ a truncation strategy to avoid bias at the boundaries. Define the intervals $A_n := [\underline{p}_{\hat{\theta}} + a_n, \bar{p}_{\hat{\theta}} - a_n]$ and $\hat{A}_n := [\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) + a_n, \max_{i\in[n]} g(\hat{\theta}^\intercal X_i) - a_n]$, which are well-defined with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$ under Assumption 9; see the proof of Proposition 7.[6] Let $\hat{N} := \sum_{i\in[n]} \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}$. The estimators of the first components are

$$\hat{V}_\tau := \left(\frac{1}{\hat{N}} \sum_{i\in[n]} [\hat{\mu}^1(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) - \hat{\mu}^0(\hat{\theta}, g(\hat{\theta}^\intercal X_i))]^2 \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n} \right) - \hat{\tau}^2_{\hat{\pi}}, \tag{16}$$

$$\hat{V}_{\tau_t} := \frac{1}{\hat{p}_1^2} \left(\frac{1}{\hat{N}} \sum_{i\in[n]} D_i[\hat{\mu}^1(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) - \hat{\mu}^0(\hat{\theta}, g(\hat{\theta}^\intercal X_i))]^2 \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n} \right) - \frac{\hat{\tau}^2_{t,\hat{\pi}}}{\hat{p}_1},$$

where

$$\hat{\mu}^d(\theta, p) := \frac{\hat{q}_{\mu,d}(\theta, p)}{\hat{h}_d(\theta, p)}, \quad \hat{q}_{\mu,d}(\theta, p) := \frac{1}{N_d\gamma_n} \sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j K\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n} \right),$$

$$\hat{h}_d(\theta, p) := \hat{f}_{\theta,d}(p) := \frac{1}{N_d\gamma_n} \sum_{j\in[n]} \mathbb{1}_{D_j=d} K\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n} \right), \quad d \in \{0, 1\};$$

and those of the second components are

$$\hat{V}_{\sigma,\pi} := \frac{1}{\hat{N}} \sum_{i\in[n]} \left(\frac{\hat{\sigma}_0^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{1 - g(\hat{\theta}^\intercal X_i)} + \frac{\hat{\sigma}_1^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{g(\hat{\theta}^\intercal X_i)} \right) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$

$$\hat{V}_{t,\sigma,\pi} := \frac{1}{\hat{p}_1^2\hat{N}} \sum_{i\in[n]} \left(\frac{g(\hat{\theta}^\intercal X_i)^2 \hat{\sigma}_0^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{1 - g(\hat{\theta}^\intercal X_i)} + g(\hat{\theta}^\intercal X_i)\hat{\sigma}_1^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) \right) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$

[6] In practice, especially for moderate sample sizes, γ_n and a_n should be chosen carefully to ensure non-negative variance estimates. The a_n should be chosen small enough to enlarge $A_n, \hat{A}_n$; for instance, one could choose κ_1 arbitrary close to zero and $\alpha := 1/(4 + \varepsilon_\alpha)$ for an $\varepsilon_\alpha > 0$ arbitrarily close to zero. As a rule, γ_n should be set small too to minimise the bias of the variance estimates by standard nonparametric theory, thereby avoiding negative values; to accommodate $\alpha < \beta$, one can set $\beta = 1/(4 + \varepsilon_\beta)$ with $0 < \varepsilon_\beta < \varepsilon_\alpha$, for example, $\varepsilon_\beta := \varepsilon_\alpha/2$. The κ_0 should be chosen to accommodate the different scales of $(g(\hat{\theta}^\intercal X_i))_{i\in[n]}$ and $(\hat{\theta}^\intercal X_i)_{i\in[n]}$ present in the estimation of the μ^d and their derivate, respectively. A small κ_0 is a safe but conservative choice. Note that asymptotically the effect of truncation disappears $(\mathbb{E}\hat{N}/n \to 1)$ as shown in Proposition 7.

where

$$\hat{\sigma}^2_d(\theta,p) \coloneqq \hat{\mu}^d_2(\theta,p) - (\hat{\mu}^d(\theta,p))^2, \quad \hat{\mu}^d_2(\theta,p) \coloneqq \frac{\hat{q}_{\mu_2,d}(\theta,p)}{\hat{h}_d(\theta,p)},$$
$$\hat{q}_{\mu_2,d}(\theta,p) \coloneqq \frac{1}{N_d\gamma_n}\sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j^2 K\left(\frac{g(\theta^\intercal X_j)-p}{\gamma_n}\right), \quad d\in\{0,1\},$$

is an estimator of $\sigma^2_d(\theta,p) = \mu^d_2(\theta,p) - (\mu^d(\theta,p))^2$ with

$$\mu^d_2(\theta,p) \coloneqq \mathbb{E}\left[Y^2 \,\middle|\, D=d, \pi(X,\theta)=p\right], \quad d\in\{0,1\}.$$

Last, the probability limits of the derivatives are estimated by

$$\hat{q}_d^\intercal \coloneqq \frac{1}{\hat{N}}\sum_{i\in[n]} \hat{\Lambda}^d(\hat{\theta},X_i)\mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}, \quad \hat{q}_{t,d}^\intercal \coloneqq \frac{1}{\hat{N}}\sum_{i\in[n]} D_i\hat{\Lambda}^d(\hat{\theta},X_i)\mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$
$$\hat{\Lambda}^d(\theta,x) \coloneqq \widehat{\left(\frac{\partial\mu^d}{\partial\theta^\intercal}\right)}(\theta,g(\theta^\intercal x)) + \widehat{\left(\frac{\partial\mu^d}{\partial p}\right)}(\theta,g(\theta^\intercal x))g'(\theta^\intercal x)x^\intercal,$$

where

$$\widehat{\left(\frac{\partial\mu^d}{\partial\theta_k}\right)}(\theta,p) \coloneqq \frac{\widehat{\left(\frac{\partial q_{\mu,d}}{\partial\theta_k}\right)}(\theta,p)\hat{h}_d(\theta,p) - \hat{q}_{\mu,d}(\theta,p)\widehat{\left(\frac{\partial h_d}{\partial\theta_k}\right)}(\theta,p)}{(\hat{h}_d(\theta,p))^2},$$
$$\widehat{\left(\frac{\partial q_{\mu,d}}{\partial\theta_k}\right)}(\theta,p) \coloneqq \frac{(g^{-1})'(p)}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j X_{j,k} K'\left(\frac{\theta^\intercal X_j - g^{-1}(p)}{\gamma_n}\right),$$
$$\widehat{\left(\frac{\partial h_d}{\partial\theta_k}\right)}(\theta,p) \coloneqq \frac{(g^{-1})'(p)}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} X_{j,k} K'\left(\frac{\theta^\intercal X_j - g^{-1}(p)}{\gamma_n}\right),$$

with θ_k ($X_{j,k}$) being the kth coordinate of θ (X_j) for $k\in[K]$, and

$$\widehat{\left(\frac{\partial\mu^d}{\partial p}\right)}(\theta,p) \coloneqq \frac{\left(\frac{\partial}{\partial p}\hat{q}_{\mu,d}(\theta,p)\right)\hat{h}_d(\theta,p) - \hat{q}_{\mu,d}(\theta,p)\frac{\partial}{\partial p}\hat{h}_d(\theta,p)}{(\hat{h}_d(\theta,p))^2},$$
$$\frac{\partial}{\partial p}\hat{q}_{\mu,d}(\theta,p) = -\frac{1}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j K'\left(\frac{g(\theta^\intercal X_j)-p}{\gamma_n}\right),$$
$$\frac{\partial}{\partial p}\hat{h}_d(\theta,p) = -\frac{1}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} K'\left(\frac{g(\theta^\intercal X_j)-p}{\gamma_n}\right)$$

for $d\in\{0,1\}$. An intercept in the propensity score model can be accommodated by defining $X_{j,K+1} \coloneqq 1$ for $j\in[n]$ and considering derivatives with respect to θ_{K+1} too.

B. Proofs

In this section, we prove the main results, Propositions 1, 2 and 4 to 6 and Theorems 1 to 4 together with supporting Lemmas 1 to 4. The proofs of Propositions 3 and 7 and Lemmas 5 to 7 are in the Supplement. For simplicity, we give the proofs for the caliper choice $\delta_n = s\frac{\log n}{n}$, $s := 1$, and provide remarks for the choices $\delta_n = \overline{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ and $\delta_n = \widehat{\overline{\underline{\Delta}}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ when necessary.

We adopt the following notation. Let $D^{(n)} := (D_i)_{i\in[n]}$ and $PS^{(n)} := (\pi(X_i))_{i\in[n]}$. For $a, b \in \mathbb{R}, a < b$, let

$$F_d[a,b] := \mathbb{P}\left(\pi(X) \in [a,b] \mid D = d\right),$$
$$\mathbb{F}_{N_d}[a,b] := \frac{1}{N_d}\sum_{i:D_i=d} \mathbb{1}_{\pi(X_i)\in[a,b]}, \tag{17}$$

be the conditional (empirical) measures of intervals $[a,b]$ for $d \in \{0,1\}$. Similarly, under Assumption 3, define the conditional (empirical) measures

$$F_{d,\theta}[a,b] := \mathbb{P}\left(\pi(X,\theta) \in [a,b] \mid D = d\right),$$
$$\mathbb{F}_{N_d,\theta}[a,b] := \frac{1}{N_d}\sum_{i:D_i=d} \mathbb{1}_{\pi(X_i,\theta)\in[a,b]} \qquad \text{for } \theta \in \Theta. \tag{18}$$

Let $\mathbb{G}_{N_d} := \sqrt{N_d}(\mathbb{F}_{N_d} - F_d)$ be the empirical process of $((\pi(X_i))_{i:D_i=d} \mid D^{(n)}) \overset{\text{i.i.d.}}{\sim} F_d$, and $[a \pm b]$ denote the interval $[a-b, a+b]$. In the proofs, the value of constants may change from equation to equation without explicit notice.

Proof of Proposition 1. Spacings and Their Order. Let $U_{(1)} \le U_{(2)} \le \ldots \le U_{(N_1)}$ be the order statistics of $(U_1, \ldots, U_{N_1} \mid D^{(n)}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$. Let $\tilde{U}_1 := U_{(1)}$, $\tilde{U}_i := U_{(i)} - U_{(i-1)}$ for $i = 2, \ldots, N_1$ and $\tilde{U}_{N_1+1} := 1 - U_{(N_1)}$ be the spacings generated by $(U_i)_{i\in[N_1]}$. Let $\tilde{U}_{(1)} \le \tilde{U}_{(2)} \le \ldots \le \tilde{U}_{(N_1+1)}$ be the ordered spacings. Shorack and Wellner (2009, Chapter 21) prove that $\mathbb{E}\left[\tilde{U}_{(N_1+1)} \,\middle|\, D^{(n)}\right] = \frac{1}{N_1+1}\sum_{i=1}^{N_1+1}\frac{1}{N_1+2-i}$, which we apply as follows.

Bounding $\mathbb{E}\overline{\underline{\Delta}}_n$ with Spacings. Let $\underline{\Delta}_i := \min_{j:D_j\neq D_i} |\pi(X_i) - \pi(X_j)|$ for $i \in [n]$, so that $\overline{\underline{\Delta}}_n \le \sum_{d\in\{0,1\}} \max_{i:D_i=d} \underline{\Delta}_i$. Consider $\max_{i:D_i=0} \underline{\Delta}_i$. Let $\pi_{(1)} \le \pi_{(2)} \le \ldots \le \pi_{(N_1)}$ be the order statistics of $(\pi(X_i))_{i:D_i=1}$, and let $\tilde{\pi}_1 := \pi_{(1)} - \underline{p}$, $\tilde{\pi}_i := \pi_{(i)} - \pi_{(i-1)}$ for $i \in \{2, \ldots, N_1\}$ and $\tilde{\pi}_{N_1+1} := \bar{p} - \pi_{(N_1)}$ be the corresponding spacings for $\underline{p}, \bar{p}$ of Assumption 2. Let $\tilde{\pi}_{(1)} \le \tilde{\pi}_{(2)} \le \ldots \le \tilde{\pi}_{(N_1+1)}$ be the order statistics of these spacings. Every propensity score $\pi(X_j)$ of the control units falls either to the left of $\pi_{(1)}$ or to the right of $\pi_{(N_1)}$ or between two propensity scores $\pi_{(i)}$ and $\pi_{(i-1)}$ for some $i \in \{2, 3, \ldots, N_1\}$. In all three cases, the closest treated propensity score to $\pi(X_j)$ is within $\tilde{\pi}_{(N_1+1)}$-distance. Hence, $\max_{i:D_i=0} \underline{\Delta}_i \leq \tilde{\pi}_{(N_1+1)}$.

Given $D^{(n)}$, the $\pi(X_i)$ restricted to $i : D_i = 1$ are i.i.d., with $(\pi(X_i) \mid D_i = 1) \sim F_1$. By Assumption 2, $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$, thus F_1 is strictly increasing on $[\underline{p}, \bar{p}]$, and therefore has a strictly increasing inverse F_1^{-1} on $[F_1(\underline{p}), F_1(\bar{p})]$. Because F_1^{-1} is increasing, $((\pi_{(i)})_{i\in[N_1]} \mid D^{(n)}) \sim ((F_1^{-1}(U_{(i)}))_{i\in[N_1]} \mid D^{(n)})$ by the quantile transform. We distinguish three cases.

- Case 1: $\tilde{\pi}_{(N_1+1)} = \pi_{(i)} - \pi_{(i-1)}$ for some $i \in \{2, \dots, N_1\}$. Then we bound $\tilde{\pi}_{(N_1+1)}$ by noting that $(\pi_{(i)} - \pi_{(i-1)} \mid D^{(n)}) \sim (F_1^{-1}(U_{(i)}) - F_1^{-1}(U_{(i-1)}) \mid D^{(n)})$, which is bounded by $\left\|(F_1^{-1})'\right\|_\infty (U_{(i)} - U_{(i-1)}) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_i$, because F_1^{-1} is Lipschitz with constant $\left\|(F_1^{-1})'\right\|_\infty$ with $(F_1^{-1})'(u) = \frac{1}{f_1(F_1^{-1}(u))}$ finite as $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$ by Assumption 2.
- Case 2: $\tilde{\pi}_{(N_1+1)} = \pi_{(1)} - \underline{p}$. Then write $\underline{p} = F_1^{-1}(F_1(\underline{p})) = F_1^{-1}(0)$, so $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty (U_{(1)} - 0) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_1$.
- Case 3: $\tilde{\pi}_{(N_1+1)} = \bar{p} - \pi_{(N_1)}$. Then write $\bar{p} = F_1^{-1}(F_1(\bar{p})) = F_1^{-1}(1)$, so $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty (1 - U_{(N_1)}) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_{N_1+1}$.

Conclude that $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty \tilde{U}_{(N_1+1)}$. Thus,

$$\begin{aligned}
\mathbb{E} \max_{i:D_i=0} \underline{\Delta}_i &\leq \mathbb{E}\mathbb{E}\left[\tilde{\pi}_{(N_1+1)} \mid D^{(n)}\right] \lesssim \mathbb{E}\left[\frac{1}{N_1+1}\sum_{i=1}^{N_1+1} \frac{1}{N_1+2-i}\right] \\
&= \sum_{n_1=0}^{n}\left[\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i}\right]\binom{n}{n_1} p_1^{n_1}(1-p_1)^{n-n_1} \\
&= (1-p_1)^n + \sum_{n_1=1}^{n}\left[\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i}\right]\binom{n}{n_1} p_1^{n_1}(1-p_1)^{n-n_1} \qquad (19)
\end{aligned}$$

by Shorack and Wellner (2009) where $p_1 = \mathbb{P}(D = 1)$. The first term in (19) decays exponentially. In the second term of (19), the integrand in the square brackets is asymptotic to $\frac{\log n_1}{n_1+1} \leq \frac{\log n}{n_1+1}$. That is, there exist some constants $c, \bar{n}_1 > 0$ such that $\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i} \leq c\frac{\log n}{n_1+1}$ if $n_1 > \bar{n}_1$. Since the integrand in the square brackets is bounded by one, it follows that the second term in (19) is bounded by

$$\sum_{n_1=1}^{\bar{n}_1}\binom{n}{n_1} p_1^{n_1}(1-p_1)^{n-n_1} + c(\log n)\mathbb{E}\left[(1+N_1)^{-1}\right],$$

where the first term is $O\left(n^{\bar{n}_1}(1-p_1)^n\right) = O\left(\log n/n\right)$ and the second term is $O\left(\log n/n\right)$ by Cribari-Neto et al. (2000). Similar arguments hold for $\mathbb{E}\max_{i:D_i=1}\underline{\Delta}_i$ by symmetry. ■

Proof of Proposition 2. We have for the R_{di} in (32) of Lemma 1,

$$\min_{i\in[n]} M_i = \min_{i\in[n]} N_{1-D_i}F_{1-D_i}[\pi(X_i)\pm\delta_n](1+R_{1-D_i,i}). \tag{20}$$

If $\min_{i\in[n]}(1+R_{1-D_i,i})\geq 0$, which happens with probability tending to one, then (20) is larger than or equal to

$$2\left(\min_{d\in\{0,1\}}\inf_{p\in[\underline{p},\bar{p}]} f_d(p)\right)(N_0\wedge N_1)\delta_n\min_{i\in[n]}(1+R_{1-D_i,i}).$$

By Assumption 2, $\inf_{p\in[\underline{p},\bar{p}]} f_d(p)>0$. By the strong law of large numbers and the continuous mapping theorem, $(1/\log n)(N_0\wedge N_1)\delta_n = (N_0\wedge N_1)/n \xrightarrow{a.s.} (1-p_1)\wedge p_1 > 0$. By Lemma 1, $\max_{i\in[n]}|R_{1-D_i,i}| = o_P(1)$. Then $\min_{i\in[n]} M_i \simeq (1+o_P(1))\log n$, from which the lower bound in Proposition 2, and thus $\mathbb{P}\left(\min_{i\in[n]} M_i\geq 1\right)\to 1$, follows. The same reasoning applies to the upper bound in Proposition 2.

When the caliper is $\delta_n = \overline{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1}\vee\frac{\log N_1}{N_1+1}$, $\min_{i\in[n]} M_i\geq 1$. Lemma 1(iv)–(vi), the continuous mapping theorem, combined with the law of large numbers and Proposition 1 prove the assertion. ■

Proof of Proposition 5. Follows along arguments in the proof of Proposition 2 and Lemma 1 (iv)–(vi). When the caliper is $\delta_n = \widehat{\overline{\underline{\Delta}}}_n \vee \frac{\log N_0}{N_0+1}\vee\frac{\log N_1}{N_1+1}$, it is bounded by $\widehat{\overline{\underline{\Delta}}}_n + \frac{\log N_0}{N_0+1} + \frac{\log N_1}{N_1+1}$, where $\mathbb{E}\frac{\log N_0}{N_0+1} = O\left(\frac{\log n}{n}\right)$ by Cribari-Neto et al. (2000). A spacings argument on $(\pi(X_i,\hat{\theta}))_{i\in[n]}$, similarly to the proof of Proposition 1, combined with Assumptions 6 and 9 yields $\widehat{\overline{\underline{\Delta}}}_n = O_P\left(\frac{\log n}{n}\right)$. Then arguments in Proposition 2 and Lemma 1(iv)–(vi) prove Proposition 5. ■

Proof of Theorem 1. By Assumption 2, $\tau(\pi(X))$ of (5) is well-defined, with $\mathbb{E}\tau(\pi(X))=\tau$ by Assumption 1. By Assumption 5, the μ^d are Lipschitz continuous on the compact set $[\underline{p},\bar{p}]$, hence are bounded, and then so is $V_\tau<\infty$. Then the central limit theorem implies $\sqrt{n}(\overline{\tau(\pi(X))}-\tau)\rightsquigarrow\mathcal{N}(0,V_\tau)$. Combine this with Lemma 2, to get

$$\begin{bmatrix} V_\tau^{-1/2}\sqrt{n}(\overline{\tau(\pi(X))}-\tau) \\ V_E^{-1/2}\sqrt{n}E \end{bmatrix} \rightsquigarrow \mathcal{N}(0,\mathrm{I}_2),$$

along subsequences, where I_2 is the 2-by-2 identity matrix. By Lemma 4, $V_E\xrightarrow{P} V_{\sigma,\pi}$, which is finite by Assumptions 2 and 4. Then the continuous mapping theorem and Slutsky's lemma imply $(V_\tau+V_{\sigma,\pi})^{-1/2}\sqrt{n}(\overline{\tau(\pi(X))}-\tau+E)\rightsquigarrow\mathcal{N}(0,1)$. The event $\{\min_{i\in[n]} M_i >$

$0\}$ happens with probability tending to one by Proposition 2. On this event, $\sqrt{n}|B| \lesssim \sqrt{n}\delta_n = \frac{\log n}{\sqrt{n}} = o(1)$ by Assumption 5. Thus $\sqrt{n}B = o_P(1)$.

When the caliper is $\delta_n = \underline{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, Lemma 2 continues to apply. Then the continuous mapping theorem, the law of large numbers and Proposition 1 imply $\sqrt{n}|B| = o_P(1)$. ■

Proof of Theorem 2. A decomposition similar to (4)–(8) holds, whereby

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_{\mathrm{t},\pi} - \tau_{\mathrm{t}}) &= \sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}}) + \sqrt{n}E_{\mathrm{t}} + \sqrt{n}B_{\mathrm{t}} \\
\overline{\tau_{\mathrm{t}}(\pi(X))} &:= \frac{1}{N_1}\sum_{i\in[n]} D_i\tau(\pi(X_i)) \\
E_{\mathrm{t}} &:= \frac{1}{N_1}\sum_{i\in[n]} E_{\mathrm{t},i}, \quad E_{\mathrm{t},i} := (\mathbb{1}_{M_i>0}D_i - (1-D_i)w_i)\varepsilon_i \\
B_{\mathrm{t}} &:= \frac{1}{N_1}\sum_{i\in[n]} B_{\mathrm{t},i}, \\
B_{\mathrm{t},i} &:= D_i(\mathbb{1}_{M_i>0} - 1)(\mu^{D_i}(\pi(X_i)) + \mu^{1-D_i}(\pi(X_i))) \\
&\quad + D_i\frac{\mathbb{1}_{M_i>0}}{M_i}\sum_{j\in\mathcal{J}(i)}(\mu^0(\pi(X_i)) - \mu^0(\pi(X_j))).
\end{aligned}$$

As $\mathbb{E}D(\tau(\pi(X)) - \tau_{\mathrm{t}}) = \mathbb{E}\left[\tau(\pi(X)) - \tau_{\mathrm{t}} \mid D = 1\right]p_1 = 0$, $\sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}})$ is mean zero with finite variance by Assumptions 1 and 5. By the central limit theorem, continuous mapping and Slutsky's lemma, $\sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}}) = (N_1/n)^{-1}n^{-1/2}\sum_{i\in[n]} D_i(\tau(\pi(X_i)) - \tau_{\mathrm{t}}) \rightsquigarrow \mathcal{N}(0, V_{\tau_{\mathrm{t}}})$ since $(N_1/n) \xrightarrow{a.s.} p_1$. The $\sqrt{n}E_{\mathrm{t}}$ has mean zero and a Lindeberg-Feller central limit theorem establishes that $\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(V_{E_{\mathrm{t}}}^{-1/2}\sqrt{n}E_{\mathrm{t}} \le x \;\middle|\; D^{(n)}, PS^{(n)}\right) - \Phi(x)\right| \xrightarrow{P} 0$, where $V_{E_{\mathrm{t}}} := \frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0}D_i - (1-D_i)w_i)^2\sigma^2_{D_i}(\pi(X_i))$, similarly to arguments in Lemma 2. By arguments similar to those of Lemma 4, $V_{E_{\mathrm{t}}} \xrightarrow{P} V_{\mathrm{t},\sigma,\pi}$. By Proposition 2 and arguments in Theorem 1, $\sqrt{n}B_{\mathrm{t}} = o_P(1)$. ■

Proof of Theorem 3. Let $w_i(\hat{\theta}) := \sum_{j\in\mathcal{J}_{\hat{\theta}}(i)}\frac{1}{M_j(\hat{\theta})}$, $i \in [n]$, and, as in (4)–(8), decompose

$$\sqrt{n}(\hat{\tau}_{\hat{\pi}} - \tau) = \frac{1}{\sqrt{n}}\sum_{i\in[n]}\left\{\mu^1(\hat{\theta}, \pi(X_i,\hat{\theta})) - \mu^0(\hat{\theta}, \pi(X_i,\hat{\theta})) - \tau\right\} \tag{21}$$

$$+\frac{1}{\sqrt{n}}\sum_{i\in[n]}(2D_i - 1)(\mathbb{1}_{M_i(\hat{\theta})>0} + w_i(\hat{\theta}))\varepsilon_i(\hat{\theta}) \tag{22}$$

$$+\frac{1}{\sqrt{n}}\sum_{i\in[n]}(2D_i - 1)(\mathbb{1}_{M_i(\hat{\theta})>0} - 1)(\mu^{1-D_i}(\hat{\theta}, \pi(X_i,\hat{\theta})) - \mu^{D_i}(\hat{\theta}, \pi(X_i,\hat{\theta}))) \tag{23}$$

$$+\frac{1}{\sqrt{n}}\sum_{i\in[n]}(2D_i - 1)\frac{\mathbb{1}_{M_i(\hat{\theta})>0}}{M_i(\hat{\theta})}\sum_{j\in\mathcal{J}_{\hat{\theta}}(i)}\left[\mu^{1-D_i}(\hat{\theta}, \pi(X_i,\hat{\theta})) - \mu^{1-D_i}(\hat{\theta}, \pi(X_j,\hat{\theta}))\right]. \tag{24}$$

We show first that (21) and (22) are, asymptotically, jointly normal and independent, and then that (23) and (24) are asymptotically negligible.

Terms (21) *and* (22). We apply the following result with (21) and (22) corresponding to V_n and W_n respectively. Let V_n, W_n, $n = 1, 2, \ldots$ be two sequences of random variables defined on some probability space. To show that $(V_n, W_n) \rightsquigarrow (V, W) \sim \mathcal{N}(0, \Sigma)$, for a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_V^2, \sigma_W^2)$, it suffices that $\mathbb{E}h_1(V_n)h_2(W_n) \to (\mathbb{E}h_1(V))(\mathbb{E}h_2(W))$ for all bounded continuous functions $h_1,\ h_2 : \mathbb{R} \to \mathbb{R}$. Let $\mathcal{F}_{n0}$ be a sub-σ-algebra such that V_n is $\mathcal{F}_{n0}$-measurable for all $n \geq 1$. As $\mathbb{E}h_1(V_n)h_2(W_n) = \mathbb{E}\left[h_1(V_n)\mathbb{E}\left[h_2(W_n) \mid \mathcal{F}_{n0}\right]\right]$, if suffices, by the Portmanteau lemma, that $V_n \rightsquigarrow \mathcal{N}(0, \sigma_V^2)$ and $\mathbb{P}\left(W_n \leq w \mid \mathcal{F}_{n0}\right) \xrightarrow{P} \Phi(w/\sigma_W)$ for all $w \in \mathbb{R}$.

Convergence of (21). Expand (21) as

$$\frac{1}{\sqrt{n}} \sum_{i \in [n]} (\mu^1_{\theta_0}(\pi(X_i, \theta_0)) - \mu^0_{\theta_0}(\pi(X_i, \theta_0)) - \tau) \tag{25}$$

$$+\frac{1}{\sqrt{n}} \sum_{i \in [n]} \Big\{ \mu^1(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^0(\hat{\theta}, \pi(X_i, \hat{\theta})) - \left[\mu^1_{\theta_0}(\pi(X_i, \theta_0)) - \mu^0_{\theta_0}(\pi(X_i, \theta_0))\right] \Big\}. \tag{26}$$

By Assumptions 1 and 5, (25) converges weakly to $\mathcal{N}(0, V_\tau)$ by the standard central limit theorem. By Assumptions 3 and 7, $\Lambda^d(\tilde{\theta}, x)$ is well-defined for all $(\tilde{\theta}, x) \in \mathrm{Nb}(\theta_0, \epsilon) \times \mathcal{X}$. Then by the mean-value theorem, $\mu^d(\hat{\theta}, \pi(x, \hat{\theta})) = \mu^d_{\theta_0}(\pi(x, \theta_0)) + \Lambda^d(\tilde{\theta}^d, x)(\hat{\theta} - \theta_0)$, for some $\tilde{\theta}^d$ on the line segment between $\hat{\theta}$ and θ_0. Rewrite (26) as

$$\left(\frac{1}{n} \sum_{i \in [n]} \left[\Lambda^1(\tilde{\theta}^1, X_i) - \Lambda^0(\tilde{\theta}^0, X_i)\right] \right) \sqrt{n}(\hat{\theta} - \theta_0). \tag{27}$$

By Assumptions 3 and 7, $\tilde{\theta} \mapsto \Lambda^d(\tilde{\theta}, x)$ is continuous and uniformly bounded for all $(\tilde{\theta}, x) \in \mathrm{Nb}(\theta_0, \epsilon) \times \mathcal{X}$, therefore $\frac{1}{n} \sum_{i \in [n]} \Lambda^d(\tilde{\theta}^d, X_i) \xrightarrow{P} q_d^\intercal$ for some finite $q_d \in \mathbb{R}^K$. Then (27) and Slutsky's lemma imply that (26) converges weakly to $\mathcal{N}(0, (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0))$ by Assumption 9. But, by Assumption 9 again, (25) is independent of $\sqrt{n}(\hat{\theta} - \theta_0)$, thus (21), being the sum of (25) and (26), converges weakly to $\mathcal{N}(0, V_\tau + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0))$.

Conditional Convergence of (22). Let $\mathcal{F}_{n0} \coloneqq \sigma\{D_1, \ldots, D_n, \pi(X_1, \hat{\theta}), \ldots, \pi(X_n, \hat{\theta}), \hat{\theta}\}$, so that (21) is $\mathcal{F}_{n0}$-measurable. We show that given $\mathcal{F}_{n0}$, (22) converges weakly to a normal variate in probability. We construct a martingale array and apply Lemma 7. Let $\xi_{ni} \coloneqq (2D_i - 1)(1 + w_i(\hat{\theta}))\varepsilon_i(\hat{\theta})/\sqrt{n}$ and

$$\mathcal{F}_{ni} \coloneqq \sigma\{D_1, \ldots, D_n, \pi(X_1, \hat{\theta}), \ldots, \pi(X_n, \hat{\theta}), \varepsilon_1(\hat{\theta}), \ldots, \varepsilon_i(\hat{\theta}), \hat{\theta}\}$$

for $i \in [n]$. Assume temporarily that $\min_{i \in [n]} M_i(\hat{\theta}) > 0$, so that (22) is equal to $\sum_{i=1}^n \xi_{ni}$. One can verify that $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$ are martingale differences relative to the filtration

$\mathcal{F}_{n1} \subset \mathcal{F}_{n2} \subset \ldots \subset \mathcal{F}_{nn}$, using that $\mu^{D_i}(\hat{\theta}, \pi(X_i, \hat{\theta}))$ is $\mathcal{F}_{n,i-1}$-measurable and Assumption 9(ii), which implies that the observations are i.i.d. given $\hat{\theta}$.

First, we verify the variance condition (56) of Lemma 7. Consider $\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \mid \mathcal{F}_{n,i-1}\right]$. We have

$$\mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right] = (1 + w_i(\hat{\theta}))^2 \sigma^2_{D_i}(\hat{\theta}, \pi(X_i, \hat{\theta}))/n,$$

where we used Assumption 9(ii) again, and that $w_i(\hat{\theta})$ is $\mathcal{F}_{n,i-1}$-measurable. But then $\mathbb{E}\left[\xi_{ni}^2 \mid \mathcal{F}_{n,i-1}\right]$ is $\mathcal{F}_{n0}$-measurable for all $i \in [n]$ for all $n \geq 1$, thus for condition (56) it suffices that $\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \mid \mathcal{F}_{n,i-1}\right]$ converges in $\mathbb{P}$-probability to a finite constant. Assumption 3(iii) implies that $\max_{i\in[n]} |\pi(X_i, \hat{\theta}) - \pi(X_i, \theta_0)| = O_P\left(\|\hat{\theta} - \theta_0\|\right)$. Then, under Assumption 8, we can write $\sigma_d^2(\hat{\theta}, \pi(X_i, \hat{\theta})) = \sigma_d^2(\theta_0, \pi(X_i, \theta_0)) + S_i$, where $\max_{i\in[n]} |S_i| = O_P\left(\|\hat{\theta} - \theta_0\|\right)$. Write

$$\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right] = \frac{1}{n}\sum_{i\in[n]} (1 + w_i(\hat{\theta}))^2 \sigma^2_{D_i}(\theta_0, \pi(X_i, \theta_0)) + \frac{1}{n}\sum_{i\in[n]} (1 + w_i(\hat{\theta}))^2 S_i. \quad (28)$$

Under Assumptions 3, 6 and 9, the arguments in the proof of Lemma 4 continue to apply in view of Lemma 1(iv)–(vi). Then for any sequence $(Q_i)_{i\in[n]}$ of random variables and any fixed constant $r \in \mathbb{R}$, we have, by Lemma 1(iv)–(vi),

$$\begin{aligned}\sum_{i\in[n]} w_i(\hat{\theta})^r Q_i = (1 + o_P(1)) \Bigg\{ \left(\frac{N_1}{N_0}\right)^r \sum_{i:D_i=0} \left(\frac{f_{1,\hat{\theta}}(\pi(X_i, \hat{\theta}))}{f_{0,\hat{\theta}}(\pi(X_i, \hat{\theta}))}\right)^r Q_i \\ + \left(\frac{N_0}{N_1}\right)^r \sum_{i:D_i=1} \left(\frac{f_{0,\hat{\theta}}(\pi(X_i, \hat{\theta}))}{f_{1,\hat{\theta}}(\pi(X_i, \hat{\theta}))}\right)^r Q_i \Bigg\}. \quad (29)\end{aligned}$$

The second term in (28) is bounded by

$$O_P\left(\|\hat{\theta} - \theta_0\|\right)\left(1 + \frac{1}{n}\sum_{i\in[n]} w_i(\hat{\theta})^2\right).$$

Assumption 6, bounding the ratios $f_{d,\theta}(p)/f_{1-d,\theta}(p)$ uniformly in $p \in [\underline{p}_\theta, \bar{p}_\theta]$ and $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$, combined with (29) and $(N_d/N_{1-d})^r \xrightarrow{a.s.} \left(\frac{p_d}{1-p_d}\right)^r$, where $p_0 := 1 - p_1$, bounds $\frac{1}{n}\sum_{i\in[n]} w_i(\hat{\theta})^2$ by a $(1 + o_P(1))$-term up to a constant factor. Thus, the second term in (28) is $O_P\left(\|\hat{\theta} - \theta_0\|\right) = o_P(1)$ under Assumption 9. Put $Q_i := \sigma^2_{D_i}(\theta_0, \pi_i)$ and apply a mean-value expansion in θ around θ_0 to the right side of (29). This is feasible under Assumptions 3 and 6, which also bound the derivative uniformly in $(x, \tilde{\theta}) \in \mathcal{X} \times \mathrm{Nb}(\theta_0, \epsilon)$. Then Assumptions 8 and 9 imply that the first term in (28) is $V_{\sigma,\pi} + o_P(1)$ under Assumption 3, as in the proof of Lemma 4, where $V_{\sigma,\pi}$ is finite by Assumptions 6 and 8. Conclude that (56) holds.

Second, we verify the Lindeberg-condition (57) of Lemma 7. We need to show

$$\sum_{i=1}^{n} \mathbb{E}\left[\xi_{ni}^2 \mathbb{1}_{|\xi_{ni}|\geq\eta} \,\middle|\, \mathcal{F}_{n0}\right] \xrightarrow{P} 0 \quad \text{for each } \eta > 0.$$

As $\mathbb{1}_{|\xi_{ni}|\geq\eta}$ is bounded by ξ_{ni}^2/η^2,

$$\sum_{i=1}^{n} \mathbb{E}\left[\xi_{ni}^2 \mathbb{1}_{|\xi_{ni}|\geq\eta} \,\middle|\, \mathcal{F}_{n0}\right] \leq \sum_{t=1}^{n} \frac{\mathbb{E}\left[\xi_{ni}^4 \,\middle|\, \mathcal{F}_{n0}\right]}{\eta^2} = \frac{1}{n\eta^2} \sum_{i\in[n]} \frac{(1+\tilde{w}_i(\hat{\theta}))^4 \mathbb{E}\left[\varepsilon_i(\hat{\theta})^4 \,\middle|\, \mathcal{F}_{n0}\right]}{n},$$

with $\mathbb{E}\left[\varepsilon_i(\hat{\theta})^4 \,\middle|\, \mathcal{F}_{n0}\right] = \sigma_{D_i}^4(\hat{\theta}, \pi(X_i, \hat{\theta}))$ by Assumption 9(ii), which is bounded uniformly in $i \in [n]$ by Assumption 8. In view of (29), $\frac{1}{n^2}\sum_{i\in[n]}(1+\tilde{w}_i(\hat{\theta}))^4 = o_P(1)$, so (57) is met.

Conclude that under the temporary assumption $\min_{i\in[n]} M_i(\hat{\theta}) > 0$, Lemma 7 applies, so (22) converges weakly to $\mathcal{N}(0, V_{\sigma,\pi})$ in probability. To remove this assumption, define the $\mathcal{F}_{n0}$-measurable set $A_n := \{\min_{i\in[n]} M_i(\hat{\theta}) > 0\}$. On A_n, (22) is equal to $\sum_{i\in[n]} \xi_{ni}$. As $\mathbb{P}(A_n) \to 1$ by Proposition 5, the desired convergence follows.

Vanishing (23). By Assumption 7, the $p \mapsto \mu^d(\theta, p)$ are continuous on a compact set $[\underline{p}_\theta, \bar{p}_\theta]$ for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$. By Assumption 9, $\mathbb{P}\left(\hat{\theta} \in \mathrm{Nb}(\theta_0, \epsilon)\right) \to 1$. By Proposition 5, $\mathbb{P}(A_n) \to 1$, thus (23) is $o_P(1)$.

Vanishing (24). By Assumption 7, $|\mu^d(\theta, p') - \mu^d(\theta, p)| \leq L|p' - p|$ for all $p', p \in [\underline{p}_\theta, \bar{p}_\theta]$ for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$. Then

$$\max_{i\in[n]} \max_{j\in\mathcal{J}_{\hat{\theta}}(i)} |\mu^{1-D_i}(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^{1-D_i}(\hat{\theta}, \pi(X_j, \hat{\theta}))| \lesssim \max_{i\in[n]} \max_{j\in\mathcal{J}_{\hat{\theta}}(i)} |\pi(X_i, \hat{\theta}) - \pi(X_j, \hat{\theta})|.$$

By the construction of $\mathcal{J}_{\hat{\theta}}(i)$, the right side is bounded by δ_n, where $\sqrt{n}\delta_n \to 0$.

When the caliper is $\delta_n = \widehat{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, the above arguments continue to hold. Specifically, in showing the conditional convergence of (22), $w_i(\hat{\theta})$ is still $\mathcal{F}_{n,i-1}$-measurable; (23) is exactly zero, and, in (24), $\sqrt{n}\delta_n \leq \sqrt{n}\left(\widehat{\underline{\Delta}}_n + \frac{\log n}{N_0+1} + \frac{\log n}{N_1+1}\right) = o_P(1)$ in view of the proof of Proposition 5. ∎

Proof of Theorem 4. Follows those of Theorem 2 and Theorem 3. ∎

Proof of Proposition 6. Assumptions 1, 3 and 9(ii) hold by construction as $\mathcal{X}$ is bounded. The assumptions of Example 1 imply Assumption 9(i) by standard asymptotic theory (Van der Vaart, 1998). We assume $K = 2$ covariates in the following so that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$; the general case $K \geq 2$ follows analogously. First, we establish some general results. Let θ_k ($\theta_{0,k}$) denote the kth entry of θ (θ_0). For $t \in \mathcal{T} := \{\theta^\intercal x : \theta \in \Theta, x \in \mathcal{X}\}$, $\theta^\intercal X$ has density $f_{\theta^\intercal X}(t) = \int_{\mathcal{X}_1} \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1$, which is strictly positive by assumptions on Ψ, the density of X.[7] Let $h : \mathcal{X} \to \mathbb{R}^J$, $J \geq 1$, be an arbitrary integrable function. We have

$$\mathbb{E}\left[h(X_1, X_2) \mid \theta^\intercal X = t\right] = \frac{1}{f_{\theta^\intercal X}(t)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1. \tag{30}$$

Combine this with the tower property of expectations to get the conditional density

$$f_{\theta^\intercal X|D}(t \mid d) = \begin{cases} \frac{1}{1-p_1} \int_{\mathcal{X}_1} (1 - g_{\theta_0}(x_1, \theta, t)) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 0, \\ \frac{1}{p_1} \int_{\mathcal{X}_1} g_{\theta_0}(x_1, \theta, t) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 1, \end{cases}$$

where $g_{\theta_0}(x_1, \theta, t) := g\left(\theta_{0,1} x_1 + \theta_{0,2} \frac{t-\theta_1 x_1}{\theta_2}\right) \in (0,1)$ as $0 < g(t') < 1$ for all t' in the bounded $\mathcal{T}$. Hence, $f_{\theta^\intercal X|D}$ is strictly positive. It is also continuously differentiable in t by assumptions on g and Ψ. Moreover, $\theta \mapsto f_{\theta^\intercal X|D}(t \mid d)$ is continuously differentiable at $\tilde{\theta} \in \mathrm{Nb}(\theta_0, \epsilon)$ with bounded derivative for a $\theta_{0,2} \neq 0$ as $\sup_{t \in \mathbb{R}} g'(t) < \infty$ and the derivatives of Ψ are bounded. One can also show that $\mathbb{E}\left[h(X_1, X_2) \mid D = d, \theta^\intercal X = t\right]$ is

$$\begin{cases} \frac{1-p_1}{f_{\theta^\intercal X|D}(t \mid 0)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) (1 - g_{\theta_0}(x_1, \theta, t))\, \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 0, \\ \frac{p_1}{f_{\theta^\intercal X|D}(t \mid 1)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) g_{\theta_0}(x_1, \theta, t) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 1, \end{cases} \tag{31}$$

which is continuously differentiable in t and θ by assumptions on g, Ψ, and the properties of $f_{\theta^\intercal X|D}$ derived above, provided h is continuously differentiable. We are now ready to verify the remaining assumptions.

Assumption 6. The distributions are

$$F_{d,\theta}(p) = \begin{cases} \frac{1}{1-p_1} \int_{\mathcal{X}} \mathbb{1}_{g(\theta^\intercal x) \leq p} (1 - g(\theta_0^\intercal x)) \Psi(x) \mathrm{d}x & \text{if } d = 0, \\ \frac{1}{p_1} \int_{\mathcal{X}} \mathbb{1}_{g(\theta^\intercal x) \leq p}\, g(\theta_0^\intercal x) \Psi(x) \mathrm{d}x & \text{if } d = 1. \end{cases}$$

Since g is increasing and $\mathcal{T}_\theta := \{\theta^\intercal x : x \in \mathcal{X}\}$ is compact, for all $\theta \in \Theta$ there exist $0 < \underline{p}_\theta < \bar{p}_\theta < 1$ such that $F_{d,\theta}(\underline{p}_\theta) = 0$ and $F_{d,\theta}(\bar{p}_\theta) = 1$. Specifically, $\underline{p}_\theta = g(\inf \mathcal{T}_\theta)$ and $\bar{p}_\theta = g(\sup \mathcal{T}_\theta)$. On $[\underline{p}_\theta, \bar{p}_\theta]$, the $F_{d,\theta}$ admit densities

$$f_{d,\theta}(p) = f_{\theta^\intercal X|D}(g^{-1}(p) \mid d)(g^{-1})'(p) = \frac{f_{\theta^\intercal X|D}(g^{-1}(p) \mid d)}{g'(g^{-1}(p))},$$

[7] If X included an intercept, then we would have $f_{\theta^\intercal X}(t) = \int_{\mathcal{X}_1} \Psi\left(x_1, \frac{t-\theta_3-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1$, where θ_3 is the coefficient on the intercept. Below, the right side of (30), $f_{\theta^\intercal X|D}(t \mid D)$ and (31) would need to be adjusted in a similar manner to accommodate an intercept.

where g^{-1} is well-defined, as well as its derivative by the inverse function theorem. Then the assumptions on g and the properties of $f_{\theta^\intercal X|D}$ imply that Assumption 6 holds.

Assumption 7. Under Assumption 1, the tower property of expectation gives

$$\mu^d(\theta, p) = \mathbb{E}\left[Y \mid D = d, \theta^\intercal X = g^{-1}(p)\right] = \mathbb{E}\left[m_d(X) \mid D = d, \theta^\intercal X = g^{-1}(p)\right].$$

But then the properties of (31), g and m_d imply that Assumption 7 holds.

Assumption 8. The lower bound in (i) is satisfied by assumptions on ν_d in Example 1. The Lipschitz condition in (i) and (ii) both follow by the mean-value theorem as the σ_d^2 and σ_d^4 are polynomials in terms of the form $\mathbb{E}\left[h(X_1, X_2) \mid D = d, \theta^\intercal X = g^{-1}(p)\right]$ for continuously differentiable functions h by assumptions on m_d and ν_d, so (31) applies. ■

Proof of Proposition 4 . Follows that of Proposition 6. ■

Lemma 1 (Convergence of Ratios)**.** *Suppose that the caliper δ_n satisfies (3). For the measures in (17), define*

$$R_{di} := \frac{\mathbb{F}_{N_d}[\pi(X_i) \pm \delta_n]}{F_d[\pi(X_i) \pm \delta_n]} - 1, \tag{32}$$

$$\tilde{R}_{di} := \frac{F_d[\pi(X_i) \pm \delta_n]}{2\delta_n f_d(\pi(X_i))} - 1, \tag{33}$$

$$\check{R}_{dji} := \frac{f_d(\pi(X_j))}{f_d(\pi(X_i))} - 1 \tag{34}$$

for $j \in \mathcal{J}(i)$, $i \in \{i \in [n] : D_i = 1 - d\}$ and $d \in \{0, 1\}$. If Assumption 2 holds, then

(i) $\max_{i:D_i=1-d} |R_{di}| = o_P(1)$*; and*

(ii) $\max_{i:D_i=1-d} |\tilde{R}_{di}| = o_P(1)$*; and*

(iii) $\max_{i:D_i=1-d} \max_{j\in\mathcal{J}(i)} |\check{R}_{dji}| = o_P(1)$

as $n \to \infty$ for all $d \in \{0, 1\}$. Suppose that Assumption 3 holds and the caliper δ_n satisfies (13). For the measures in (18), define

$$R^\theta_{di} := \frac{\mathbb{F}_{N_d,\theta}[\pi(X_i, \theta) \pm \delta_n]}{F_{d,\theta}[\pi(X_i, \theta) \pm \delta_n]} - 1, \tag{35}$$

$$\tilde{R}^\theta_{di} := \frac{F_{d,\theta}[\pi(X_i, \theta) \pm \delta_n]}{2\delta_n f_{d,\theta}(\pi(X_i, \theta))} - 1, \tag{36}$$

$$\check{R}^\theta_{dji} := \frac{f_{d,\theta}(\pi(X_j, \theta))}{f_{d,\theta}(\pi(X_i, \theta))} - 1 \tag{37}$$

for $j \in \mathcal{J}_\theta(i)$, $i \in \{i \in [n] : D_i = 1 - d\}$, $d \in \{0, 1\}$ and $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$. If Assumption 6 holds, then for any $\hat{\theta}$ satisfying Assumption 9, *(iv)* $\max_{i:D_i=1-d} |R^{\hat\theta}_{di}| = o_P(1)$*; and*

(v) $\max_{i:D_i=1-d} |\tilde R^{\hat\theta}_{di}| = o_P(1)$*; and*

(vi) $\max_{i:D_i=1-d} \max_{j\in\mathcal{J}_{\hat\theta}(i)} |\check R^{\hat\theta}_{dji}| = o_P(1)$

as $n\to\infty$ *for all* $d\in\{0,1\}$.

Proof. Assertion (i). Consider

$$|R_{1i}| \le \sup_{p\in[\underline{p},\bar p]} \left|\frac{\mathbb{F}_{N_1}[p\pm\delta_n]}{F_1[p\pm\delta_n]} - 1\right| = \sup_{p\in[\underline{p},\bar p]} \frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{\sqrt{N_1}F_1[p\pm\delta_n]} =: W_1. \tag{38}$$

Fix a constant $\zeta>0$. We bound $\mathbb{P}\left(W_1>\zeta \mid D^{(n)}\right)$ using ratio and tail bounds of empirical processes. To this end, note that for any finite $\delta_n>0$, $\mathcal{C}_{\delta_n} := \{[p\pm\delta_n] : p\in[\underline{p},\bar p]\}$ is a VC-class, with VC-dimension equal to two (e.g. Van der Vaart and Wellner (1996, Example 2.6.1)). Let $\gamma_n := 2\delta_n \inf_{p\in[\underline{p},\bar p]} f_1(p)$, so that $\inf_{p\in[\underline{p},\bar p]} F_1[p\pm\delta_n]\ge\gamma_n$. The event $\{W_1>\zeta\}$ is equal to

$$\left\{\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]} : p\in[\underline{p},\bar p], F_1[p\pm\delta_n]\ge\gamma_n\right\} > \sqrt{N_1}\zeta\right\} \subset$$

$$\left(\left\{\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]} : p\in[\underline{p},\bar p], F_1[p\pm\delta_n]\ge\gamma_n, F_1[p\pm\delta_n]\le\frac{1}{2}\right\} > \sqrt{N_1}\zeta\right\}\cup\right. \tag{39}$$

$$\left.\left\{\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]} : p\in[\underline{p},\bar p], F_1[p\pm\delta_n]\ge\gamma_n, F_1[p\pm\delta_n]>\frac{1}{2}\right\} > \sqrt{N_1}\zeta\right\}\right). \tag{40}$$

First, we bound the probability of the event in (39). Take $A := \{p\in[\underline{p},\bar p] : F_1[p\pm\delta_n]\ge\gamma_n\}$, $B := \{p\in[\underline{p},\bar p] : F_1[p\pm\delta_n]\le\frac{1}{2}\}$. If $\gamma_n>\frac{1}{2}$, $A\cap B$ is empty and by the convention $\sup\emptyset=-\infty$, the set (39) has measure zero. So assume without loss of generality that $\gamma_n\le F_1[p\pm\delta_n]\le\frac{1}{2}$. Thus, $\frac{\gamma_n}{2}\le\sigma_1^2[p\pm\delta_n] := (F_1[p\pm\delta_n])(1-F_1[p\pm\delta_n])\le\frac{1}{4}$, where note that $\sigma_1^2[p\pm\delta_n] < F_1[p\pm\delta_n]$ for $F_1[p\pm\delta_n]\ge\gamma_n>0$. As $N_1\gamma_n \xrightarrow{a.s.} \infty$ and $N_1^{-1}\log(e\vee\log(e\vee N_1)) = o(\gamma_n)$ a.s., conditions (2.2)–(2.3) in Alexander (1987, Theorem 2.1) hold a.s.. Therefore, the bounds in the proof of Theorem 5.1 therein apply. Hence, as in (7.66)–(7.67) of Alexander (1987),

$$\mathbb{P}\left(\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]} : p\in[\underline{p},\bar p], F_1[p\pm\delta_n]\ge\gamma_n, F_1[p\pm\delta_n]\le\frac{1}{2}\right\} > \sqrt{N_1}\zeta \;\middle|\; D^{(n)}\right)$$

$$\le \mathbb{P}\left(|\mathbb{G}_{N_1}[p\pm\delta_n]| > (\sigma_1^2[p\pm\delta_n])\sqrt{N_1}\zeta \text{ for some } p\in[\underline{p},\bar p] : \frac{\gamma_n}{2}\le\sigma_1^2[p\pm\delta_n]\le\frac{1}{4} \;\middle|\; D^{(n)}\right)$$

$$\le 36\int_{\gamma_n/2}^{1/4} t^{-1}e^{-\zeta^2N_1t/512}\mathrm{d}t + 68e^{-\zeta N_1\gamma_n/256}$$

$$\le \frac{36}{\zeta^2N_1\gamma_n}e^{-\zeta^2N_1\gamma_n/1024} + 68e^{-\zeta N_1\gamma_n/256}.$$

Second, the probability of the event in (40) is bounded by

$$\mathbb{P}\left(\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]}: p\in[\underline{p},\bar{p}], F_1[p\pm\delta_n]>\frac{1}{2}\right\}>\sqrt{N_1}x\,\middle|\, D^{(n)}\right)$$
$$\leq \mathbb{P}\left(\sup\left\{|\mathbb{G}_{N_1}[p\pm\delta_n]|: p\in[\underline{p},\bar{p}], F_1[p\pm\delta_n]>\frac{1}{2}\right\}>\frac{\sqrt{N_1}}{2}x\,\middle|\, D^{(n)}\right)$$
$$\leq \mathbb{P}\left(\sup_{p\in[\underline{p},\bar{p}]}|\mathbb{G}_{N_1}[p\pm\delta_n]|>\frac{\sqrt{N_1}}{2}x\,\middle|\, D^{(n)}\right) \tag{41}$$

as the supremum over a larger set cannot decrease. Van der Vaart and Wellner (1996, Theorem 2.14.9) bound (41) by $cN_1\zeta^2 e^{-\frac{N_1\zeta^2}{2}}$. Therefore,

$$\mathbb{P}\left(W_1>\zeta \mid D^{(n)}\right)\leq \frac{c_1}{\zeta^2 N_1\gamma_n}e^{-\zeta^2 N_1\gamma_n/1024}+c_2 e^{-\zeta N_1\gamma_n/256}+c_3 N_1\zeta^2 e^{-N_1\zeta^2/2}, \tag{42}$$

on the set where $N_0, N_1\geq 1$, which happens with probability tending to one. The left side is bounded by one, the right side converges to zero in probability; then the left side also converges to zero in expectation. Similar arguments hold for $W_0 := \sup_{p\in[\underline{p},\bar{p}]}\frac{|\mathbb{G}_{N_0}|[p\pm\delta_n]|}{\sqrt{N_0}F_0[p\pm\delta_n]}$, bounding $\max_{i:D_i=1}|R_{0i}|$. Conclude that $\max_{i:D_i=1-d}|R_{di}| = o_P(1)$. When the caliper is $\delta_n = \underline{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1}\vee\frac{\log N_1}{N_1+1}$, the same arguments yield the assertion by letting $\gamma_n := \gamma_{N_d} := 2\inf_{p\in[\underline{p},\bar{p}]} f_d(p)\frac{\log N_d}{N_d+1}$ when bounding W_d, $d\in\{0,1\}$.

Assertion (ii). By Assumption 2, the f_d are continuous on the compact set $[\underline{p},\bar{p}]$, hence are uniformly continuous. By the mean-value theorem, uniform continuity of f_d implies uniform differentiability of F_d. By uniform differentiability of F_0, $\sup_{j:D_j=1}|F_0[\pi(X_j)\pm\delta_n]-2\delta_n f_0(\pi(X_j))| = o_P(\delta_n)$ for $\delta_n = o_P(1)$. To see this, fix a constant $\zeta>0$. By uniform differentiability of F_0, for all ζ there exists a constant $\bar{\delta}>0$ such that

$$\sup_{p\in[\underline{p},\bar{p}]}\frac{|F_0(p+\delta_n)-F_0(p)-\delta_n f_0(p)|}{\delta_n}\leq\zeta$$

whenever $\delta_n\leq\bar{\delta}$. The event $\{\sup_{p\in[\underline{p},\bar{p}]}\frac{|F_0(p+\delta_n)-F_0(p)-\delta_n f_0(p)|}{\delta_n}>\zeta\}$ is equal to

$$\left\{\sup_{p\in[\underline{p},\bar{p}]}\frac{|F_0(p+\delta_n)-F_0(p)-\delta_n f_0(p)|}{\delta_n}>\zeta,\delta_n\leq\bar{\delta}\right\}$$
$$\cup\left\{\sup_{p\in[\underline{p},\bar{p}]}\frac{|F_0(p+\delta_n)-F_0(p)-\delta_n f_0(p)|}{\delta_n}>\zeta,\delta_n>\bar{\delta}\right\}.$$

The first event has measure zero by uniform differentiability. The probability of the second event is dominated by $\mathbb{P}\left(\delta_n>\bar{\delta}\right)$, which is $o(1)$. Then the statement follows by noting that $F_0[p\pm\delta_n] = F_0(p+\delta_n)-F_0(p)+F_0(p)-F_0(p-\delta_n)$ and that $\max_{i\in[n]}\max_{j\in\mathcal{J}(i)}|f_0(\pi(X_i))-f_0(\pi(X_j))| = o_P(1)$ (see proof of Assertion (iii)), so $2f_0(p) = f_0(p)+f_0(p-\delta_n)+f_0(p)-$

$f_0(p-\delta_n) = f_0(p) + f_0(p-\delta_n) + o_P(1)$. As $\inf_{p\in[\underline{p},\bar{p}]} f_0(p) > 0$ by Assumption 2, Assertion (ii) follows. When the caliper is $\delta_n = \underline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, Proposition 1, the continuous mapping theorem and the law of large numbers imply $\delta_n = o_P(1)$, whence the assertion follows by the above arguments.

Assertion (iii). As f_0 is uniformly continuous by Assumption 2,

$$\max_{i\in[n]} \max_{j\in\mathcal{J}(i)} |f_0(\pi(X_j)) - f_0(\pi(X_i))| = o_P(1).$$

To see this, fix a constant $\zeta > 0$. By uniform continuity of f_0, for all ζ there exists an $\eta > 0$ such that $|f_0(p) - f_0(p')| \leq \zeta$ whenever $|p-p'| \leq \eta$ for all $p, p' \in [\underline{p}, \bar{p}]$. The event $\{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta\}$ is equal to $\{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta, |\pi(X_i) - \pi(X_j)| \leq \eta\} \cup \{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta, |\pi(X_i) - \pi(X_j)| > \eta\}$. The first event has measure zero by uniform continuity. As $j \in \mathcal{J}(i)$, the probability of the second event is dominated by $\mathbb{P}(|\pi(X_i) - \pi(X_j)| > \eta) \leq \mathbb{P}(\delta_n > \eta)$, which is $o(1)$. Hence $\max_{i:D_i=0} \max_{j\in\mathcal{J}(i)} |\check{R}_{0ji}| = o_P(1)$, because $\inf_{p\in[\underline{p},\bar{p}]} f_0(p) > 0$ by Assumption 2. When the caliper is $\delta_n = \underline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, arguments proving Assertion (ii) apply.

Assertions (iv), (v), (vi) follow along the same arguments by conditioning on $\hat{\theta}$, exploiting Assumptions 6 and 9 and that the constants of the bounds of Alexander (1987) and Van der Vaart and Wellner (1996) do not depend on the underlying distribution. This holds for any caliper choice in (13) as $\hat{\underline{\Delta}}_n = o_P(1)$ by the proof of Proposition 5. ■

Lemma 2 (Error Term). *Suppose Assumptions 2 and 4 hold, and the caliper δ_n satisfies (3). Then*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(V_E^{-1/2}\sqrt{n}E^{CDO} \leq x \,\middle|\, D^{(n)}, PS^{(n)} \right) - \Phi(x) \right| \xrightarrow{P} 0 \qquad \text{as } n \to \infty,$$

for $V_E := \frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^2\sigma^2_{D_i}(\pi(X_i))$ and standard normal distribution function Φ.

Proof. We apply a Lindeberg–Feller central limit theorem as E_i, given $(D^{(n)}, PS^{(n)})$, are independently, but not identically, distributed with mean zero across $i \in [n]$ (the M_i, w_i are constants given $(D^{(n)}, PS^{(n)})$). By Assumption 2, the μ^d, and hence ε, are well-defined. By definition of V_E, $\mathbb{V}\left[\sum_{i\in[n]} E_i/\sqrt{nV_E} \mid D^{(n)}, PS^{(n)}\right] = 1$. Thus, we only need to verify the Lindeberg–Feller condition:

$$\sum_{i\in[n]} \mathbb{E}\left[(E_i/\sqrt{nV_E})^2 \mathbb{1}_{|E_i/\sqrt{nV_E}|\geq\eta} \,\middle|\, D^{(n)}, PS^{(n)} \right] \xrightarrow{P} 0 \quad \text{for all constants } \eta > 0. \tag{43}$$

Since $\mathbb{1}_{|E_i/\sqrt{nV_E}|\geq\eta}$ is bounded by $E_i^2/(\eta^2 nV_E)$ on $\{E_i^2/(\eta^2 nV_E) \geq 1\}$, the expectation in (43) satisfies

$$\frac{1}{nV_E}\mathbb{E}\left[E_i^2 \mathbb{1}_{|E_i|\geq\eta\sqrt{nV_E}} \,\middle|\, D^{(n)}, PS^{(n)}\right] \leq \frac{1}{nV_E}\mathbb{E}\left[E_i^4/(\eta^2 nV_E) \,\middle|\, D^{(n)}, PS^{(n)}\right]$$
$$= \frac{\mathbb{E}\left[E_i^4 \mid D^{(n)}, PS^{(n)}\right]}{(\eta nV_E)^2}, \tag{44}$$

where $\mathbb{E}\left[E_i^4 \mid D^{(n)}, PS^{(n)}\right] = (\mathbb{1}_{M_i>0} + w_i)^4 \mathbb{E}\left[\varepsilon_i^4 \mid D_i, \pi(X_i)\right]$, with $\mathbb{E}\left[\varepsilon_i^4 \mid D_i, \pi(X_i)\right] \leq \sup_{d\in\mathbb{B}, p\in[0,1]} \mathbb{E}\left[\varepsilon_i^4 \mid D_i = d, \pi(X_i) = p\right] < \infty$ by Assumption 4. By Lemma 4, $V_E \xrightarrow{P} V_{\sigma,\pi} \in (0, \infty)$, so (44) is bounded by $(1 + o_P(1))n^{-1}\left(\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4\right)$ up to a constant factor. Fix a constant $C > 0$. By Markov's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4 > C\right) \leq C^{-1}\max_{i\in[n]}\mathbb{E}(\mathbb{1}_{M_i>0} + w_i)^4.$$

Below, we show that $n^{-\varrho}\max_{i\in[n]}\mathbb{E}(\mathbb{1}_{M_i>0} + w_i)^4 = O(1)$ for any $\varrho > 0$, so that

$$\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4 = O_P(n^{\varrho}).$$

Then (44) is bounded by $(1 + o_P(1))O_P(n^{\varrho-1})$, which is $o_P(1)$ for $\varrho < 1$, so (43) is met.

As $w_i \geq 0$ and $(1 + x)^4 \leq (2x)^4$ for $x \geq 1$,

$$(\mathbb{1}_{M_i>0} + w_i)^4 \leq 2^4 + 2^4 w_i^4 \mathbb{1}_{w_i>1}.$$

We have $w_i \leq M_i \max_{j\in\mathcal{J}(i)} M_j^{-1}$ and hence

$$\mathbb{E}w_i^4\mathbb{1}_{w_i>1} \leq \mathbb{E}\left[\mathbb{1}_{w_i>1}M_i^4 \max_{j\in\mathcal{J}(i)} M_j^{-4}\right].$$

This yields the result by Lemma 3.

When the caliper is $\delta_n = \underline{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, it is $\sigma\left\{D^{(n)}, PS^{(n)}\right\}$-measurable, hence the M_i, w_i are constants given $(D^{(n)}, PS^{(n)})$. Lemmas 3 and 4 complete the proof. ■

Lemma 3 (Lindeberg–Feller Bound). *Suppose that the caliper δ_n satisfies (3) and that Assumption 2 holds. Then for any finite fixed constant integer $r \geq 2$ and any finite fixed constant $\varrho > 0$,*

$$\max_{i\in[n]}\mathbb{E}\mathbb{1}_{w_i>1}\max_{j\in\mathcal{J}(i)}\left(\frac{M_i}{M_j}\right)^r = o(n^{\varrho}). \tag{45}$$

Proof. If $w_i > 1$ for some $i \in [n]$, then $M_i \geq 1$ and hence also $M_j \geq 1$ for all $j \in \mathcal{J}(i)$, as $j \in \mathcal{J}(i)$ if and only if $i \in \mathcal{J}(j)$. This in turn implies $N_0, N_1 \geq 1$. Thus, $\mathbb{1}_{w_i>1} \leq \mathbb{1}_{M_j\geq 1}\mathbb{1}_{N_0\geq 1}\mathbb{1}_{N_1\geq 1}$ for all $j \in \mathcal{J}(i)$ for all $i \in [n]$. By definition, $M_i = N_{1-d}\mathbb{F}_{N_{1-d}}[\pi(X_i)\pm\delta_n]$ if $D_i = d$. Then, in the notation of (17) and (32), the left side of (45) is bounded by

$$\sum_{d\in\{0,1\}} \max_{i:D_i=d} \mathbb{E}\left[\left(\frac{N_{1-d}\mathbb{1}_{N_d\geq 1}}{N_d}\right)^r \max_{j\in\mathcal{J}(i)}\left(\frac{F_{1-d}[\pi(X_i)\pm\delta_n]}{F_d[\pi(X_j)\pm\delta_n]}\frac{1+R_{1-d,i}}{1+R_{d,j}}\mathbb{1}_{N_{1-d}\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right]$$
$$\lesssim \sum_{d\in\{0,1\}} \max_{i:D_i=d} \mathbb{E}\left[\left(\frac{N_{1-d}\mathbb{1}_{N_d\geq 1}}{N_d}\right)^r \max_{j\in\mathcal{J}(i)}\left(\frac{1+R_{1-d,i}}{1+R_{d,j}}\mathbb{1}_{N_{1-d}\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right], \tag{46}$$

where the second line follows from Assumption 2: because f_0, f_1 have the same support, are bounded away from zero and infinity, we have, for $c \neq 0$, $0 < \frac{2c\inf_{p\in[\underline{p},\bar{p}]} f_{1-d}(p)}{2c\sup_{p\in[\underline{p},\bar{p}]} f_d(p)} \leq F_{1-d}[a\pm c]/F_d[b\pm c] \leq \frac{2c\sup_{p\in[\underline{p},\bar{p}]} f_{1-d}(p)}{2c\inf_{p\in[\underline{p},\bar{p}]} f_d(p)} < \infty$. We address the case $d = 0$ in (46); $d = 1$ follows by symmetry. For the expectation in (46), two applications of the Cauchy–Schwarz inequality give

$$\mathbb{E}\left[\left(\frac{N_1\mathbb{1}_{N_0\geq 1}}{N_0}\right)^r \max_{j\in\mathcal{J}(i)}\left(\frac{1+R_{1i}}{1+R_{0j}}\mathbb{1}_{N_1\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right]$$
$$\leq \sqrt{\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r}\sqrt{\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1}\mathbb{E}\left(\max_{j\in\mathcal{J}(i)}(\mathbb{1}_{M_j\geq 1}/(1+R_{0j}))^r\right)^4}}. \tag{47}$$

Here $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r} \leq n^{2r}\mathbb{E}(\mathbb{1}_{N_0\geq 1}/N_0)^{2r}$ with

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{N_0\geq 1}/N_0)^{2r} &= \sum_{n_0=1}^{n}\binom{n}{n_0}(1-p_1)^{n_0}p_1^{n-n_0}\left(\frac{1}{n_0}\right)^{2r}\\
&= \sum_{n_0=0}^{n-1}\frac{n}{n_0+1}\binom{n-1}{n_0}(1-p_1)^{n_0+1}p_1^{n-(n_0+1)}\left(\frac{1}{n_0+1}\right)^{2r}\\
&= n\frac{1-p_1}{p_1}\sum_{n_0=0}^{n-1}\binom{n-1}{n_0}(1-p_1)^{n_0}p_1^{n-n_0}\left(\frac{1}{n_0+1}\right)^{2r+1}.
\end{aligned}$$

The last line is $O(n^{-2r})$ by Cribari-Neto et al. (2000), thus $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r} = O(1)$.

Next, we bound the second factor of (47). We have $\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} \leq 2^{4r} + c\mathbb{E}|R_{1i}|^{4r}\mathbb{1}_{N_1\geq 1}$ for some constant $c > 0$, where $\max_{i:D_i=0}|R_{1i}| \leq W_1$ for W_1 in (38). Because $W_1 \geq 0$,

$$\mathbb{E}\left[W_1^{4r} \mid D^{(n)}\right] = \int_0^\infty \mathbb{P}\left(W_1^{4r} > w \mid D^{(n)}\right)\mathrm{d}w = \int_0^\infty \mathbb{P}\left(W_1 > w^{\frac{1}{4r}} \mid D^{(n)}\right)\mathrm{d}w.$$

By (42) of Lemma 1,

$$\mathbb{E}\left[W_1^{4r} \mid D^{(n)}\right] \leq \frac{c_0}{N_1\gamma_n}\int_0^\infty \frac{1}{w^{\frac{2}{4r}}} e^{-w^{\frac{2}{4r}}N_1\gamma_n/1024}\mathrm{d}w + c_1\int_0^\infty e^{-w^{\frac{1}{4r}}N_1\gamma_n/256}\mathrm{d}w \tag{48}$$

$$+ cN_1\int_0^\infty w^{\frac{2}{4r}} e^{-\frac{N_1 w^{\frac{2}{4r}}}{2}}\mathrm{d}w. \tag{49}$$

The integral in the first term of (48) is $\frac{c_0}{N_1\gamma_n}\frac{1}{\lambda_{n,N_1}}\int_0^\infty t^{-1}\lambda_{n,N_1}e^{-\lambda_{n,N_1}t}t^{2r-1}\mathrm{d}t$, where $\lambda_{n,N_1} := \frac{N_1\gamma_n}{1024}$ is strictly positive for $N_1 \geq 1$. This integral is the $(2r-2)$th moment of an Exponential(λ_{n,N_1}) variable, which is well-defined for $r \geq 2$ and for finite integer r is bounded by $\left(\frac{1}{\lambda_{n,N_1}}\right)^{2r-2}$ up to a constant factor. Hence, the first term of (48) is bounded by $c_0\left(\frac{1}{N_1\gamma_n}\right)^{2r} \simeq c_0\left(\frac{n}{N_1\log n}\right)^{2r}$. Similar arguments show that the second integrals of (48) and (49) are bounded by $c_1\left(\frac{1}{N_1\gamma_n}\right)^{4r}$ and $c\left(\frac{1}{N_1}\right)^{2r}$, respectively. Conclude that

$$\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} \leq O(1) + c_0 n^{2r}\mathbb{E}\left(\frac{\mathbb{1}_{N_1\geq 1}}{N_1}\right)^{2r} + c_1 n^{4r}\mathbb{E}\left(\frac{\mathbb{1}_{N_1\geq 1}}{N_1}\right)^{4r}.$$

By arguments bounding $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r}$ of (47), the right side is $O(1)$.

Finally, the last factor of (47) satisfies

$$\max_{j\in\mathcal{J}(i)}(\mathbb{1}_{M_j\geq 1}/(1+R_{0j}))^r = \max_{j\in\mathcal{J}(i)}\left(\frac{N_0F_0[\pi(X_j)\pm\delta_n]}{M_j}\mathbb{1}_{M_j\geq 1}\right)^r \leq (2n\|f_0\|_\infty\delta_n)^r \tag{50}$$

$$\lesssim (\log n)^r.$$

As $(\log n)^r = o(n^\varrho)$ for any $\varrho > 0$, (45) follows.

When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, (48) holds with $\gamma_n := 2\inf_{p\in[\underline{p},\bar{p}]} f_1(p)\frac{\log N_1}{N_1+1}$ in view of the proof of Lemma 1, showing $\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} = O(1)$. Consider (50), wherein $\delta_n^r \leq 2^r\left(\overline{\Delta}_n^r + (2\log n)^r\left(\left(\frac{1}{N_0+1}\right)^r + \left(\frac{1}{N_1+1}\right)^r\right)\right)$. Here, $\mathbb{E}\left(\frac{1}{N_d+1}\right)^r = O(n^{-r})$ by Cribari-Neto et al. (2000). Arguments in the proof of Proposition 1 and Lemma 5 imply $n^r\mathbb{E}\overline{\Delta}_n^r = O((\log n)^r)$. Conclude that $n^r\mathbb{E}\delta_n^r = O((\log n)^r)$ and hence (45) follows from (50). ■

Lemma 4 (Semiparametric Efficiency)**.** *If Assumptions 2 and 4 hold, and the caliper δ_n satisfies (3), then $V_E \xrightarrow{P} V_{\sigma,\pi}$ as $n\to\infty$.*

Proof. By Proposition 2, $\min_{i\in[n]} M_i > 0$ with probability tending to one, so with probability tending to one,

$$V_E = \frac{1}{n}\sum_{i\in[n]}(1+w_i)^2\sigma_{D_i}^2(\pi(X_i)) = \frac{1}{n}\sum_{i\in[n]}(1+2w_i+w_i^2)\sigma_{D_i}^2(\pi(X_i))$$

$$= \mathbb{E}\sigma_D^2(\pi(X)) + o_P(1) + \frac{2}{n}\sum_{i\in[n]}w_i\sigma_{D_i}^2(\pi(X_i)) + \frac{1}{n}\sum_{i\in[n]}w_i^2\sigma_{D_i}^2(\pi(X_i)) \tag{51}$$

by the law of large of large numbers. In the notation of (17), we have, by definition,

$$\sum_{i\in[n]} w_i^r \sigma_{D_i}^2(\pi(X_i)) = \left(\frac{N_1}{N_0}\right)^r \sum_{i:D_i=0} \left(\frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n]}\right)^r \sigma_{D_i}^2(\pi(X_i))$$
$$+\left(\frac{N_0}{N_1}\right)^r \sum_{i:D_i=1} \left(\frac{1}{N_0}\sum_{j\in\mathcal{J}(i)} \frac{1}{\mathbb{F}_{N_1}[\pi(X_j)\pm\delta_n]}\right)^r \sigma_{D_i}^2(\pi(X_i)). \quad (52)$$

Write $\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n] = (1+R_{0j})F_0[\pi(X_j)\pm\delta_n]$ and $F_0[\pi(X_j)\pm\delta_n] = 2\delta_n f_0(\pi(X_j))(1+\tilde{R}_{0j})$ for $R_{0j}, \tilde{R}_{0j}$ of (32), (33). By Lemma 1,

$$\max_{j:D_j=1} |R_{0j}| = o_P(1) \quad \text{and} \quad \max_{j:D_j=1} |\tilde{R}_{0j}| = o_P(1).$$

Then we can write

$$\begin{aligned}
\frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n]} &= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{F_0[\pi(X_j)\pm\delta_n]}\frac{1}{1+R_{0j}}\\
&= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{2\delta_n f_0(\pi(X_j))}\frac{1}{1+\tilde{R}_{0j}}\frac{1}{1+R_{0j}}\\
&= \frac{1+o_P(1)}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{2\delta_n f_0(\pi(X_j))}
\end{aligned}$$

where the $o_P(1)$ terms are uniform in $i \in [n]$. Write $f_0(\pi(X_j)) = f_0(\pi(X_i))(1+\check{R}_{0ji})$ for $\check{R}_{0ji}$ of (34). By Lemma 1, $\max_{i:D_i=0}\max_{j\in\mathcal{J}(i)} |\check{R}_{0ji}| = o_P(1)$. Then, by the continuous mapping theorem,

$$\begin{aligned}
\frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{2\delta_n f_0(\pi(X_j))} &= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)} \frac{1}{2\delta_n f_0(\pi(X_i))}\frac{1}{1+\check{R}_{0ji}}\\
&= (1+o_P(1))\frac{f_1(\pi(X_i))}{f_0(\pi(X_i))}\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))},
\end{aligned}$$

because $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$ by Assumption 2. Write

$$\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))} = \frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{F_1[\pi(X_i)\pm\delta_n]}\frac{F_1[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))},$$

where $\max_{i:D_i=0}\left|\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{F_1[\pi(X_i)\pm\delta_n]} - 1\right| = o_P(1)$ and $\max_{i:D_i=0}\left|\frac{F_1[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))} - 1\right| = o_P(1)$ by Lemma 1. By symmetry, similar arguments apply to the second term of (52). Then (52), divided by n, is equal to

$$\begin{aligned}\frac{1+o_P(1)}{n}&\Bigg[\left(\frac{N_1}{N_0}\right)^r\sum_{i:D_i=0}\left(\frac{f_1(\pi(X_i))}{f_0(\pi(X_i))}\right)^r\sigma^2_{D_i}(\pi(X_i))\\&+\left(\frac{N_0}{N_1}\right)^r\sum_{i:D_i=1}\left(\frac{f_0(\pi(X_i))}{f_1(\pi(X_i))}\right)^r\sigma^2_{D_i}(\pi(X_i))\Bigg]\\&\xrightarrow{P}\mathbb{E}\left[(1-D)\left(\frac{p_1}{1-p_1}\frac{f_1(\pi(X))}{f_0(\pi(X))}\right)^r\sigma^2_D(\pi(X))\right]\qquad(53)\\&+\mathbb{E}\left[D\left(\frac{1-p_1}{p_1}\frac{f_0(\pi(X))}{f_1(\pi(X))}\right)^r\sigma^2_D(\pi(X))\right],\qquad(54)\end{aligned}$$

by the weak law of large numbers as $(N_d/N_{1-d})^r \xrightarrow{a.s.} \left(\frac{p_d}{1-p_d}\right)^r$ and the f_d/f_{1-d} are uniformly bounded by Assumption 2 and the σ_d^2 are bounded by Assumption 4. For $r=1$, (53) is

$$\begin{aligned}\mathbb{E}\left[\frac{p_1}{1-p_1}\frac{f_1(\pi(X))}{f_0(\pi(X))}\sigma_0^2(\pi(X))\,\middle|\, D=0\right](1-p_1)&=\int_{\underline{p}}^{\bar{p}}p_1\frac{f_1(p)}{f_0(p)}\sigma_0^2(p)f_0(p)\mathrm{d}p\\&=\mathbb{E}\pi(X)\sigma_0^2(\pi(X)),\end{aligned}$$

where we used that $f_1(p)=\frac{p}{p_1}f_{\pi(X)}(p)$, with $f_{\pi(X)}$ being the density of $\pi(X)$. Noting that $f_0(p)=\frac{1-p}{1-p_1}f_{\pi(X)}(p)$, (54) is $\mathbb{E}(1-\pi(X))\sigma_1^2(\pi(X))$. For $r=2$, the same arguments yield $\mathbb{E}\frac{\pi(X)^2}{1-\pi(X)}\sigma_0^2(\pi(X))$ for (53), and $\mathbb{E}\frac{(1-\pi(X))^2}{\pi(X)}\sigma_1^2(\pi(X))$ for (54). Note that $\mathbb{E}\sigma_D^2(\pi(X))=\mathbb{E}(1-\pi(X))\sigma_0^2(\pi(X))+\mathbb{E}\pi(X)\sigma_1^2(\pi(X))$. Collect the terms to get the assertion, which also holds for $\delta_n=\underline{\overline{\Delta}}_n\vee\frac{\log N_0}{N_0+1}\vee\frac{\log N_1}{N_1+1}$ in view of Lemma 1. ■

Supplement

Proof of Proposition 3. The semiparametric lower bound is the variance of the efficient influence function. When the observed sample is $((Y_i, D_i, X_i))_{i\in[n]}$, the efficient influence function of ATE is

$$\chi_{\mathcal{X}}(Y,D,X) \coloneqq \frac{D(Y-\mu^1_{\mathcal{X}}(X))}{\pi(X)} - \frac{(1-D)(Y-\mu^0_{\mathcal{X}}(X))}{1-\pi(X)} + \mu^1_{\mathcal{X}}(X) - \mu^0_{\mathcal{X}}(X) - \tau,$$

see Hahn (1998). Then $V = V_{\mathrm{eff}}$ follows under (10) in Proposition 3. One can verify that $V = \mathbb{V}\left[\chi(Y,D,\pi(X))\right]$, where

$$\begin{aligned}\chi(Y,D,\pi(X)) \coloneqq& \frac{D(Y-\mu^1(\pi(X)))}{\pi(X)} - \frac{(1-D)(Y-\mu^0(\pi(X)))}{1-\pi(X)}\\ &+ \mu^1(\pi(X)) - \mu^0(\pi(X)) - \tau.\end{aligned}$$

Assumption 1 and $\mathbb{E}\left[D \mid X\right] = \pi(X)$ imply that

$$\mathbb{E}\chi_{\mathcal{X}}(Y,D,X)(\chi(Y,D,\pi(X)) - \chi_{\mathcal{X}}(Y,D,X)) = 0,$$

thus

$$\begin{aligned}V = \mathbb{V}\left[\chi(Y,D,\pi(X))\right] &= \mathbb{V}\left[\chi(Y,D,\pi(X)) - \chi_{\mathcal{X}}(Y,D,X)\right] + \mathbb{V}\left[\chi_{\mathcal{X}}(Y,D,X)\right]\\ &\geq \mathbb{V}\left[\chi_{\mathcal{X}}(Y,D,X)\right] = V_{\mathrm{eff}}.\end{aligned}$$

For ATT, $V_{\mathrm{t,eff}} \leq V_{\mathrm{t}}$ follows similarly as the efficient influence function of ATT under unknown propensity score (Hahn, 1998) is

$$\begin{aligned}\chi_{\mathrm{t},\mathcal{X}}(Y,D,X) \coloneqq& \frac{D}{p_1}(Y-\mu^1_{\mathcal{X}}(X)) - \frac{1-D}{p_1}\frac{\pi(X)}{1-\pi(X)}(Y-\mu^0_{\mathcal{X}}(X))\\ &+ \frac{D}{p_1}(\mu^1_{\mathcal{X}}(X) - \mu^0_{\mathcal{X}}(X) - \tau_{\mathrm{t}}).\end{aligned}$$

Under Assumption 1, one verifies that $V_{\mathrm{t}} = \mathbb{V}\left[\chi_{\mathrm{t}}(Y,D,\pi(X))\right]$, where

$$\begin{aligned}\chi_{\mathrm{t}}(Y,D,\pi(X)) \coloneqq& \frac{D}{p_1}(Y-\mu^1(\pi(X))) - \frac{1-D}{p_1}\frac{\pi(X)}{1-\pi(X)}(Y-\mu^0(\pi(X)))\\ &+ \frac{D}{p_1}(\mu^1(\pi(X)) - \mu^0(\pi(X)) - \tau_{\mathrm{t}}),\end{aligned}$$

and that

$$\mathbb{E}\chi_{\mathrm{t},\mathcal{X}}(Y,D,X)(\chi_{\mathrm{t}}(Y,D,\pi(X)) - \chi_{\mathrm{t},\mathcal{X}}(Y,D,X)) = 0,$$

thus proving the assertion.

■

Lemma 5 (Moment Bounds of Ordered Uniform Spacings). *Let the order statistics of* $(U_1, U_2, \dots, U_n) \overset{i.i.d.}{\sim}$ *Uniform*$(0,1)$ *be* $U_{(1)} \le U_{(2)} \le \dots \le U_{(n)}$. *Let* $\tilde{U}_1 \coloneqq U_{(1)}$, $\tilde{U}_i \coloneqq U_{(i)} - U_{(i-1)}$ *for* $i = 2, \dots, n$ *and* $\tilde{U}_{n+1} \coloneqq 1 - U_{(n)}$ *be the spacings generated by* $(U_i)_{i\in[n]}$. *Let* $\tilde{U}_{(1)} \le \tilde{U}_{(2)} \le \dots \le \tilde{U}_{(n+1)}$ *be the ordered spacings. Then for any finite fixed integer* $1 \le a < \frac{n+1}{2}$ *and* $n \ge 2$,

$$\mathbb{E}\tilde{U}_{(r)}^a = \begin{cases} O\left(\left(\frac{1}{n}\right)^{2a}\right) & \text{for } r = 1 \\ O\left(\left(\frac{\log r}{n}\right)^{a}\right) & \text{for all } r = 2, 3, \dots, n+1. \end{cases}$$

In particular, $\mathbb{E}\tilde{U}_{(r)}^a = o(1)$ *for all* $r \in [n+1]$ *and finite fixed integer* $1 \le a < \frac{n+1}{2}$.

Proof. By Shorack and Wellner (2009, Chapter 21), $\tilde{U}_{(r)}$ is distributed as $\frac{Z_{r:n+1}}{\sum_{i\in[n+1]} Z_i}$, for $(Z_1, Z_2, \dots, Z_{n+1}) \overset{\text{i.i.d.}}{\sim}$ Exponential(1) with order statistics $Z_{1:n+1} \le Z_{2:n+1} \le \dots \le Z_{n+1:n+1}$. The Cauchy–Schwarz inequality gives

$$\mathbb{E}\tilde{U}_{(r)}^a \le \sqrt{\mathbb{E}Z_{r:n+1}^{2a}\mathbb{E}\left(\sum_{i\in[n+1]} Z_i\right)^{-2a}}.$$

Here, $\mathbb{E}Z_{1:n+1}^{2a} = O\left(n^{-2a}\right)$ and $\mathbb{E}Z_{r:n+1}^{2a} = O\left((\log r)^{2a}\right)$ for $r \ge 2$ and for finite fixed integer $a \ge 1$ by Lemma 6. The sum of $n+1$ i.i.d. Exponential(1) variates follows a Gamma$(n+1, 1)$ distribution. Thus, $(\sum_{i\in[n+1]} Z_i)^{-1}$ follows an Inverse-Gamma$(n+1,1)$ distribution, whose $2a$-th moment is equal to $\frac{(n-2a)!}{n!} \lesssim n^{-2a}$ for $2a < n+1$, where the last inequality follows from $\sqrt{2\pi}n^{n+1/2}e^{-n} \le n! \le en^{n+1/2}e^{-n}$ (Robbins, 1955), because $n - 2a + 1/2 \ge 0$ for a positive integer a. ∎

Lemma 6 (Moments of Exponential(1) Order Statistics). *Let* $(Z_1, Z_2, \dots, Z_n) \overset{i.i.d.}{\sim}$ *Exponential(1) with order statistics* $Z_{1:n} \le Z_{2:n+1} \le \dots \le Z_{n:n}$. *Then for finite fixed integer* $k \ge 1$, *we have for all* $n \ge 2$,

$$\mathbb{E}Z_{r:n}^k = k! \sum_{t_1=1}^{r} \frac{1}{n+1-t_1} \sum_{t_2=1}^{t_1} \frac{1}{n+1-t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} \frac{1}{n+1-t_{k-1}} \sum_{t_k=1}^{t_{k-1}} \frac{1}{n+1-t_k} \tag{55}$$

for all $r = 1, 2, \dots, n$. *The right side in* (55) *is* $O\left((\log r)^k\right)$ *for* $r \ge 2$, *and* $\mathbb{E}Z_{1:n}^k = O\left(n^{-k}\right)$.

Proof. The right side of (55) and $\mathbb{E}Z_{1:n}^k = O\left(n^{-k}\right)$ are obtained by solving the recursion in Balakrishnan and Gupta (1998, Theorems 1 and 2). The innermost sum in (55) satisfies

$\sum_{t_k=1}^{t_{k-1}} \frac{1}{n+1-t_k} \leq \sum_{j=1}^{r} \frac{1}{n+1-j}$ as $t_k \leq r$. Because every fraction in (55) is positive, we can upper bound the right side of (55) by

$$k! \left(\sum_{j=1}^{r} \frac{1}{n+1-j} \right) \left(\sum_{t_1=1}^{r} \frac{1}{n+1-t_1} \sum_{t_2=1}^{t_1} \frac{1}{n+1-t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} \frac{1}{n+1-t_{k-1}} \right).$$

Apply the same bound for the remaining $k-1$ sums noting that $1 \leq t_1 \leq t_1 \leq \ldots \leq t_{k-1} \leq r$, to obtain the bound $k! \left(\sum_{j=1}^{r} \frac{1}{n+1-j} \right)^k$ on (55). The proof is complete as $\sum_{j=1}^{r} \frac{1}{n+1-j}$ is $O(\log r)$ for $r \in [n]$ as $n \to \infty$. ■

Lemma 7 (Conditional Martingale Central Limit Theorem)**.** *Let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be a sequence of probability spaces. Let $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn} : \Omega_n \to \mathbb{R}$ be martingale differences with respect to sub-σ-algebras $\mathcal{F}_{n1} \subset \mathcal{F}_{n2} \subset \ldots \subset \mathcal{F}_{nn} \subset \mathcal{F}_n$. Let $\mathcal{F}_{n0} \subset \mathcal{F}_{n1}$ be a sub-σ-algebra. For $k = 1, 2, \ldots, n$, let $\sigma_{nk}^2 := \mathbb{E}_n \left[\xi_{nk}^2 \mid \mathcal{F}_{n,k-1} \right]$. If there exists a finite constant $\sigma > 0$ such that*

$$\mathbb{P}_n \left(\left| \sum_{k=1}^{n} \sigma_{nk}^2 - \sigma^2 \right| > \epsilon \,\middle|\, \mathcal{F}_{n0} \right) \xrightarrow{\mathbb{P}_n} 0 \quad \textit{for all constants } \epsilon > 0 \textit{ and} \tag{56}$$

$$\sum_{k=1}^{n} \mathbb{E}_n \left[\xi_{nk}^2 \mathbb{1}_{|\xi_{nk}| \geq \eta} \,\middle|\, \mathcal{F}_{n0} \right] \xrightarrow{\mathbb{P}_n} 0 \quad \textit{for all constants } \eta > 0, \tag{57}$$

then $\mathbb{P}_n \left(\sigma^{-1} \sum_{k=1}^{n} \xi_{nk} \leq x \mid \mathcal{F}_{n0} \right) \xrightarrow{\mathbb{P}_n} \Phi(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$, where Φ is the standard normal distribution function.

Proof. Follows from Billingsley (1995, Theorem 35.12) by conditioning on $\mathcal{F}_{n0}$ throughout. ■

Proof of Proposition 7. Existence of $A_n, \hat{A}_n$. We show that $A_n, \hat{A}_n$ are well-defined with probability tending to one. For A_n, this happens if and only if

$$\mathbb{P}\left(\underline{p}_{\hat{\theta}} + a_n < \bar{p}_{\hat{\theta}} - a_n \right) = \mathbb{P}\left(2a_n < \bar{p}_{\hat{\theta}} - \underline{p}_{\hat{\theta}} \right) \to 1.$$

If $\underline{p}_{\hat{\theta}} = \underline{p} + o_P(1)$ and $\bar{p}_{\hat{\theta}} = \bar{p} + o_P(1)$, the probability is $\mathbb{P}\left(2a_n < \bar{p} - \underline{p} + o_P(1)\right) = \mathbb{P}\left(2 < a_n^{-1}(\bar{p} - \underline{p}) + o_P(a_n^{-1})\right)$, which goes to one as $\bar{p} > \underline{p}$ and $a_n \downarrow 0$. To show that these conditions hold, define $T_n(x) := \hat{\theta}^\intercal x$ and $T(x) := \theta_0^\intercal x$, so by definition $\bar{p} = \sup_{x \in \mathcal{X}} g(T(x)) = g(\sup_{x \in \mathcal{X}} T(x))$ and $\bar{p}_{\hat{\theta}} = \sup_{x \in \mathcal{X}} g(T_n(x)) = g(\sup_{x \in \mathcal{X}} T_n(x))$ because g is increasing by Assumption 11. Since g is continuous by Assumption 11, it suffices by the continuous mapping theorem to show $\sup_{x \in \mathcal{X}} T_n(x) \xrightarrow{P} \sup_{x \in \mathcal{X}} T(x)$. Because $\mathcal{X}$ is bounded and $\hat{\theta} \xrightarrow{P} \theta_0$ by Assumption 9, $\sup_{x \in \mathcal{X}} |T_n(x) - T(x)| \lesssim \|\hat{\theta} - \theta_0\| = o_P(1)$, so that $\bar{p}_{\hat{\theta}} = \bar{p} + o_P(1)$. Similar arguments yield $\underline{p}_{\hat{\theta}} = \underline{p} + o_P(1)$.

For $\hat{A}_n$, the desired result follows from that for A_n above, and that $\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \underline{p}_{\hat{\theta}} + o_P(1)$ and $\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \bar{p}_{\hat{\theta}} + o_P(1)$. Because $F_{\hat{\theta}}^{-1}$, the inverse of $F_{\hat{\theta}}(p) = p_1 F_{1,\hat{\theta}}(p) + (1-p_1)F_{0,\hat{\theta}}(p)$ which is the distribution of $(\pi(X,\hat{\theta}) \mid \hat{\theta})$ under Assumption 9, is strictly increasing by Assumption 6, $(\min_{i\in[n]} \pi(X_i,\hat{\theta}) \mid \hat{\theta})$ is distributed as $F_{\hat{\theta}}^{-1}(U_{(1)})$, where $U_{(1)}$ is the sample minimum of $((U_1, U_2, \dots, U_n) \mid \hat{\theta}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$. Then $(\min_{i\in[n]} \pi(X_i,\hat{\theta}) - \underline{p}_{\hat{\theta}} \mid \hat{\theta})$ is distributed as $F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(F_{\hat{\theta}}(\underline{p}_{\hat{\theta}})) = F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(0)$ given $\hat{\theta}$. By Assumption 6, $F_{\hat{\theta}}^{-1}$ is Lipschitz with constant $\left\|(F_{\hat{\theta}}^{-1})'\right\|_\infty$ with $(F_{\hat{\theta}}^{-1})'(u) = \frac{1}{f_{\hat{\theta}}(F_{\hat{\theta}}^{-1}(u))}$ finite for $\inf_{p\in \underline{p}_{\hat{\theta}}, \bar{p}_{\hat{\theta}}} f_{\hat{\theta}}(p) > 0$ by Assumption 6 for $\hat{\theta} \in \text{Nb}(\theta_0, \epsilon)$. Thus, $F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(0) \lesssim U_{(1)}$. Here, $\mathbb{E}U_{(1)}$ goes to zero by the proof of Proposition 1. Hence, by Assumption 11, $\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \underline{p}_{\hat{\theta}} + o_P(1)$ and $\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \bar{p}_{\hat{\theta}} + o_P(1)$ similarly. In the following, we prove the consistency of the variance component estimators for $V_{\hat{\pi}}$; similar arguments give the result for $V_{\mathrm{t},\hat{\pi}}$. We show below that $\hat{N}/n \xrightarrow{P} 1$ (since (62) is $o_P(1)$). Therefore, in the following, we prove the consistency of the estimators $\hat{V}_\tau, \hat{V}_{\tau_\mathrm{t}}, \hat{V}_{\sigma,\pi}, \hat{V}_{\mathrm{t},\sigma,\pi}, \hat{q}_d$ and $\hat{q}_{\mathrm{t},d}$ normalised by n rather than $\hat{N}$.

Consistency of $\hat{V}_\tau$. By Theorem 3 and the continuous mapping theorem, $(\hat{\tau}_{\hat{\pi}})^2 \xrightarrow{P} \tau^2$. For short, put $\pi_i := g(\theta_0^\intercal X_i)$, $\hat{\pi}_i := g(\hat{\theta}^\intercal X_i)$ and $\hat{\mathbb{1}}_i := \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in \hat{A}_n}$. The first term in (16) is

$$\frac{1}{n}\sum_{i\in[n]}[\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i)]^2\hat{\mathbb{1}}_i \tag{58}$$

$$+\frac{1}{n}\sum_{i\in[n]}[\hat{\mu}^1(\hat{\theta},\hat{\pi}_i) - \hat{\mu}^0(\hat{\theta},\hat{\pi}_i) - (\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i))]^2\hat{\mathbb{1}}_i \tag{59}$$

$$+\frac{2}{n}\sum_{i\in[n]}[\hat{\mu}^1(\hat{\theta},\hat{\pi}_i) - \hat{\mu}^0(\hat{\theta},\hat{\pi}_i) - (\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i))][\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i)]\hat{\mathbb{1}}_i. \tag{60}$$

Here, (58) converges to $\mathbb{E}(\mu^1(\theta_0,\pi_i) - \mu^0(\theta_0,\pi_i))^2$. To see this, first note that under Assumption 9,

$$\left|\frac{1}{n}\sum_{i\in[n]}[\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i)]^2\hat{\mathbb{1}}_i - \frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i) - \mu^0(\theta_0,\pi_i)]^2\hat{\mathbb{1}}_i\right| \xrightarrow{P} 0,$$

which follows from a mean-value expansion of $[\mu^1(\hat{\theta},\hat{\pi}_i) - \mu^0(\hat{\theta},\hat{\pi}_i)]^2$ in $\hat{\theta}$, similarly to the treatment of (26) in the proof of Theorem 3. Second, as the $\mu^d(\theta,\cdot)$, $\theta \in \text{Nb}(\theta_0,\epsilon)$, are bounded by Assumption 5,

$$\left|\frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i) - \mu^0(\theta_0,\pi_i)]^2\hat{\mathbb{1}}_i - \frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i) - \mu^0(\theta_0,\pi_i)]^2\right| \tag{61}$$

is of the order

$$\frac{1}{n}\sum_{i\in[n]}(1-\hat{\mathbb{1}}_i) = \frac{1}{n}\sum_{i\in[n]}\left(\mathbb{1}_{g(\hat\theta^\intercal X_i)\notin\hat A_n} - \mathbb{E}\left[\mathbb{1}_{g(\hat\theta^\intercal X_i)\notin\hat A_n}\,\middle|\,\hat\theta\right]\right) + \frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\mathbb{1}_{g(\hat\theta^\intercal X_i)\notin\hat A_n}\,\middle|\,\hat\theta\right]. \tag{62}$$

with probability tending to one. Here the first term has mean zero and variance bounded by $1/n$, so it converges to zero in the first mean, and then so in probability. By Assumption 9, the second term of (62) is

$$\mathbb{P}\left(g(\hat\theta^\intercal X_i)\notin\hat A_n\,\middle|\,\hat\theta\right) = \mathbb{P}\left(g(\hat\theta^\intercal X_i) < \min_{i\in[n]} g(\hat\theta^\intercal X_i) + a_n\,\middle|\,\hat\theta\right) \tag{63}$$

$$+\,\mathbb{P}\left(\max_{i\in[n]} g(\hat\theta^\intercal X_i)) - a_n < g(\hat\theta^\intercal X_i)\,\middle|\,\hat\theta\right) \tag{64}$$

for some $i\in[n]$. Let $\underline{G}_{\hat\theta} \coloneqq \min_{i\in[n]} g(\hat\theta^\intercal X_i)$. To bound (63), we have, by Shanmugam and Arnold (1988),

$$\mathbb{P}\left(g(\hat\theta^\intercal X_i)\le \underline{G}_{\hat\theta}+a_n\,\middle|\,\hat\theta,\underline{G}_{\hat\theta}\right) = \frac{1}{n}+\frac{n-1}{n}\frac{F_{\hat\theta}(\underline{G}_{\hat\theta}+a_n)-F_{\hat\theta}(\underline{G}_{\hat\theta})}{1-F_{\hat\theta}(\underline{G}_{\hat\theta})} \tag{65}$$

under Assumptions 9 and 11, where $F_{\hat\theta} \coloneqq p_1F_{\hat\theta,1}+(1-p_1)F_{\hat\theta,0}$ is the distribution function of $g(\hat\theta^\intercal X)$ given $\hat\theta$. By Assumption 6, $F_{\hat\theta}$ is continuous, and by arguments on $\hat A_n$ above $\underline{G}_{\hat\theta} = \underline{p}_{\hat\theta} + o_P(1)$. Then $a_n \downarrow 0$ implies that (65) is $o_P(1)$, which in turn implies that (63), being bounded by one, is also $o_P(1)$. Term (64) is $o_P(1)$ by similar arguments, noting that $\left\{\max_{i\in[n]} g(\hat\theta^\intercal X_i)) - a_n < g(\hat\theta^\intercal X_i)\right\} = \left\{\min_{i\in[n]} -g(\hat\theta^\intercal X_i)) + a_n > -g(\hat\theta^\intercal X_i)\right\}$. Thus, (62) is $o_P(1)$. Conclude that (58) converges in probability to $V_\tau+\tau^2$. Write (59) as

$$\frac{1}{n}\sum_{i\in[n]}[\hat\mu^1(\hat\theta,\hat\pi_i)-\mu^1(\hat\theta,\hat\pi_i)]^2\hat{\mathbb{1}}_i + \frac{2}{n}\sum_{i\in[n]}(\hat\mu^1(\hat\theta,\hat\pi_i)-\mu^1(\hat\theta,\hat\pi_i))(\hat\mu^0(\hat\theta,\hat\pi_i)-\mu^0(\hat\theta,\hat\pi_i))\hat{\mathbb{1}}_i + \frac{1}{n}\sum_{i\in[n]}[\hat\mu^0(\hat\theta,\hat\pi_i)-\mu^0(\hat\theta,\hat\pi_i)]^2\hat{\mathbb{1}}_i.$$

As Assumption 10(i) bounds both $\hat\mu^d$ and μ^d, $\sup_{p\in A_n}|\hat\mu^d(\hat\theta,p)-\mu^d(\hat\theta,p)| \xrightarrow{P} 0$ implies that all three terms in the last display are $o_P(1)$ (note that $\hat A_n\subset A_n$). This convergence can be established along the same lines as that of other estimators, which are detailed below. Similar arguments show that (60) is $o_P(1)$. Conclude that $\hat V_\tau \xrightarrow{P} V_\tau$.

Consistency of $\hat V_{\sigma,\pi}$. First, $0 < g(\hat\theta^\intercal x) < 1$ for all x in compact $\mathcal{X}$ and for all $\hat\theta\in \mathrm{Nb}(\theta_0,\epsilon)$. Then under Assumption 11, a mean-value expansion and $\hat\theta \xrightarrow{P} \theta_0$, implied by Assumption 9, gives that $\sup_{x\in\mathcal{X}} |1/g(\hat\theta^\intercal x) - 1/g(\theta_0^\intercal x)| = o_P(1)$ and similarly for $1/(1-g(\hat\theta^\intercal x))$. Second, the Lipschitz condition in Assumption 8(i) implies

$$\left|\frac{1}{n}\sum_{i\in[n]} \sigma_d^2(\hat\theta, \hat\pi_i)\hat{\mathbb{1}}_i - \frac{1}{n}\sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i)\hat{\mathbb{1}}_i\right| \xrightarrow{P} 0,$$

and, as in (61) above, we also have

$$\left|\frac{1}{n}\sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i)\hat{\mathbb{1}}_i - \frac{1}{n}\sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i)\right| \xrightarrow{P} 0.$$

Thus, if $\sup_{p\in A_n} |\hat\sigma_d^2(\hat\theta, p) - \sigma_d^2(\hat\theta, p)| \xrightarrow{P} 0$, then the law of large numbers gives that $\hat V_{\sigma,\pi} \xrightarrow{P} V_{\sigma,\pi}$. This holds if both $\sup_{p\in A_n} |\hat\mu^d(\hat\theta, p) - \mu^d(\hat\theta, p)| \xrightarrow{P} 0$ and $\sup_{p\in A_n} |\hat\mu_2^d(\hat\theta, p) - \mu_2^d(\hat\theta, p)| \xrightarrow{P} 0$ since the former implies $\sup_{p\in A_n} |(\hat\mu^d(\hat\theta, p))^2 - (\mu^d(\hat\theta, p))^2| \xrightarrow{P} 0$ under Assumption 10(i). See below for a proof of such convergence (e.g. establishing (103)).

Consistency of $\hat q_d$. By definition of q_d, the conditions

$$\|\frac{1}{n}\sum_{i\in[n]} \Lambda^d(\hat\theta, X_i) - \frac{1}{n}\sum_{i\in[n]} \Lambda^d(\hat\theta, X_i)\hat{\mathbb{1}}_i\| \xrightarrow{P} 0 \tag{66}$$

$$\sup_{p\in A_n} \left|\widehat{\left(\frac{\partial\mu^d}{\partial p}\right)}(\hat\theta, p) - \frac{\partial\mu^d}{\partial p}(\hat\theta, p)\right| \xrightarrow{P} 0 \tag{67}$$

$$\sup_{p\in A_n} \left|\widehat{\left(\frac{\partial\mu^d}{\partial\theta_k}\right)}(\hat\theta, p) - \frac{\partial\mu^d}{\partial\theta_k}(\hat\theta, p)\right| \xrightarrow{P} 0 \quad \text{for all } k = 1, 2, \ldots, K, \tag{68}$$

together with Assumptions 10 and 11, implying the boundedness of $\mathcal{X}$ and g', ensure $\hat q_d \xrightarrow{P} q_d$. Assumption 7, 11(i), and Assumption 10 imply that Λ^d is bounded uniformly. Therefore, (66) is satisfied, because the arguments treating $\hat{\mathbb{1}}_i$ in the case of $\hat V_\tau$ above (e.g. (61)) apply.

In the following, we show that (67) and (68) hold, wherein we also give a detailed proof of the uniform consistency of the nonparametric estimators assumed above. For simplicity of exposition, we only give the proof for the (derivatives of) $\mu(\theta, p) := \mathbb{E}[Y \mid g(\theta^\intercal X) = p] = \mathbb{E}[Y \mid \theta^\intercal X = g^{-1}(p)]$ and $\mu_2(\theta, p) := \mathbb{E}[Y^2 \mid g(\theta^\intercal X) = p]$. The proof when we also condition on $D = d$ follows along the same lines. To this end, let $h(\theta, p) := p_1 h_1(\theta, p) + (1 - p_1)h_0(\theta, p)$ be the density of $\pi(X, \theta)$, $q_\mu(\theta, p) := \mu(\theta, p)h(\theta, p)$, $q_{\mu_2}(\theta, p) := \mu_2(\theta, p)h(\theta, p)$ and their corresponding estimators be obtained by setting $\mathbb{1}_{D_j=d} := 1$ for all $j \in [n]$ in the formulae for $\hat h_d$, $\hat q_{\mu,d}$, $\hat q_{\mu_2,d}$, respectively. For short, we also let $h'(\theta, p) := (\partial/\partial p)h(\theta, p)$, $\hat h'(\theta, p) := (\partial/\partial p)\hat h(\theta, p)$ and likewise for $q'_\mu, \hat q'_\mu$. Furthermore, we let $\widehat{\left(\frac{\partial h}{\partial\theta_k}\right)}$ and $\widehat{\left(\frac{\partial q_\mu}{\partial\theta_k}\right)}$ be obtained by setting $\mathbb{1}_{D_j=d} := 1$ in the formula for $\widehat{\left(\frac{\partial h_d}{\partial\theta_k}\right)}$ and $\widehat{\left(\frac{\partial q_{\mu,d}}{\partial\theta_k}\right)}$, respectively.

For short, we put $\dot{h}_k(\theta, p) := (\partial/\partial\theta_k)h(\theta, p)$, $\hat{\dot{h}}_k(\theta, p) := \widehat{\left(\frac{\partial h}{\partial \theta_k}\right)}(\theta, p)$ and likewise for $\dot{q}_{\mu,k}(\theta, p), \hat{\dot{q}}_{\mu,k}(\theta, p)$. For a function $r : \Theta \times [0, 1] \to \mathbb{R}$, we let $\|r\|_{A_n} := \sup_{p\in A_n} |r(\hat{\theta}, p)|$.

Condition (67). First we show that the numerator of

$$\widehat{\left(\frac{\partial \mu}{\partial p}\right)}(\hat{\theta}, p) := \frac{\hat{q}'_\mu(\hat{\theta}, p)\hat{h}(\hat{\theta}, p) - \hat{q}_\mu(\hat{\theta}, p)\hat{h}'(\hat{\theta}, p)}{(\hat{h}(\hat{\theta}, p))^2} \tag{69}$$

converges to that of

$$\frac{\partial \mu}{\partial p}(\hat{\theta}, p) = \frac{q'_\mu(\hat{\theta}, p)h(\hat{\theta}, p) - q_\mu(\hat{\theta}, p)h'(\hat{\theta}, p)}{h(\hat{\theta}, p)^2}. \tag{70}$$

Consider

$$\begin{aligned}
\|\hat{q}'_\mu\hat{h} - \hat{q}_\mu\hat{h}' - (q'_\mu h - q_\mu h')\|_{A_n} &\le \|\hat{q}'_\mu\hat{h} - q'_\mu h\|_{A_n} + \|\hat{q}_\mu\hat{h}' - q_\mu h'\|_{A_n} \\
\|\hat{q}'_\mu\hat{h} - q'_\mu h\|_{A_n} &= \|(\hat{q}'_\mu - q'_\mu + q'_\mu)(\hat{h} - h + h) - q'_\mu h\|_{A_n} \\
&\le \|\hat{q}'_\mu - q'_\mu\|_{A_n}\|\hat{h} - h\|_{A_n} + \|\hat{q}'_\mu - q'_\mu\|_{A_n}\|h\|_{A_n} \\
&\quad + \|q'_\mu\|_{A_n}\|\hat{h} - h\|_{A_n} \\
\|\hat{q}_\mu\hat{h}' - q_\mu h'\|_{A_n} &\le \|\hat{h}' - h'\|_{A_n}\|\hat{q}_\mu - q_\mu\|_{A_n} + \|\hat{h}' - h'\|_{A_n}\|q_\mu\|_{A_n} \\
&\quad + \|h'\|_{A_n}\|\hat{q}_\mu - q_\mu\|_{A_n}.
\end{aligned}$$

By Assumption 6(iv), $\|h\|_{A_n}$ is bounded with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$ and combining it with Assumptions 7 and 11(ii), the same holds for $\|h'\|_{A_n}$, $\|q_\mu\|_{A_n} = \|\mu h\|_{A_n}$ and $\|q'_\mu\|_{A_n} = \|\mu' h + \mu h'\|_{A_n}$. It follows that if all

$$\|\hat{q}'_\mu - q'_\mu\|_{A_n} \xrightarrow{P} 0, \tag{71}$$

$$\|\hat{h}' - h'\|_{A_n} \xrightarrow{P} 0, \tag{72}$$

$$\|\hat{q}_\mu - q_\mu\|_{A_n} \xrightarrow{P} 0, \tag{73}$$

$$\|\hat{h} - h\|_{A_n} \xrightarrow{P} 0, \tag{74}$$

then the numerator of (69) converges to that of (70) uniformly in $p \in A_n$. The denominator of (69) satisfies

$$\|(\hat{h})^2 - h^2\|_{A_n} = \|(\hat{h} - h)(\hat{h} + h + h - h)\|_{A_n} \le \|\hat{h} - h\|_{A_n}\left\{\|\hat{h} - h\|_{A_n} + 2\|h\|_{A_n}\right\}.$$

By Assumption 6(iv), $\|h\|_{A_n} < \infty$ with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$, hence (74) implies that the denominator of (69) converges to that of (70) uniformly in $p \in A_n$. Thus, both the numerator and the denominator of (69) converges to those of (70). Because

$\inf_{p\in[\underline{p}_{\hat\theta},\bar p_{\hat\theta}]} h(\hat\theta,p)>0$ for all $\hat\theta\in \mathrm{Nb}(\theta_0,\epsilon)$, it follows that (69) converges to (70) uniformly in $p\in A_n$. In the following, we show that (71)–(74) hold.

Condition (67)*, part* (71). The proof consists in showing

$$\mathbb{E}\mathbb{E}\left[\sup_{p\in A_n}\left|\hat q_\mu'(\hat\theta,p)-\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right]\right|\,\middle|\,\hat\theta\right]\to 0 \quad \text{and} \tag{75}$$

$$\sup_{p\in A_n}\left|\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right]-q_\mu'(\hat\theta,p)\right|\xrightarrow{P}0. \tag{76}$$

We show (75) following Bierens (1994). For the imaginary unit i, let $\psi(t):=\int e^{itx}K(x)\mathrm{d}x$, $t\in\mathbb{R}$, be the characteristic function of K, which is $\psi(t)=e^{-t^2/2}$ for the Gaussian kernel K, so that $K(x)=(2\pi)^{-1}\int e^{-itx}\psi(t)\mathrm{d}t$ by the inversion formula for characteristic functions. Then $K'(x)=(2\pi)^{-1}\int(-it)e^{-itx}\psi(t)\mathrm{d}t$, hence

$$\begin{aligned}\hat q_\mu'(\hat\theta,p)&=-\frac{1}{n\gamma_n^2}\sum_{j\in[n]}Y_j(2\pi)^{-1}\int(-it)e^{-it(g(\hat\theta^\intercal X_j)-p)/\gamma_n}\psi(t)\mathrm{d}t\\&=(2\pi)^{-1}\int\left(\frac1n\sum_{j\in[n]}Y_je^{-itg(\hat\theta^\intercal X_j)}\right)e^{itp}it\psi(\gamma_nt)\mathrm{d}t,\end{aligned}$$

where we used a change of variables and Fubini's theorem (the integral is bounded as $|Y_j|<\bar y$ almost surely by Assumption 10(i), $|it|\le 1$, $\sup_{t\in\mathbb{R}}|\psi(t)|<\infty$ and the exponential is bounded too). As $|e^{itp}|\le 1$ and $|it|\le|t|$, we have that

$$\begin{aligned}&\mathbb{E}\left[\sup_{p\in A_n}\left|\hat q_\mu'(\hat\theta,p)-\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right]\right|\,\middle|\,\hat\theta\right]\\&\le(2\pi)^{-1}\int\mathbb{E}\left[\left|\left(\frac1n\sum_{j\in[n]}Y_je^{-itg(\hat\theta^\intercal X_j)}\right)-\mathbb{E}\left[Y_je^{-itg(\hat\theta^\intercal X_j)}\,\middle|\,\hat\theta\right]\right|\,\middle|\,\hat\theta\right]|t||\psi(\gamma_nt)|\mathrm{d}t.\end{aligned}$$

As $e^{-ia}=\cos(a)-i\sin(a)$ for $a\in\mathbb{R}$, and $\mathbb{E}|W|\le\sqrt{\mathbb{E}W^2}$ for any square-integrable random variable W, the expectation on the right of the last display is bounded by

$$\mathbb{V}\left[\frac1n\sum_{j\in[n]}Y_j\cos(g(\hat\theta^\intercal X_j))\,\middle|\,\hat\theta\right]^{1/2}+\mathbb{V}\left[\frac1n\sum_{j\in[n]}Y_j\sin(g(\hat\theta^\intercal X_j))\,\middle|\,\hat\theta\right]^{1/2}\le 2\sqrt{\frac{\mathbb{E}Y_j^2}{n}},$$

where we used that by Assumption 9(ii) the elements in the sum are i.i.d. given $\hat\theta$, so the covariances are zero, and that $\mathbb{V}\left[Y_j\cos(g(\hat\theta^\intercal X_j))\,\middle|\,\hat\theta\right]\le\mathbb{E}\left[Y_j^2\,\middle|\,\hat\theta\right]=\mathbb{E}Y_j^2$. Thus,

$$\begin{aligned}\mathbb{E}\left[\sup_{p\in A_n}\left|\hat q_\mu'(\hat\theta,p)-\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right]\right|\,\middle|\,\hat\theta\right]&\le\sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2n}}\int|t||\psi(\gamma_nt)|\mathrm{d}t\\&\le\sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2n\gamma_n^4}}\int|t||\psi(t)|\mathrm{d}t.\end{aligned}$$

As $\mathbb{E}Y_j^2 < \infty$ by Assumption 10(i) and $\int |t||\psi(t)|\mathrm{d}t = \int |t|e^{-t^2/2} < \infty$ for the Gaussian K, the right side is of the order $1/(\gamma_n^2\sqrt{n}) = (\kappa_0)^{-2}n^{2\beta-1/2} = o(1)$ for $\beta < 1/4$.

Next, we show (76). As the summands are identically distributed given $\hat\theta$ by Assumption 9(ii), the tower property of expectations gives

$$\begin{aligned}
\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right] &= -\frac{1}{\gamma_n^2}\mathbb{E}\left[\mathbb{E}\left[Y \mid g(\hat\theta^\intercal X),\hat\theta\right]K'((g(\hat\theta X)-p)/\gamma_n)\,\middle|\,\hat\theta\right]\\
&= -\frac{1}{\gamma_n^2}\mathbb{E}\left[\mu(\hat\theta, g(\hat\theta^\intercal X))K'((g(\hat\theta X)-p)/\gamma_n)\,\middle|\,\hat\theta\right]\\
&= -\frac{1}{\gamma_n^2}\int_{\underline{p}_{\hat\theta}}^{\bar p_{\hat\theta}} \mu(\hat\theta,\tilde p)K'((\tilde p - p)/\gamma_n)h(\hat\theta,\tilde p)\mathrm{d}\tilde p\\
&= -\frac{1}{\gamma_n^2}\int_{\underline{p}_{\hat\theta}}^{\bar p_{\hat\theta}} q_\mu(\hat\theta,\tilde p)K'((\tilde p - p)/\gamma_n)\mathrm{d}\tilde p\\
&= -\frac{1}{\gamma_n}\int_{(\underline{p}_{\hat\theta}-p)/\gamma_n}^{(\bar p_{\hat\theta}-p)/\gamma_n} q_\mu(\hat\theta,\gamma_n v + p)K'(v)\mathrm{d}v
\end{aligned}$$

by definition of $h(\hat\theta,\cdot)$ as the density of $(g(\hat\theta^\intercal X) \mid \hat\theta)$ under Assumptions 9(ii) and 11(i), and $q_\mu = \mu h$. Integration by parts gives

$$\mathbb{E}\left[\hat q_\mu'(\hat\theta,p)\,\middle|\,\hat\theta\right] = -\frac{1}{\gamma_n}\left\{q(\hat\theta,\bar p_{\hat\theta})K((\bar p_{\hat\theta}-p)/\gamma_n) - q(\hat\theta,\underline{p}_{\hat\theta})K((\underline{p}_{\hat\theta}-p)/\gamma_n)\right\} \tag{77}$$

$$+\int_{(\underline{p}_{\hat\theta}-p)/\gamma_n}^{(\bar p_{\hat\theta}-p)/\gamma_n} q_\mu'(\hat\theta,\gamma_n v + p)K(v)\mathrm{d}v. \tag{78}$$

As $q_\mu(\hat\theta,\cdot)$ is bounded for $\hat\theta \in \mathrm{Nb}(\theta_0,\epsilon)$, with probability tending to one (77) is of the order

$$\gamma_n^{-1}\left[\sup_{p\in A_n} K((\bar p_{\hat\theta}-p)/\gamma_n) + \sup_{p\in A_n} K((\underline{p}_{\hat\theta}-p)/\gamma_n)\right] = 2\gamma_n^{-1}K(a_n/\gamma_n).$$

As $\gamma_n^{-1}K(a_n/\gamma_n) \to 0$, (77) is $o_P(1)$. Now we show that (78) converges uniformly to $q_\mu'(\hat\theta,p)$ adapting the proof of Schuster and Yakowitz (1979, Lemma 1). The kernel K being a density integrating to one implies

$$q_\mu'(\hat\theta,p) = q_\mu'(\hat\theta,p)\gamma_n^{-1}\left\{\int_{-\infty}^{p-\bar p_{\hat\theta}} K(u/\gamma_n)\mathrm{d}u + \int_{p-\bar p_{\hat\theta}}^{p-\underline{p}_{\hat\theta}} K(u/\gamma_n)\mathrm{d}u + \int_{p-\underline{p}_{\hat\theta}}^{\infty} K(u/\gamma_n)\mathrm{d}u\right\}.$$

Combine this with a change of variables in (78) (with $u := -\gamma_n v$ noting that K is symmetric about zero), to get

$$\begin{aligned} &\sup_{p\in A_n}\left|\int_{(\underline{p}_{\hat\theta}-p)/\gamma_n}^{(\bar p_{\hat\theta}-p)/\gamma_n} q'_\mu(\hat\theta,\gamma_n v+p)K(v)\mathrm{d}v - q'_\mu(\hat\theta,p)\right| \\ &\le \sup_{p\in A_n}\left|\int_{-\infty}^{p-\bar p_{\hat\theta}} q'_\mu(\hat\theta,p)\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| + \sup_{p\in A_n}\left|\int_{p-\underline{p}_{\hat\theta}}^{\infty} q'_\mu(\hat\theta,p)\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \qquad (79) \\ &\quad + \sup_{p\in A_n}\left|\int_{p-\bar p_{\hat\theta}}^{p-\underline{p}_{\hat\theta}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \qquad (80)\end{aligned}$$

The two terms in (79) are $o_P(1)$. The first one is bounded by

$$\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{(p-\bar p_{\hat\theta})/\gamma_n} K(v)\mathrm{d}v = \sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{-a_n/\gamma_n} K(v)\mathrm{d}v,$$

which vanishes as, on one hand, $\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|<\infty$ with probability tending to one as $\hat\theta \xrightarrow{P} \theta_0$ by Assumptions 6(iv) and 7, and, on the other hand, $a_n/\gamma_n\to\infty$ implies that the integral of the Gaussian K goes to zero. The second term in (79) is bounded by

$$\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{(p-\underline{p}_{\hat\theta})/\gamma_n}^{\infty} K(v)\mathrm{d}v = \sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{a_n/\gamma_n}^{\infty} K(v)\mathrm{d}v,$$

so it is also $o_P(1)$ by the same argument. To show that (80) also vanishes, let $\rho_n>0$ be a sequence satisfying $\rho_n/\gamma_n\to\infty$ and $\rho_n<a_n$, i.e. $\gamma_n\ll\rho_n<a_n$ (as $\gamma_n=\kappa_0 n^{-\beta}$ and $a_n=\kappa_1 n^{-\alpha}$, $\beta>\alpha$, we can take $\rho_n=\kappa_2 n^{-(\beta+\alpha)/2}$, $0<\kappa_2<\kappa_1$). Then (80) is bounded by

$$\begin{aligned} &\sup_{p\in A_n}\left|\int_{[p-\bar p_{\hat\theta},p-\underline{p}_{\hat\theta}]\cap\{|u|\le\rho_n\}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \qquad (81) \\ &+ \sup_{p\in A_n}\left|\int_{[p-\bar p_{\hat\theta},p-\underline{p}_{\hat\theta}]\cap\{|u|>\rho_n\}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]u^{-1}u\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \qquad (82)\end{aligned}$$

Here, (81) is bounded by

$$\begin{aligned} &\sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{\infty}\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u \\ &\le \sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|.\end{aligned}$$

By Assumptions 6(iv) and 7, $q'_\mu(\hat\theta,\cdot)$ is continuous on the compact set $[\underline{p}_{\hat\theta},\bar p_{\hat\theta}]\supset A_n$ and is therefore uniformly continuous. Thus, $|u|\le\rho_n$ for a small enough $\rho_n<a_n$ implies that $p-u,p\in[\underline{p}_{\hat\theta},\bar p_{\hat\theta}]$ for all $p\in A_n$. Hence, $\sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|=o_P(1)$.

In the integral of (82), $u \in [p - \bar{p}_{\hat{\theta}}, p - \underline{p}_{\hat{\theta}}]$, so $p - u \in [\underline{p}_{\hat{\theta}}, \bar{p}_{\hat{\theta}}]$, and thus (82) is bounded by $2\sup_{p\in[\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}]} |q'_\mu(\hat{\theta}, p)|$ times

$$\sup_{p\in A_n} \left| \int_{[p-\bar{p}_{\hat{\theta}},p-\underline{p}_{\hat{\theta}}]\cap\{|u|>\rho_n\}} \frac{K(u/\gamma_n)}{\gamma_n} \mathrm{d}u \right| \leq \sup_{p\in A_n} \int_{\{|u|>\rho_n\}} \frac{K(u/\gamma_n)}{\gamma_n} \mathrm{d}u \leq \int_{\{|v|>\rho_n/\gamma_n\}} K(v)\mathrm{d}v,$$

where we used that $K > 0$. As $\gamma_n \ll \rho_n$, and $K(v) \downarrow 0$ as $|v| \to \infty$, the right integral tends to zero. As Assumptions 6(iv) and 7 control $\sup_{p\in[\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}]} |q'_\mu(\hat{\theta}, p)|$, (82) is $o_P(1)$. Thus, (76) holds. Conclude that (71) holds.

Condition (67), *part* (72). Follows directly along the lines of (71), setting $Y_j := 1$ for all $j \in [n]$, in the formulae of (71).

Condition (67), *part* (73). Analogously to $\hat{q}'_\mu$ above, we can write

$$\begin{aligned}\hat{q}_\mu(\hat{\theta}, p) &= \frac{1}{n\gamma_n} \sum_{j\in[n]} Y_j (2\pi)^{-1} \int e^{-it(g(\hat{\theta}^\intercal X_j)-p)/\gamma_n} \psi(t)\mathrm{d}t \\ &= (2\pi)^{-1} \int \left(\frac{1}{n} \sum_{j\in[n]} Y_j e^{-itg(\hat{\theta}^\intercal X_j)} \right) e^{itp} \psi(\gamma_n t)\mathrm{d}t,\end{aligned}$$

and then

$$\mathbb{E}\left[\sup_{p\in A_n} \left| \hat{q}_\mu(\hat{\theta}, p) - \mathbb{E}\left[\hat{q}_\mu(\hat{\theta}, p) \,\middle|\, \hat{\theta} \right] \right| \,\middle|\, \hat{\theta} \right] \leq \sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2 n}} \int |\psi(\gamma_n t)|\mathrm{d}t \leq \sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2 n \gamma_n^2}} \int |\psi(t)|\mathrm{d}t.$$

Assumption 10(i) and $\int |\psi(t)|\mathrm{d}t = \int e^{-t^2/2} = 2\pi$ for the Gaussian kernel K mean that the right side is of the order $1/(\gamma_n\sqrt{n}) = (\kappa_0)^{-1} n^{\beta-1/2} = o(1)$ for $\beta < 1/4$. Next, as for $\hat{q}'_\mu$,

$$\begin{aligned}\mathbb{E}\left[\hat{q}_\mu(\hat{\theta}, p) \,\middle|\, \hat{\theta} \right] &= \frac{1}{\gamma_n} \int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}} \mu(\hat{\theta}, \tilde{p}) K((\tilde{p} - p)/\gamma_n) h(\hat{\theta}, \tilde{p})\mathrm{d}\tilde{p} \\ &= \frac{1}{\gamma_n} \int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}} q_\mu(\hat{\theta}, \tilde{p}) K((\tilde{p} - p)/\gamma_n)\mathrm{d}\tilde{p} = \int_{(\underline{p}_{\hat{\theta}}-p)/\gamma_n}^{(\bar{p}_{\hat{\theta}}-p)/\gamma_n} q_\mu(\hat{\theta}, \gamma_n v + p) K(v)\mathrm{d}v.\end{aligned}$$

This converges uniformly to $q_\mu(\hat{\theta}, p)$ in $p \in A_n$ by the same arguments as (78) does to $q'_\mu(\hat{\theta}, p)$, given Assumptions 6 and 7 ensuring the boundedness and continuity of $q_\mu(\hat{\theta}, \cdot)$ with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$.

Condition (67), *part* (74). Follows from (73) by setting $Y_j := 1$ for all $j \in [n]$.

Condition (68). To establish the uniform convergence of

$$\widehat{\left(\frac{\partial \mu}{\partial \theta_k}\right)}(\hat{\theta}, p) := \frac{\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p)\hat{h}(\hat{\theta}, p) - \hat{q}_\mu(\hat{\theta}, p)\hat{\dot{h}}_k(\hat{\theta}, p)}{(\hat{h}(\hat{\theta}, p))^2} \tag{83}$$

to

$$\frac{\partial \mu}{\partial \theta_k}(\hat{\theta}, p) := \frac{\dot{q}_{\mu,k}(\hat{\theta}, p)h(\hat{\theta}, p) - \hat{q}_\mu(\hat{\theta}, p)\dot{h}_k(\hat{\theta}, p)}{(\hat{h}(\hat{\theta}, p))^2} \tag{84}$$

in $p \in A_n$, we can follow the same steps which led to (71)–(74) of Condition (67), because Assumptions 6(iv) and 7 ensure that with probability tending to one, $\|h\|_{A_n}$, $\|\dot{h}_k\|_{A_n}$, $\|q_\mu\|_{A_n}$, $\|\dot{q}_{\mu,k}\|_{A_n}$ are all bounded. As (73) and (74) were proved above, it is sufficient to show both

$$\|\hat{\dot{q}}_{\mu,k} - \dot{q}_{\mu,k}\|_{A_n} \xrightarrow{P} 0 \quad \text{and} \tag{85}$$

$$\|\hat{\dot{h}}_k - \dot{h}_k\|_{A_n} \xrightarrow{P} 0. \tag{86}$$

Condition (68)*, part* (85)*.* We proceed by showing

$$\mathbb{E}\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p) - \mathbb{E}\left[\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p) \,\middle|\, \hat{\theta}\right]\right| \,\middle|\, \hat{\theta}\right] \to 0 \quad \text{and} \tag{87}$$

$$\sup_{p\in A_n}\left|\mathbb{E}\left[\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p) \,\middle|\, \hat{\theta}\right] - \dot{q}_{\mu,k}(\hat{\theta}, p)\right| \xrightarrow{P} 0. \tag{88}$$

As for $\hat{q}'_\mu$ above,

$$\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p) = \frac{(g^{-1})'(p)}{n\gamma_n^2}\sum_{j\in[n]} Y_j X_{j,k}(2\pi)^{-1}\int(-it)e^{-it(\hat{\theta}^\intercal X_j - g^{-1}(p))/\gamma_n}\psi(t)\mathrm{d}t \tag{89}$$

$$= (g^{-1})'(p)(2\pi)^{-1}\int\left(\frac{1}{n}\sum_{j\in[n]} Y_j X_{j,k} e^{-it\hat{\theta}^\intercal X_j}\right)e^{itg^{-1}(p)}(-it)\psi(\gamma_n t)\mathrm{d}t. \tag{90}$$

Since $|-it| \le |t|$ and $\mathcal{X}$ is bounded by Assumption 10, Euler's formula and Assumption 9(ii) imply that the bound of (75) also apply here up to a constant, which bounds $\mathcal{X}$, times $\sup_{p\in A_n}|(g^{-1})'(p)|$. As $\sup_{p\in A_n}|(g^{-1})'(p)| \le 1/\|g'\|_\infty < \infty$ by Assumption 11, (87) holds.

Condition (68)*, part* (85)*,* (88)*.* We begin by deriving $\dot{q}_{\mu,k}(\hat{\theta}, p) = ((\partial/\partial\theta_k)(\mu h))(\hat{\theta}, p)$. For simplicity, assume that we only have two covariates, both continuous, and we are interested in the derivative with respect to the first coordinate of θ ($k := 1$). It is straightforward to generalise the arguments below for the general case. By the tower property of expectation,

$$\mu(\theta, p) = \mathbb{E}\left[Y \mid \theta^\intercal X = g^{-1}(p)\right] = \mathbb{E}\left[m(X) \mid \theta^\intercal X = g^{-1}(p)\right].$$

In view of (30) of Proposition 6, we can write

$$\mathbb{E}\left[m(X) \mid \theta^\intercal X = t\right] = \frac{1}{v(\theta, t)}\int_{\mathcal{X}_1} m\left(x_1, \frac{t - \theta_1 x_1}{\theta_2}\right)\Psi\left(x_1, \frac{t - \theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1 \tag{91}$$

$$v(\theta, t) := \int_{\mathcal{X}_1}\Psi\left(x_1, \frac{t - \theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1, \tag{92}$$

where $v(\theta,\cdot)$ is the density of $\theta^\intercal X$ satisfying $v(\theta,\cdot)>0$ by Assumption 10. Thus, $h(\theta,\cdot)$, being the density of $g(\theta^\intercal X)$, is equal to

$$h(\theta,p)=(g^{-1})'(p)\int_{\mathcal{X}_1}\Psi\left(x_1,\frac{g^{-1}(p)-\theta_1x_1}{\theta_2}\right)\mathrm{d}x_1. \tag{93}$$

By Assumption 9(ii), (91) and (93) remain valid once we replace θ with $\hat\theta$. It then follows for continuously differentiable Ψ and m (Assumptions 10 and 11), that for $k=1$,

$$\begin{aligned}
q_\mu(\theta,p)&=\mu(\theta,p)h(\theta,p)\\
&=(g^{-1})'(p)\int_{\mathcal{X}_1}m\left(x_1,\frac{g^{-1}(p)-\theta_1x_1}{\theta_2}\right)\Psi\left(x_1,\frac{g^{-1}(p)-\theta_1x_1}{\theta_2}\right)\mathrm{d}x_1\\
\dot q_{\mu,k}(\hat\theta,p)&=(g^{-1})'(p)\int_{\mathcal{X}_1}\frac{\mathrm{d}}{\mathrm{d}\theta_1}\left[m\left(x_1,\frac{g^{-1}(p)-\hat\theta_1x_1}{\hat\theta_2}\right)\Psi\left(x_1,\frac{g^{-1}(p)-\hat\theta_1x_1}{\hat\theta_2}\right)\right]\mathrm{d}x_1\\
&=\frac{\mathrm{d}}{\mathrm{d}p}\int_{\mathcal{X}_1}-x_1m\left(x_1,\frac{g^{-1}(p)-\hat\theta_1x_1}{\hat\theta_2}\right)\Psi\left(x_1,\frac{g^{-1}(p)-\hat\theta_1x_1}{\hat\theta_2}\right)\mathrm{d}x_1,
\end{aligned}\tag{94}$$

where in the last step we used that $(g^{-1})'>0$. We proceed by showing the desired convergence. Let $[\underline{t}_{\hat\theta},\bar t_{\hat\theta}]\coloneqq[g^{-1}(\underline{p}_{\hat\theta}),g^{-1}(\bar p_{\hat\theta})]$. We have for $k=1$ by the tower property

$$\begin{aligned}
\mathbb{E}\left[\hat{\dot q}_{\mu,k}(\hat\theta,p)\,\middle|\,\hat\theta\right]&=\frac{g^{-1}(p)}{\gamma_n^2}\mathbb{E}\left[Y_jX_{j,k}K'((\hat\theta^\intercal X_j-g^{-1}(p)/\gamma_n)\,\middle|\,\hat\theta\right]\\
=\frac{(g^{-1})'(p)}{\gamma_n^2}&\mathbb{E}\left[\mathbb{E}\left[m(X_j)X_{j,1}\mid\hat\theta^\intercal X_j,\hat\theta\right]K'((\hat\theta^\intercal X_j-g^{-1}(p))/\gamma_n)\,\middle|\,\hat\theta\right]\\
=\frac{(g^{-1})'(p)}{\gamma_n^2}\int_{\underline{t}_{\hat\theta}}^{\bar t_{\hat\theta}}K'\left(\frac{t-g^{-1}(p)}{\gamma_n}\right)&\int_{\mathcal{X}_1}x_1m\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)\Psi\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)\mathrm{d}x_1\,\mathrm{d}t\\
&=(g^{-1})'(p)\\
\times\int_{\mathcal{X}_1}x_1\left[\frac{1}{\gamma_n^2}\int_{\underline{t}_{\hat\theta}}^{\bar t_{\hat\theta}}K'\left(\frac{t-g^{-1}(p)}{\gamma_n}\right)\right.&\left.m\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)\Psi\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)\mathrm{d}t\right]\mathrm{d}x_1,
\end{aligned}$$

where we used that (91) holds not only for m, but for any generic function of X by Proposition 6. Let $\lambda_n(x_1,t)\coloneqq m\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)\Psi\left(x_1,\frac{t-\hat\theta_1x_1}{\hat\theta_2}\right)$. We show that the term in the square brackets converges to

$$-\frac{\partial}{\partial t}\lambda_n(x_1,g^{-1}(p))=-\left(\frac{\mathrm{d}}{\mathrm{d}p}\lambda_n(x_1,g^{-1}(p))\right)/[(g^{-1})'(p)]$$

uniformly in $x_1 \in \mathcal{X}_1$ and $p \in A_n$, which completes the proof for (88) in light of (94). Put $\lambda_n'(x_1,t) := \frac{\partial}{\partial t}\lambda_n(x_1,t)$. Integration by parts gives

$$\frac{1}{\gamma_n^2}\int_{\underline{t}_{\hat\theta}}^{\bar t_{\hat\theta}} K'\left(\frac{t-g^{-1}(p)}{\gamma_n}\right)\lambda_n(x_1,t)\mathrm{d}t = \frac{1}{\gamma_n}\int_{(\underline{t}_{\hat\theta}-g^{-1}(p))/\gamma_n}^{(\bar t_{\hat\theta}-g^{-1}(p))/\gamma_n} K'(v)\lambda_n(x_1,\gamma_n v+g^{-1}(p))\mathrm{d}t$$

$$= \frac{1}{\gamma_n}\left\{\lambda_n(x_1,\bar t_{\hat\theta})K((\bar t_{\hat\theta}-g^{-1}(p))/\gamma_n) - \lambda_n(x_1,\underline{t}_{\hat\theta})K((\underline{t}_{\hat\theta}-g^{-1}(p))/\gamma_n)\right\} \tag{95}$$

$$-\int_{(\underline{t}_{\hat\theta}-g^{-1}(p))/\gamma_n}^{(\bar t_{\hat\theta}-g^{-1}(p))/\gamma_n} \lambda_n'(x_1,\gamma_n v+g^{-1}(p))K(v)\mathrm{d}v. \tag{96}$$

Now λ_n is bounded with probability tending to one by Assumption 10. Recall that $\underline{t}_{\hat\theta} = g^{-1}(\underline{p}_{\hat\theta}), \bar t_{\hat\theta} = g^{-1}(\bar p_{\hat\theta})$. As g^{-1} is increasing, and K reaches its maximum at zero and satisfies $\lim_{|u|\to\infty} K(u)\to 0$, (95) is of the order

$$\gamma_n^{-1}\left[\sup_{p\in A_n} K\left(\frac{g^{-1}(\bar p_{\hat\theta})-g^{-1}(p)}{\gamma_n}\right) + \sup_{p\in A_n} K\left(\frac{g^{-1}(\underline{p}_{\hat\theta})-g^{-1}(p)}{\gamma_n}\right)\right]$$

$$= \gamma_n^{-1}\left[K\left(\frac{g^{-1}(\bar p_{\hat\theta})-g^{-1}(\bar p_{\hat\theta}-a_n)}{\gamma_n}\right) + K\left(\frac{g^{-1}(\underline{p}_{\hat\theta})-g^{-1}(\underline{p}_{\hat\theta}+a_n)}{\gamma_n}\right)\right]. \tag{97}$$

By a mean-value expansion, the first term in (97) is of the order

$$\gamma_n^{-1}K\left(\left(\inf_{p\in[\underline{p}_{\hat\theta},\bar p_{\hat\theta}]}(g^{-1})'(p)\right)a_n/\gamma_n\right).$$

As $\inf_{p\in[\underline{p}_{\hat\theta},\bar p_{\hat\theta}]}(g^{-1})'(p) > 0$, this is $o_P(1)$ for $a_n/\gamma_n \to \infty$ and Gaussian K. The same applies to the second term of (97), so (95) is $o_P(1)$. Last, we show that (96) converges to $-\lambda_n'(x_1,g^{-1}(p))$. Again, we can write

$$\lambda_n'(x_1,g^{-1}(p)) = \frac{\lambda_n'(x_1,g^{-1}(p))}{\gamma_n}$$

$$\times\left\{\int_{-\infty}^{g^{-1}(p)-\bar t_{\hat\theta}} K(u/\gamma_n)\mathrm{d}u + \int_{g^{-1}(p)-\bar t_{\hat\theta}}^{p-\underline{t}_{\hat\theta}} K(u/\gamma_n)\mathrm{d}u + \int_{g^{-1}(p)-\underline{t}_{\hat\theta}}^{\infty} K(u/\gamma_n)\mathrm{d}u\right\}.$$

A change of variables in (96) ($u := -\gamma_n v$) and K being symmetric about zero give

$$\sup_{x_1\in\mathcal{X}_1, p\in A_n}\left|\int_{(\underline{t}_{\hat\theta}-g^{-1}(p))/\gamma_n}^{(\bar t_{\hat\theta}-g^{-1}(p))/\gamma_n}\lambda_n'(x_1,\gamma_n v+g^{-1}(p))K(v)\mathrm{d}v-\lambda_n'(x_1,g^{-1}(p))\right|$$
$$\leq \sup_{x_1\in\mathcal{X}_1, p\in A_n}\left|\int_{-\infty}^{g^{-1}(p)-\bar t_{\hat\theta}}\lambda_n'(x_1,g^{-1}(p))\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{98}$$
$$+\sup_{x_1\in\mathcal{X}_1, p\in A_n}\left|\int_{g^{-1}(p)-\underline{t}_{\hat\theta}}^{\infty}\lambda_n'(x_1,g^{-1}(p))\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{99}$$
$$+\sup_{x_1\in\mathcal{X}_1, p\in A_n}\left|\int_{g^{-1}(p)-\bar t_{\hat\theta}}^{g^{-1}(p)-\underline{t}_{\hat\theta}}[\lambda_n'(x_1,g^{-1}(p)-u)-\lambda_n'(x_1,g^{-1}(p))]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{100}$$

Since $\mathcal{X}_1$ is bounded by Assumption 10, $\sup_{x_1\in\mathcal{X}_1,p\in A_n}|\lambda_n'(x_1,g^{-1}(p))|$ is bounded with probability tending to one. Hence, as g^{-1} is increasing, $K\geq 0$ and $\bar t_{\hat\theta}=g^{-1}(\bar p_{\hat\theta})$, (98) is of the order $\int_{-\infty}^{(g^{-1}(\bar p_{\hat\theta}-a_n)-g^{-1}(p))/\gamma_n}K(v)\mathrm{d}v$. Here,

$$(g^{-1}(\bar p_{\hat\theta}-a_n)-g^{-1}(p))/\gamma_n\leq(\inf_{p\in[0,1]}(g^{-1})'(p))(-a_n/\gamma_n)\leq(1/\|g'\|_\infty)(-a_n/\gamma_n)$$

by a mean-value expansion. As the infimum is positive, (98) is $o_P(1)$ for $a_n/\gamma_n\to\infty$, and so is (99) by similar arguments. The term (100) can be treated analogously to (80). First note that taking the supremum over $p\in A_n$ is the same as taking the supremum over $t\in[g^{-1}(\underline{p}_{\hat\theta}+a_n),g^{-1}(\bar p_{\hat\theta}-a_n)]=:T_n$ as g^{-1} is increasing. Take a $\gamma_n\ll\rho_n<(1/\|g'\|_\infty)a_n$ (e.g. $\rho_n:=(1/\|g'\|_\infty)\kappa_2 n^{-(\alpha+\beta)/2}$ for some $0<\kappa_2<\kappa_1$). Then (100) is bounded by

$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\left|\int_{[t-\bar t_{\hat\theta},t-\underline{t}_{\hat\theta}]\cap\{|u|\leq\rho_n\}}[\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{101}$$
$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\left|\int_{[t-\bar t_{\hat\theta},t-\underline{t}_{\hat\theta}]\cap\{|u|>\rho_n\}}[\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t))]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{102}$$

Here, (101) is bounded by

$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\sup_{|u|\leq\rho_n}|\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)|\sup_{x_1\in\mathcal{X}_1,t\in T_n}\int_{-\infty}^{\infty}\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u$$
$$\leq\sup_{x_1\in\mathcal{X}_1,t\in T_n}\sup_{|u|\leq\rho_n}|\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)|.$$

By Assumptions 10 and 11, $\lambda_n'(x_1,\cdot)$ is continuous uniformly in x_1 and is therefore uniformly continuous on the compact set $[g^{-1}(\underline{p}_{\hat\theta}),g^{-1}(\bar p_{\hat\theta})]\supset[g^{-1}(\underline{p}_{\hat\theta}+a_n),g^{-1}(\bar p_{\hat\theta}-a_n)]$. Thus, $|u|\leq\rho_n$ for a small enough ρ_n implies that $t-u,t\in[g^{-1}(\underline{p}_{\hat\theta}),g^{-1}(\bar p_{\hat\theta})]$ for all $t\in T_n$. (Note that ρ_n is small enough if and only if $|u|\leq\rho_n$ implies

$$|u|\leq\min\left\{g^{-1}(\underline{p}_{\hat\theta}+a_n)-g^{-1}(\underline{p}_{\hat\theta}),g^{-1}(\bar p_{\hat\theta})-g^{-1}(\bar p_{\hat\theta}-a_n)\right\}.$$

By a mean-value expansion, the right side is smaller than or equal to $\inf_{p\in[0,1]}(g^{-1})'(p)a_n \leq (1/\|g'\|_\infty)a_n$. But then $\rho_n < (1/\|g'\|_\infty)a_n$ is small enough.) As a consequence,

$$\sup_{x\in\mathcal{X}_1, t\in T_n} \sup_{|u|\leq\rho_n} |\lambda_n'(x_1, t-u) - \lambda_n'(x_1, t)| = o_P(1),$$

hence (101) is $o_P(1)$. As $2\sup_{x\in\mathcal{X}_1, t\in[\underline{t}_{\hat\theta}, \bar{t}_{\hat\theta}]} |\lambda_n(x_1,t)|$ is bounded by probability tending to one as $\hat\theta_2 \xrightarrow{P} \theta_{0,2} \neq 0$ by Assumptions 10 and 11, (102) can be shown to be $o_P(1)$ similarly to (82). Thus, (100) is $o_P(1)$ which shows the desired convergence of (96). Hence, (88) holds for $k = 1$. The case for $k = 2$ follows by writing

$$\mathbb{E}[m(X) \mid \theta^\intercal X = t] = (1/v(\theta,t)) \int_{\mathcal{X}_2} m\left(\frac{t-\theta_2 x_2}{\theta_1}, x_2\right) \Psi\left(\frac{t-\theta_2 x_2}{\theta_1}, x_2\right) \mathrm{d}x_2$$
$$v(\theta,t) \coloneqq \int_{\mathcal{X}_2} \Psi\left(\frac{t-\theta_2 x_2}{\theta_1}, x_2\right) \mathrm{d}x_2$$

instead of (91) and (92). By Assumption 10, these behave equally well. Conclude that the desired convergence of $\hat{\hat{q}}_{\mu,k}$ (Condition (68), part (85)) holds.

Condition (68)*, part* (86)*.* Follows along the same lines as (85) by setting $Y_j \coloneqq 1$ for all $j \in [n]$. Conclude that Condition (68) holds.

Remaining Uniform Consistency Results. Showing

$$\sup_{p\in A_n} |\hat\mu(\hat\theta, p) - \mu(\hat\theta, p)| \xrightarrow{P} 0, \tag{103}$$
$$\sup_{p\in A_n} |\hat\mu_2(\hat\theta, p) - \mu_2(\hat\theta, p)| \xrightarrow{P} 0 \tag{104}$$

completes the proof of Proposition 7. As $h(\hat\theta, \cdot)$ is bounded away from zero with probability tending to one by Assumption 6, (103) is implied by (73) and (74). Likewise, (104) is implied by $\|\hat{q}_{\mu_2} - q_{\mu_2}\|_{A_n} \xrightarrow{P} 0$, which can be shown as (73). ∎

References

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006. ISSN 0012-9682, 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://doi.wiley.com/10.1111/j.1468-0262.2006.00655.x.

Alberto Abadie and Guido W. Imbens. Matching on the Estimated Propensity Score. *Econometrica*, 84(2):781–807, 2016. ISSN 0012-9682. doi: 10.3982/ECTA11293. URL https://www.econometricsociety.org/doi/10.3982/ECTA11293.

Kenneth S. Alexander. Rates of Growth and Sample Moduli for Weighted Empirical Processes Indexed by Sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

N. Balakrishnan and Shanti S. Gupta. 2 Higher Order Moments of Order Statistics from Exponential and Right-Truncated Exponential Distributions and Applications to Life-Testing Problems. In *Order Statistics: Applications*, volume 17 of *Handbook of Statistics*, pages 25–59. Elsevier, 1998. doi: https://doi.org/10.1016/S0169-7161(98)17004-9. URL https://www.sciencedirect.com/science/article/pii/S0169716198170049. ISSN: 0169-7161.

Richard Kwasi Bannor, Gupta Amarnath Krishna Kumar, Helena Oppong-Kyeremeh, and Camillus Abawiera Wongnaa. Adoption and Impact of Modern Rice Varieties on Poverty in Eastern India. *Rice Science*, 27(1):56–66, January 2020. ISSN 16726308. doi: 10.1016/j.rsci.2019.12.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1672630819301106.

Herman J. Bierens. *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press, 1 edition, July 1994. ISBN 978-0-521-56511-0 978-0-521-41900-0 978-0-511-59927-9. doi: 10.1017/CBO9780511599279. URL https://www.cambridge.org/core/product/identifier/9780511599279/type/book.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 3rd ed edition, 1995. ISBN 978-0-471-00710-4.

Elenka Brenna. Should I care for My Mum or for My Kid? Sandwich Generation and Depression Burden in Italy. *Health Policy*, 125(3):415–423, March 2021. ISSN 01688510. doi: 10.1016/j.healthpol.2020.11.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0168851020302979.

Kristen Capogrossi and Wen You. The Influence of School Nutrition Programs on the Weight of Low-Income Children: A Treatment Effect Analysis. *Health Economics*, 26(8):980–1000, 2017.

Youngmin Cho. The Effects of Nonstandard Work Schedules on Workers' Health: A Mediating Role of Work-to-Family Conflict. *International Journal of Social Welfare*, 27(1):74–87, January 2018. ISSN 1369-6866, 1468-2397. doi: 10.1111/ijsw.12269. URL https://onlinelibrary.wiley.com/doi/10.1111/ijsw.12269.

William G. Cochran. Matching in Analytical Studies. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):684–691, June 1953. ISSN 0002-9572. doi: 10.2105/AJPH.43.6_Pt_1.684. URL https://ajph.aphapublications.org/doi/full/10.2105/AJPH.43.6_Pt_1.684.

William G. Cochran and Donald B. Rubin. Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446, 1973. ISSN 0581572X. URL http://www.jstor.org/stable/25049893.

Francisco Cribari-Neto, N. Garcia, and Klaus L. P. Vasconcellos. A Note on Inverse Moments of Binomial Variates. *Brazilian Review of Econometrics*, 20:269–277, 2000.

Rajeev Dehejia and Sadek Wahba. Propensity Score Matching Methods for Non-experimental Causal Studies. Technical Report w6829, National Bureau of Economic Research, Cambridge, MA, December 1998. URL http://www.nber.org/papers/w6829.pdf.

Rajeev H. Dehejia and Sadek Wahba. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/3211745.

P. M. Densen, P. Padget, B. Webster, C. S. Nicol, and C. Rich. Studies in Cardiovascular Syphilis. II. Methodologic Problems in the Evaluation of Therapy. *American Journal of Syphilis, Gonorrhea, and Venereal Diseases*, 36(1):64–76, January 1952. ISSN 0096-6738.

Eva Christine Erhardt. Microfinance Beyond Self-Employment: Evidence for Firms in Bulgaria. *Labour Economics*, 47: 75–95, August 2017. ISSN 09275371. doi: 10.1016/j.labeco.2017.04.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0927537116302561.

Markus Frölich. Matching Estimators and Optimal Bandwidth Choice. *Statistics and Computing*, 15(3):197–215, July 2005. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-005-1309-6. URL http://link.springer.com/10.1007/s11222-005-1309-6.

Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315, March 1998. ISSN 00129682. doi: 10.2307/2998560. URL https://www.jstor.org/stable/2998560?origin=crossref.

James J. Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies*, 65(2):261–294, 1998. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2566973.

Jonas Heese, Mozaffar Khan, and Karthik Ramanna. Is the SEC Captured? Evidence from Comment-Letter Reviews. *Journal of Accounting and Economics*, 64(1):98–122, August 2017. ISSN 01654101. doi: 10.1016/j.jacceco.2017.06.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0165410117300344.

Martin Huber, Michael Lechner, and Andreas Steinmayr. Radius Matching on the Propensity Score with Bias Adjustment: Tuning Parameters and Finite Sample Behaviour. *Empirical Economics*, 49(1):1–31, August 2015a. ISSN 0377-7332, 1435-8921. doi: 10.1007/s00181-014-0847-1. URL http://link.springer.com/10.1007/s00181-014-0847-1.

Martin Huber, Michael Lechner, and Conny Wunsch. Workplace Health Promotion and Labour Market Performance of Employees. *Journal of Health Economics*, 43:170–189, September 2015b. ISSN 01676296. doi: 10.1016/j.jhealeco.2015.07.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629615000776.

Jonathan Izudi, Imelda K. Tamwesigire, and Francis Bajunirwe. Does Completion of Sputum Smear Monitoring Have an Effect on Treatment Success and Cure Rate Among Adult Tuberculosis Patients in Rural Eastern Uganda? A Propensity Score-Matched Analysis. *PLOS ONE*, 14(12):e0226919, December 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0226919. URL https://dx.plos.org/10.1371/journal.pone.0226919.

Shao-Hsun Keng and Sheng-Jang Sheu. The Effect of Stimulants and Their Combined Use with Cigarettes on Mortality: The Case of Betel Quid. *The European Journal of Health Economics*, 14(4):677–695, August 2013. ISSN 1618-7598, 1618-7601. doi: 10.1007/s10198-012-0415-6. URL http://link.springer.com/10.1007/s10198-012-0415-6.

Yuvaraj Krishnamoorthy and Tanveer Rehman. Impact of Antenatal Care Visits on Childhood Immunization: A Propensity Score-Matched Analysis Using Nationally Representative Survey. *Family Practice*, 39(4):603–609, July 2022. ISSN 1460-2229. doi: 10.1093/fampra/cmab124. URL https://academic.oup.com/fampra/article/39/4/603/6375553.

Ying-Ying Lee. Efficient Propensity Score Regression Estimators of Multivalued Treatment Effects for the Treated. *Journal of Econometrics*, 204(2):207–222, June 2018. ISSN 03044076. doi: 10.1016/j.jeconom.2018.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407618300290.

Jerzy Neyman. *On the Applications of the Theory of Probability to Agricultural Experiments*. PhD thesis, University of Warsaw, 1924.

Anouk Patel-Campillo and V.B. Salas García. Breaking the Poverty Cycle? Conditional Cash Transfers and Higher Education Attainment. *International Journal of Educational Development*, 92:102612, July 2022. ISSN 07380593. doi: 10.1016/j.ijedudev.2022.102612. URL https://linkinghub.elsevier.com/retrieve/pii/S0738059322000621.

Herbert Robbins. A Remark on Stirling's Formula. *The American Mathematical Monthly*, 62(1):26, January 1955. ISSN 00029890. doi: 10.2307/2308012. URL http://www.jstor.org/stable/2308012?origin=crossref.

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. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41.

Paul R. Rosenbaum and Donald B. Rubin. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, 39(1):33–38, February 1985. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.1985.10479383. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1985.10479383.

Donald B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663. doi: 10.1037/h0037350. URL http://content.apa.org/journals/edu/66/5/688.

Donald B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511810725.

Luca Salmasi and Luca Pieroni. Immigration Policy and Birth Weight: Positive Externalities in Italian Law. *Journal of Health Economics*, 43:128–139, September 2015. ISSN 01676296. doi: 10.1016/j.jhealeco.2015.06.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629615000752.

E. Schuster and S. Yakowitz. Contributions to the Theory of Nonparametric Regression, with Application to System Identification. *The Annals of Statistics*, 7(1):139–149, 1979. ISSN 00905364. URL http://www.jstor.org/stable/2958838. Publisher: Institute of Mathematical Statistics.

Ramalingam Shanmugam and Barry C. Arnold. Characterizations Based on Conditional Distributions Given the Minimum Value in the Sample. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 50(3):452–459, 1988. ISSN 0581572X. URL http://www.jstor.org/stable/25050713.

Chung-Hua Shen and Yuan Chang. Ambition Versus Conscience, Does Corporate Social Responsibility Pay off? The Application of Matching Methods. *Journal of Business Ethics*, 88(S1):133–153, April 2009. ISSN 0167-4544, 1573-0697. doi: 10.1007/s10551-008-9826-9. URL http://link.springer.com/10.1007/s10551-008-9826-9.

Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Number 59 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2009. ISBN 978-0-89871-684-9. OCLC: ocn421147764.

Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010. doi: 10.1214/09-STS313. URL https://doi.org/10.1214/09-STS313.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 1-107-26372-7.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2547-6 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2. URL http://link.springer.com/10.1007/978-1-4757-2545-2.

Nerina Vecchio, Debbie Davies, and Nicholas Rohde. The Effect of Inadequate Access to Healthcare Services on Emergency Room Visits. A Comparison Between Physical and Mental Health Conditions. *PLOS ONE*, 13(8):1–15, August 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0202559. URL https://dx.plos.org/10.1371/journal.pone.0202559.

Wei Wang, Lili Lu, Mohammedhamid Mohammedosman Kelifa, Yan Yu, Anqi He, Na Cao, Si Zheng, Wenjun Yan, and Yinmei Yang. Mental Health Problems in Chinese Healthcare Workers Exposed to Workplace Violence During the COVID-19 Outbreak: A Cross-Sectional Study Using Propensity Score Matching Analysis. *Risk Management and Healthcare Policy*, Volume 13:2827–2833, December 2020. ISSN 1179-1594. doi: 10.2147/RMHP.S279170.