EFFICIENT ESTIMATION OF TREATMENT EFFECTS UNDER TREATMENT-BASED SAMPLING

KYUNGCHUL SONG AND ZHENGFEI YU

Abstract. Nonrandom sampling schemes are often used in program evaluation settings to improve the quality of inference. When observations are drawn with different frequency based on the treatment status and other demographic characteristics and a reasonable range of population shares is given, average treatment effects and average treatment effect on the treated are functions on this range. This paper establishes semiparametric efficiency bounds for these treatment effects, and proposes their efficient estimators. Then the paper proposes a method to construct an asymptotically valid confidence set by inverting a test that is efficient at each point of population shares. Results from a Monte Carlo simulation study are provided.

Key words: Treatment-based Sampling; Standard Stratified Sampling; Choice-Based Sampling; Semiparametric Efficiency; Average Treatment Effects; Optimal Sampling Designs; Partial Identification

JEL Subject Classification: C12, C14, C52.

1. Introduction

Program evaluation studies often adopt nonrandom sampling to improve the quality of inference. For example, Ashenfelter and Card (1985) analyzed data from the Comprehensive Employment and Training Act (CETA) training program using a sample constructed by combining subsamples of program participants and a sample of nonparticipants drawn from the Current Population Survey (CPS). Also, the studies of Lalonde (1986), Dehejia and Wahba (1998, 1999) and Smith and Todd (2005) investigated the National Supported Work (NSW) training program where the training group consisted of individuals eligible for the program and the comparison sample were drawn from the CPS and the Panel Study of Income...
Dynamics (PSID) surveys. Numerous studies focused on the Job Training Partnership Act (JTPA) training program (e.g. Heckman, Ichimura, Smith and Todd (1998), Heckman, Ichimura and Todd (1997)). The participants in these data sets typically represented about 50% in the study sample in comparison to 3% in the population.

This paper refers to as treatment-based sampling a standard stratified sampling process in which the samples are drawn with different frequency based on the treatment status and other observed strata of certain variables. In treatment-based sampling, the observations are not a random sample from the population. Except for the special case where the observations are drawn solely based on the treatment status and the object of interest is an average treatment effect on the treated, the treatment effects parameters are not identified without knowledge of the population shares. (Population shares here refer to the population proportions of the strata variables.)

As noted by Heckman and Todd (2009), information about the population shares is typically not available in the data set. However, a reasonable range of population shares that are plausible in the specific context of a research may be available. This prior knowledge of population shares can be obtained from aggregate demographic statistics from other published data sets such as PSID or the U.S. Census data. When the information of population shares is provided not as a point but as a range, the inference should be robust to different values of the population shares in the given range.

This paper focuses on weighted average treatment effects and weighted average treatment effects on the treated. When the population shares are given as a range, the treatment effect parameters are functions on the range of the population shares. This paper first establishes the efficiency bound for the functions, and provide efficient estimators. Finally, the paper derives the way to construct a confidence interval for the treatment effect parameter by inverting the functional equality restrictions.

The rationale for nonrandom sampling often stems from the belief that when the participants constitute a small proportion in the population, sampling relatively more from the participants will improve the quality of inference. However, this is not an accurate description because we need to consider also the contribution of the noise in the subsample to the variance of the estimator. This paper makes this point clear by developing an optimal design of treatment-based sampling. See Hahn, Hirano and Karlan (2009) for an optimal design of social experiments in a related context.

Early literature on nonrandom sampling has assumed that the conditional distribution of observations given a stratum belongs to a parametric family. (Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a, 1981b), Imbens (1992), and Imbens and Lancaster (1996).) Wooldridge (1999, 2001) studied M-estimators under nonrandom sampling which do not rely on this assumption.
Closer to this paper, Breslow, McNeney and Wellner (2003) and Tripathi (2011) investigated the problem of efficient estimation under non-random sampling schemes. Tripathi (2011) considered moment-based models under various non-random sampling schemes and proved that the empirical likelihood estimators adapted to an appropriate change of measure achieve efficiency. The stratified sampling scheme studied by Tripathi (2011) is different from this paper’s set-up because the identification of the counterfactual quantities in this paper cannot be formulated as arising from the moment condition of his paper. Neither does this paper’s framework fall into the framework of Breslow, McNeney and Wellner (2003) who considered variable probability sampling which is different from the standard stratified sampling studied here.

In the program evaluations literature, there are surprisingly few researches that deal with inference under treatment-based sampling. Chen, Hong, and Tarozzi (2008) established semiparametric efficiency bounds in a broader context where one has outcome observations with missing values and has auxiliary data that aid in identification. While the general approach of Chen, Hong, and Tarozzi (2008) applies to some stratified sampling schemes, it does not here because the event of missing values involves the treatment status here, failing the unconfoundedness condition assumed in their paper. A paper by Heckman and Todd (2009) offers a nice, simple idea to estimate treatment effect on the treated under treatment-based sampling without assuming knowledge of population shares.

The structure of this paper is as follows. Section two introduces treatment-based sampling data designs. Section three provides the main proposal of the paper. Section four discusses results from Monte Carlo simulation studies. In Section five, we conclude. All the technical results and proofs are relegated to the appendix.

2. Treatment Effects under Treatment-Based Sampling

2.1. Treatment-Based Sampling. Treatment-based sampling proceeds as follows. Let $D$ be a random variable that takes values in $\{0,1\}$, where $D = 1$ means participation in the program and $D = 0$ being left in the control group. Let $X = (V_1, V_2, W)$ be a vector of covariates. The first component $V_1$ is a continuous random vector, and $V_2$ and $W$ are discrete random vectors taking values from finite sets $V_2$ and $W$ respectively. The random vector $W$ represents part of $X$, whose support, along with that of $D$, forms strata for treatment-based sampling. For example, $W$ may be the vector of dummy variables for the service regions in the JTPA job training program. We write $V = (V_1, V_2)$ so that $X = (V, W)$.

Under treatment-based sampling, a random sample of size $N$ for the discrete vector $(D,W)$ is first collected. From each subsample with $(D_i, W_i) = (d, w)$, a random sample $\{Y_i, V_i\}_{i=1}^{n_{d,w}}$ of predetermined size $n_{d,w}$ for a vector $(Y, V)$ is collected, where $Y = \ldots$
$Y_1D + Y_0(1 - D)$ and $Y_1$ denotes the potential outcome of a person treated in the program and $Y_0$ the potential outcome of a person not treated. In this paper, we call this type of sampling *treatment-based sampling* as the strata $\{0, 1\} \times W$ involve the treatment status. When $W_i = 1$ for all $i$, so that the strata are constructed based only on the treatment status, we call this sampling *pure treatment-based sampling*. While the observations in the combined sample $\{(D_i, W_i, Y_i, V_i)\}_{i=1}^n$ are independent across $i$'s, the marginals are not identical. Hence inference based on random sampling can be misleading. Throughout this paper, it is assumed that the econometrician does not have individual observations for $(D_i, W_i)_{i=1}^N$ from the original data set, and does not have knowledge of population shares $p^*_d/w = P\{(D, W) = (d, w)\}$, although we assume that we have knowledge of reasonable ranges for them.

For an illustration of treatment-based sampling, consider a job training program implemented in $K$ different service regions. (In the case of the JTPA job training program, there were 16 service regions.) Let $W = \{1, 2, ..., K\}$, the set of index numbers representing the $K$ service regions, and $W \in W$ the service region index for the worker. Each individual worker has a treatment-region status represented by the pair $(D, W)$. For example a worker with $(D, W) = (0, 3)$ means that the worker is not treated and belongs to Service Region 3. When a service region has very few workers eligible for the program in the population, one may want to sample treated workers with a larger proportion than one represented in the population. The extent of the oversampling may differ across different service regions. Then one combines samples obtained by oversampling or undersampling the observations of $(Y, V)$ from each $(d, w)$-subsample. The resulting total sample is one from treatment-based sampling whose distribution is no longer representative of the population.

First, note that a likelihood for observations generated from standard stratified sampling can be viewed as a conditional likelihood from multinomial sampling given $\{n_{d,w}\}_{(d,w)\in\{0,1\}\times W}$. As pointed out by Imbens and Lancaster (1996) (see also Tripathi (2011)), $(D, W)$ is ancillary in both stratified sampling and multinomial sampling, and hence it suffices for semiparametric efficiency to consider only multinomial sampling with design probabilities, say, $\{q_{d,w}\}_{(d,w)\in\{0,1\}\times W}$. Furthermore, since $\{n_{d,w}\}_{(d,w)\in\{0,1\}\times W}$ is a sufficient statistic for multinomial distributions, we can assume that $\{q_{d,w}\}_{(d,w)\in\{0,1\}\times W}$ is known for the computation of semiparametric efficiency bounds. The design probabilities are often known to the researcher from the descriptions of the sampling process.

Thus we let the observations $\{(Y_i, V_i, D_i, W_i)\}_{i=1}^n$ for $(Y, V, D, W)$ be generated by the multinomial sampling scheme using known design probabilities $\{q_{d,w}\}_{(d,w)\in\{0,1\}\times W}$. In other words, we draw a stratum $(d, w)$ from $\{0, 1\} \times W$ using the multinomial distribution with known probabilities $\{q_{d,w}\}_{(d,w)\in\{0,1\}\times W}$, and then draw $(Y, V)$ from the subsample with $(D, W) = (d, w)$. We repeat the procedure until the total sample size becomes $n$. Unless
$q_{d,w} = p^*_d,\!w$ for all $(d,w) \in \{0,1\} \times W$, the observations $\{(Y_i,V_i,D_i,W_i)\}_{i=1}^n$ are not i.i.d. draws from $P$. The observations $\{(Y_i,V_i,D_i,W_i)\}_{i=1}^n$ are i.i.d., however, under probability $Q$ with density $q_{d,w} f_{Y,V|D,W}(y,v|d,w)$, where $f_{Y,V|D,W}(y,v|d,w)$ is the conditional density of $(Y,V)$ given $(D,W) = (d,w)$ with respect to a $\sigma$-finite measure, say, $\mu$. Therefore, the object of inference is a functional of $P$, while the observations are i.i.d. under $Q$.

The notation of expectation, $\mathbf{E}$, without a subscript, is assumed to be under $P$. Expectation $\mathbf{E}_Q$ denotes expectation under $Q$. Expectation $\mathbf{E}_{d,w}$ denotes the conditional expectation given $(D,W) = (d,w)$, which is identical both under $Q$ and under $P$. In pure treatment-based sampling, we suppress the notation $w$ from subscripts, for example, writing $p^*_d$ instead of $p^*_d,\!w$ and $\mathbf{E}_d$ instead of $\mathbf{E}_{d,w}$. Expectations $\mathbf{E}$ and $\mathbf{E}_d$ depend on the population shares $p^*_d,\!w$. When a potential value $p_{d,w}$ for the population shares is used in these expectations, we write $\mathbf{E}_p$ and $\mathbf{E}_{d,p}$ to make its dependence on $p_{d,w}$ explicit.

### 2.2. Treatment Effects under Treatment-Based Sampling.

The main objects of interest are the weighted average treatment effect, $\tau_{ate}$, and the weighted average treatment effect on the treated, $\tau_{tet}$, defined as follows:

\begin{equation}
\tau_{ate} = \frac{\mathbf{E}[g(X)(Y_1 - Y_0)]}{\mathbf{E}[g(X)]} \quad \text{and} \quad \tau_{tet} = \frac{\mathbf{E}[g(X)(Y_1 - Y_0)|D = 1]}{\mathbf{E}[g(X)|D = 1]},
\end{equation}

where $g$ denotes a nonnegative weighting function. Define propensity scores under $P$ and under $Q$ as follows:

$$p_d(x) = P\{D = d|X = x\},$$
$$q_d(x) = Q\{D = d|X = x\}.$$

Throughout this paper, we adopt the following conditions:

(C1) Unconfounded Treatment Assignment: $(Y_0,Y_1) \perp\!\perp D|X$, i.e., $(Y_0,Y_1)$ is conditionally independent of $D$ given $X$.

(C2) There exists $\epsilon > 0$ such that for all $d \in \{0,1\}$,

$$\epsilon < \inf_{x \in S_g} p_d(x) \quad \text{and} \quad \epsilon < \inf_{x \in S_g} q_d(x),$$

where $S_g$ denotes the support of $g$, i.e., the closure of $\{x \in S_X : g(x) > 0\}$, and $S_X$ is the support of $X$.

Condition C1 requires that $(Y_0,Y_1)$ is conditionally independent of $D$ given $X$. Hence the unconfoundedness condition holds under $P$ whereas the observations are i.i.d. under $Q$. 
Condition C2 assumes that the conditional probabilities \( p_d(x) \) and \( q_d(x) \) are bounded away from zero on the support of \( g \).

Let \( p \) be a vector whose entries are \( p_{d,w} \) with \( (d, w) \) running in the set \( \{0, 1\} \times W \), and let \( A \) be the set of values where the population share \( p \) vector is believed to belong. We assume that \( A \) is contained in the interior of the simplex:

\[
S = \{ p : \Sigma_{(d,w)\in\{0,1\}\times W} p_{d,w} = 1 \text{ and } p_{d,w} > 0 \text{ for all } (d, w) \in \{0, 1\} \times W \},
\]

so that as for the population shares, it is satisfied that for all \( (d, w) \in \{0, 1\} \times W \), \( p_{d,w} \in (0, 1) \).

Let us define for each \( p \in S \),

\[
\tau_{ate}(p) = \frac{\sum_{w \in W} E_{1,w} [g_{1,w}(X; p)Y] - \sum_{w \in W} E_{0,w} [g_{0,w}(X; p)Y]}{\sum_{w \in W} E_{1,w} [g_{1,w}(X; p)] - \sum_{w \in W} E_{0,w} [g_{0,w}(X; p)]}
\]

\[
\tau_{tet}(p) = \frac{\sum_{w \in W} p_{1,w} E_{1,w} [g(X)Y] - \sum_{w \in W} E_{0,w} [g_{0,w}(X; p)p_1(X)Y]}{\sum_{w \in W} p_{1,w} E_{1,w} [g(X)] - \sum_{w \in W} E_{0,w} [g_{0,w}(X; p)p_1(X)]},
\]

where \( g_{d,w}(X; p) = p_{d,w}g(X)/p_d(X; p) \),

\[
p_d(x; p) = \frac{f(v|d, w)p_{d,w}}{f(v|1, w)p_{1,w} + f(v|0, w)p_{0,w}},
\]

and \( f(v|d, w) \) denotes the conditional density function of \( V \) given \( (D, W) = (d, w) \). Note that we have \( p_d(x; p^*) = p_d(x) \). Under C1 and C2, it is not hard to see that we can write

\[
\tau_{ate} = \tau_{ate}(p^*) \text{ and } \tau_{tet} = \tau_{tet}(p^*).
\]

Thus the treatment effect parameters \( \tau_{ate} \) and \( \tau_{tet} \) are identified up to \( p^* \).

Under nonpure treatment-based sampling, \( \tau_{ate} \) and \( \tau_{tet} \) are not identified without knowledge of the population shares \( p^*_{d,w} \), because the marginal distribution of \( X \) is not identified from the data. However, under pure treatment-based sampling or more generally when \( p^*_{1,w} = p^*_w p^*_d \) and \( p^*_w \)'s are identical across \( w \in W \), we can identify \( \tau_{tet} \) without knowledge of \( p^*_{d,w} \). In fact, under (C1), the design of pure treatment-based sampling (i.e., the choice of \( q_d \)) does not play a role in determining the conditional distribution of \( (Y_1, Y_0) \) given \( X \). These facts about identification are summarized in Table 1.

As Wooldridge (2001) observed, the assumption of known population shares \( p^*_{d,w} \) may be motivated by the sampling environment where \( N_{d,w} = \sum_{i=1}^N 1\{(D_i, W_i) = (d, w)\} \) is very large relative to the subsample size \( n_{d,w} \). Such sampling is reasonable when it is much less costly to gather information about \( (D, W) \) than the outcome \( Y \) or full covariates \( X \). In this case, a proper large sample theory would be one with \( n_{d,w}/N_{d,w} \to_P 0 \). At the level of treatment-based samples, the asymptotic theory is tantamount to that from assuming knowledge of \( p^*_{d,w} \). However, as pointed out by Heckman and Todd (2009), the population shares \( N_{d,w}/N \) are not often available in practice.
Table 1. Identification of Treatment Effects Parameters

<table>
<thead>
<tr>
<th>$p_{d,w}'s$</th>
<th>$p_w'$</th>
<th>$\tau_{ate}$</th>
<th>$\tau_{tet}$ (non-PTS)</th>
<th>$\tau_{tet}$ (PTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>Identified</td>
<td>Identified</td>
<td>Identified</td>
<td>Identified</td>
</tr>
<tr>
<td>Unknown</td>
<td>Known</td>
<td>Non-identified</td>
<td>Identified</td>
<td>Identified</td>
</tr>
<tr>
<td>Unknown</td>
<td>Identical $p_w'$'s</td>
<td>Non-identified</td>
<td>Identified</td>
<td>Identified</td>
</tr>
<tr>
<td>Nonidentical $p_w'$'s</td>
<td>Non-identified</td>
<td>Non-identified</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Notes: PTS stands for pure treatment-based sampling and non-PTS for nonpure treatment-based sampling. In the case of non-PTS when $p_w'$'s are identical across $w \in \mathcal{W}$, the identification of $\tau_{tet}$ requires the additional condition that $p_{1,w}' = p_1 p_w'$.

3. Efficient Estimation of Treatment Effects

3.1. Efficiency Bound for the Treatment Effect Parameters. This section focuses on the semiparametric efficiency bounds for $\tau_{ate}(\cdot)$ and $\tau_{tet}(\cdot)$ on $A$. Under the assumptions of this paper, the treatment effect functions $\tau_{ate}(\cdot)$ and $\tau_{tet}(\cdot)$ are $\sqrt{n}$-estimable infinite dimensional elements. To avoid repetitive statements, we write $\tau(\cdot)$ to denote generically either $\tau_{ate}(\cdot)$ or $\tau_{tet}(\cdot)$. For any weakly regular estimator $\hat{\tau}(\cdot)$ of $\tau(\cdot)$, it is satisfied that

$$\sqrt{n}\{\hat{\tau}(\cdot) - \tau(\cdot)\} \rightsquigarrow G(\cdot) + \Delta(\cdot)$$

in $l_\infty(A)$, where $l_\infty(A)$ is the class of bounded real functions on $A$, $\rightsquigarrow$ represents weak convergence in the sense of Hoffman-Jorgensen, $G(\cdot)$ is a mean zero Gaussian process with continuous sample paths, and $\Delta(\cdot)$ is a random element that is independent of $G(\cdot)$. The finite dimensional distributions of $G(\cdot)$ are fully characterized by its covariance kernel denoted by $I^{-1}(p, \tilde{p}) = \mathbb{E}[G(p)G(\tilde{p})]$. We make the following additional assumptions.

1. $\mathbb{E}_d[g(x)] < \infty$ and $\mathbb{E}_{d,p}[g(X)] > 0$ for $(d, p) \in \{0, 1\} \times A$.
2. $\mathbb{E}_d[Y_d^2] < \infty$ for $(d, p) \in \{0, 1\} \times A$.
3. $A$ is compact.

Let us define $\beta_d(X) \equiv \mathbb{E}[Y_d|X]$, $\tau(X) \equiv \mathbb{E}[Y_1 - Y_0|X]$, and

$$e_d(p) \equiv \frac{g(X)}{(\mathbb{E}_p[g(X)])} p_d(X)(Y_d - \beta_d(X)).$$

Also, let for $d \in \{0, 1\}$, $R_{d,ate}(p)(v, w) \equiv t_{ate,p}(v, w) - \mathbb{E}_{d,w}[t_{ate,p}(X)]$, where

$$t_{ate,p}(X) \equiv \frac{g(X)(\tau(X) - \tau_{ate}(p))}{\mathbb{E}_p[g(X)]}.$$
Similarly, let $R_{1,\text{tet}}(p)(v, w) \equiv t_{\text{tet},p}(v, w) - \mathbb{E}_{1,w}[t_{\text{tet},p}(X)]$, where
\[
t_{\text{tet},p}(X) \equiv \frac{g(X)\{\tau(X) - \tau_{\text{tet}}(p)\}}{\mathbb{E}_{1,p}[g(X)]}.
\]
We simply write $R_{\text{d,ate}}(p) = R_{d,\text{ate}}(p)(X)$ and $R_{1,\text{tet}}(p) = R_{1,\text{tet}}(p)(X)$ below. The following theorem establishes the semiparametric efficiency bound for the functions $\tau_{\text{ate}}(\cdot)$ and $\tau_{\text{tet}}(\cdot)$.

**Theorem 1:** Suppose that Conditions C1-C5 hold. Then the following holds.

(i) The inverse information covariance kernels for $\tau_{\text{ate}}(\cdot)$, $I_{\text{ate}}^{-1}(\cdot, \cdot) : A \times A \rightarrow \mathbb{R}$, is equal to
\[
I_{\text{ate}}^{-1}(p, \bar{p}) = \sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w} \tilde{p}_{1,w}}{q_{1,w}} \mathbb{E}_{1,w} \left[ e_1(p)e_1(\bar{p}) + R_{\text{ate}}(p)R_{\text{ate}}(\bar{p}) \right] \right\}
+ \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w} \tilde{p}_{0,w}}{q_{0,w}} \mathbb{E}_{0,w} \left[ e_0(p)e_0(\bar{p}) + R_{\text{ate}}(p)R_{\text{ate}}(\bar{p}) \right] \right\}.
\]

(ii) The inverse information covariance kernels for $\tau_{\text{tet}}(\cdot)$, $I_{\text{tet}}^{-1}(\cdot, \cdot) : A \times A \rightarrow \mathbb{R}$, is equal to
\[
I_{\text{tet}}^{-1}(p, \bar{p}) = \frac{1}{p_1 \bar{p}_1} \sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w} \tilde{p}_{1,w}}{q_{1,w}} \mathbb{E}_{1,w} \left[ \tilde{e}_{11}(p)\tilde{e}_{11}(\bar{p}) + R_{\text{tet}}(p)R_{\text{tet}}(\bar{p}) \right] \right\}
+ \frac{1}{p_1 \bar{p}_1} \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w} \tilde{p}_{0,w}}{q_{0,w}} \mathbb{E}_{0,w} \left[ \tilde{e}_{0,1}(p)\tilde{e}_{0,1}(\bar{p}) \right] \right\},
\]
where $\tilde{e}_{s,d} \equiv g(X)p_d(X)(Y_d - \beta_d(X))/(\mathbb{E}_{1,p}[g(X)]p_s(X))$.

From Theorem 1, it follows that $\sigma_{\text{ate}}^2 = I_{\text{ate}}^{-1}(p^*, p^*)$ is the inverse of the semiparametric efficiency bound for $\tau_{\text{ate}}$ when we know the true population share $p^*$. In particular, when $\mathcal{W}$ is a singleton and the sampling is a random sampling, it follows that
\[
\sigma_{\text{ate}}^2 \equiv \frac{1}{(\mathbb{E}[g(X)])^2} \mathbb{E} \left[ g(X)^2 \left\{ \frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} \right\} + \sum_{d \in \{0,1\}} R_{d,\text{ate}}(p^*)p_d(X) \right],
\]
where $\sigma_d^2(X) = \text{Var}(Y_1|X)$ for $d \in \{0,1\}$. Note that the variance bound $\sigma_{\text{ate}}^2$ is smaller than that of Hirano, Imbens and Ridder (2003) where their variance bound is equal to
\[
\frac{1}{(\mathbb{E}[g(X)])^2} \mathbb{E} \left[ g(X)^2 \left\{ \frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} \right\} + g^2(X)(\tau(X) - \tau_{\text{ate}})^2 \right].
\]
Both the variance bounds are identical if and only if $\mathbb{E}_d[g(X)\{\tau(X) - \tau_{\text{ate}}\}] = 0$ for all $d \in \{0,1\}$. Hence knowledge of $p^*_d$ is not ancillary for $\tau_{\text{ate}}$.

Similarly, $\sigma_{\text{tet}}^2 = I_{\text{tet}}^{-1}(p^*, p^*)$ is the inverse of the semiparametric efficiency bound for $\tau_{\text{tet}}$ when we know the population share $p^*$. Under random sampling (i.e., $p_{d,w}^* = q_{d,w}$) with $g(x) = 1$, $\sigma_{\text{tet}}^2$ is smaller than the variance bound in Hahn (1998) that does not assume knowledge of $p_{d,w}^*$. Therefore, the population shares are not ancillary in general. However,
the situation becomes different when the sampling is pure treatment-based sampling. In this case, the population shares $p_d$ are ancillary. Indeed, in pure treatment-based sampling with $p_d = q_d$ with $g(x) = 1$, $\sigma^2_{\text{ate}}$ is reduced to
\[
E \left[ \frac{p_1(X)\sigma^2_0(X)}{p^*_1(X)} + \frac{\sigma^2_0(X)p^2_1(X)}{p_0(X)p^*_1(X)} \right] + \{\tau(X) - \tau_{\text{ate}}\}^2 p_1(X),
\]
which is identical to the variance bound of Hahn (1998) for $\tau_{\text{ate}}$ without knowledge of $p^*_d$. Therefore, $\sigma^2_{\text{ate}}$ can be viewed as a generalization of the variance bound of Hahn (1998) to pure treatment-based sampling.

Suppose that we know the population share $p^*_{d,w}$. Then one might seek for an optimal choice of the design probabilities $q_{d,w}$ that minimize the inverse of the semiparametric efficiency bounds $\sigma^2_{\text{ate}}$ and $\sigma^2_{\text{tet}}$. The answer is given in the following corollary.

**Corollary 1**: Suppose that Conditions C1-C5 hold. Then the optimal choice of $q_{d,w}$ is given as follows:

\[
q^*_{d,w} = \frac{\sqrt{J^*_{d,w}}}{\sum_{(d,w)\in\{0,1\}\times W} \sqrt{J^*_d}} \quad \text{(for $\tau_{\text{ate}}$)} \quad \text{and} \quad \tilde{q}^*_{d,w} = \frac{\sqrt{\tilde{J}_{d,w}}}{\sum_{(d,w)\in\{0,1\}\times W} \sqrt{\tilde{J}_{d,w}}} \quad \text{(for $\tau_{\text{tet}}$)},
\]

where

\[
J^*_{d,w} = \frac{p^*_{d, w}^2}{(E[g(X)])^2} E_{d,w} \left[ g(X)^2 \frac{\sigma^2_0(X)}{p^*_0(X)} + R^2_{d, \text{ate}}(p^*) \right] \quad \text{and} \quad \tilde{J}_{d,w} = p^*_{d, w}^2 E_{d,w} \left[ \frac{d}{p^*_{1}} \left\{ \sigma^2_1(X) + R^2_{1, \text{tet}}(p^*) \right\} + \frac{1 - d}{p^*_{1}^2} \frac{\sigma^2_0(X)p^2_1(X)}{p^*_0(X)} \right].
\]

To appreciate the result of Corollary 1, let us focus on $\tau_{\text{ate}}$. (The remarks below apply to $\tau_{\text{tet}}$ similarly.) The optimal design for $\tau_{\text{ate}}$ suggests that we sample from the $(d, w)$-subsample precisely according to the “noise” proportion $\sqrt{J^*_{d,w}}$ of the subsample $(d, w)$ in $\sum_{(d,w)\in\{0,1\}\times W} \sqrt{J^*_{d,w}}$. In other words, we sample more from a subsample that induces more sampling variability to the efficient estimator. When we have some pilot sample obtained from a two-stage sampling scheme or other data sources that can be used to draw information about $J^*_{d,w}$, the result here may serve as a guide for optimally choosing the size of the sampling fractions $q_{d,w}$.

Using $q^*_{d,w}$ yields the minimum semiparametric efficiency bound as

\[
(3.1) \quad \left\{ \sum_{(d,w)\in\{0,1\}\times W} \sqrt{J^*_{d,w}} \right\}^2.
\]
The variance in (3.1) is the minimum variance bound over all the choices of the sampling probabilities \( q_{d,w} \). The variance (3.1) can be used to compare different choices of additional stratum variables \( W_i \).

In the case of pure treatment-based sampling, we can make precise the condition for treatment-based sampling to lead to inference of better quality than random sampling. Let \( V_{RS} \) be the variance bound for \( \tau_{ate} \) under random sampling, which is equal to \( V_{TS} \) with \( p^*_d = q_d \). Then it is not hard to see that \( V_{RS} \geq V_{TS} \) if and only if

\[
\min \left\{ p^*_1, \frac{J^*_1}{J^*_1 + J^*_0} \right\} \leq q_1 \leq \max \left\{ p^*_1, \frac{J^*_1}{J^*_1 + J^*_0} \right\}.
\]

Therefore, it is not always true that sampling more from a subsample of low population proportion leads to a better result. The improvement hinges on the noise proportion \( J^*_1/(J^*_1 + J^*_0) \) as well. When \( p^*_1 \) happens to coincide with \( J^*_1/(J^*_1 + J^*_0) \), there is no way for treatment-based sampling to improve upon random sampling.

### 3.2. Efficient Estimation of Treatment Effect Process

In this section, we propose two types of estimators for \( \tau_{ate}(p) \) and \( \tau_{tet}(p) \): sample average (SA) estimators \( \tilde{\tau}^{SA}_{ate}(p) \) and \( \tilde{\tau}^{SA}_{tet}(p) \) and fixed integration (FI) estimators \( \tilde{\tau}^{FI}_{ate}(p) \) and \( \tilde{\tau}^{FI}_{tet}(p) \). The SA estimators are computationally less intensive when the dimension of \( X_i \) is relatively large, while the FI estimator exhibits a better finite sample property in simulations.

#### 3.2.1. Sample Average (SA) Estimator

The SA estimators of \( \tau_{ate}(p) \) and \( \tau_{tet}(p) \) are constructed as sample analogue estimators of \( \tau_{ate}(p) \) and \( \tau_{tet}(p) \). Let \( V = (V_1, V_2) \), where \( V_1 \) is continuous and \( V_2 \) is discrete with supports \( V_1 \subset \mathbb{R}^{t_1} \) and \( V_2 \subset \mathbb{R}^{t_2} \) respectively. Also, let

\[
S_{d,w} = \{1 \leq i \leq n : (D_i, W_i) = (d, w)\}.
\]

First, we obtain a propensity score estimator:

\[
\tilde{\pi}_{d,i}(X_i) \equiv \frac{\hat{\lambda}_{d,i}(X_i)}{\hat{\lambda}_{1,i}(X_i) + \hat{\lambda}_{0,i}(X_i)},
\]

where, with \( \hat{L}_{d,w,i} \equiv (\hat{p}_{d,w}/\hat{q}_{d,w})1\{i \in S_{d,w}\} \) and \( \hat{q}_{d,w} \equiv n_{d,w}/n \), we define

\[
\hat{\lambda}_{d,i}(X_i) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \hat{L}_{d,w,j} K_h(V_{ij} - V_{ii}) 1\{V_{ij} = V_{ii}\},
\]

and \( K_h(s_1, \ldots, s_{L_1}) = K(s_1/h, \ldots, s_{L_1}/h)/h^{L_1} \) and \( K(\cdot) \) is a multivariate kernel function. Then we construct the following SA estimator of \( \tau_{ate}(p) \) as follows:

\[
\tilde{\tau}^{SA}_{ate}(p) \equiv \frac{\sum_{w \in W} \sum_{i \in S_{1,w}} \tilde{g}_{1,w,i} Y_i}{\sum_{w \in W} \sum_{i \in S_{1,w}} \tilde{g}_{1,w,i}} - \frac{\sum_{w \in W} \sum_{i \in S_{0,w}} \tilde{g}_{0,w,i} Y_i}{\sum_{w \in W} \sum_{i \in S_{0,w}} \tilde{g}_{0,w,i}},
\]

where \( \tilde{g}_{1,w,i} \) and \( \tilde{g}_{0,w,i} \) are obtained using the estimated propensity scores \( \tilde{\pi}_{d,i}(X_i) \).
where, for a positive sequence $\delta_n \to 0$, with $\tilde{1}_{n,i} = 1 \{ \tilde{\lambda}_{1,i}(X_i) \wedge \tilde{\lambda}_{0,i}(X_i) \geq \delta_n \}$, we define

\begin{equation}
\tilde{g}_{d,w,i} = \frac{p_{d,w} \tilde{1}_{n,i} g(X_i)}{n_{d,w} \tilde{p}_{d,i}(X_i)}.
\end{equation}

Similarly, we construct a sample analogue estimator of $\tau_{tet}(p)$ as follows:

\begin{equation}
\tilde{\tau}_{tet}^{SA}(p) = \frac{\sum_{w \in W_n} \sum_{i \in S_{1,w}} g(X_i) Y_i}{\sum_{w \in W} \sum_{i \in S_{1,w}} g(X_i)} - \frac{\sum_{w \in W} \sum_{i \in S_{0,w}} \tilde{g}_{0,w,i} \tilde{p}_{1,i}(X_i) Y_i}{\sum_{w \in W} \sum_{i \in S_{0,w}} \tilde{g}_{0,w,i} \tilde{p}_{1,i}(X_i)}.
\end{equation}

In the case of pure treatment-based sampling, the knowledge of $p_d$ is ancillary. In this case, the estimator $\tilde{\tau}_{tet}^{SA}$ is reduced to the following form:

\begin{equation}
\tilde{\tau}_{tet}^{SA}(p) = \frac{\sum_{i \in S_1} g(X_i) Y_i}{\sum_{i \in S_1} g(X_i)} - \frac{\sum_{i \in S_0} Y_i \tilde{1}_{0,i} \sum_{j \in S_1} K_{ji} / \sum_{j \in S_0} K_{ji}}{\sum_{i \in S_0} \tilde{1}_{0,i} \sum_{j \in S_1} K_{ji} / \sum_{j \in S_0} K_{ji}},
\end{equation}

where $K_{ji} = K_h(V_{1j} - V_{1i}) 1\{V_{2j} = V_{2i}\}$ and $S_d = \{ 1 \leq i \leq n : D_i = d \}$. This estimator does not involve the population shares $p_d$.

3.2.2. Fixed Integration (FI) Estimator. To introduce the FI estimators, we rewrite the expression (2.2) in an integral form. Define $f(v, w) = f(v_1|v_2, w)P\{V_{2i} = (v_2, w)\}$, where $f(v_1|v_2, w)$ is the conditional density of $V_{1i}$ given $(V_{2i}, W_i) = (v_2, w)$. Let

\begin{equation}
f(v, w|1) = \frac{p_1(v, w) f(v, w)}{p_1} \quad \text{or} \quad f(v, w|1) = \frac{p_1 v 1 w f(v|1, w)}{p_1}.
\end{equation}

Then we can rewrite the identification results in (2.2) as follows:

\begin{equation}
\tau_{ate}(p) = \frac{\sum_{w \in W} \int g(v, w) \{ \beta_1(v, w) - \beta_0(v, w) \} f(v, w) dv}{\sum_{w \in W} \int g(v, w) f(v, w) dv} \quad \text{and}
\end{equation}

\begin{equation}
\tau_{tet}(p) = \frac{\sum_{w \in W} \int g(v, w) \{ \beta_1(v, w) - \beta_0(v, w) \} f(v, w|1) dv}{\sum_{w \in W} \int g(v, w) f(v, w|1) dv},
\end{equation}

where

\begin{equation}
\beta_d(v, w) = \mathbb{E}[Y_d|(V_i, W_i) = (v, w)] = \frac{p_{d,w} \mathbb{E}_{d,w} [Y|V_i = v]}{p_d(v, w)}.
\end{equation}

We construct an estimator of $f(v, w)$ as follows:

\begin{equation}
\hat{f}(v, w) \equiv \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} K_{h,i}(v) + \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} K_{h,i}(v),
\end{equation}

where $K_{h,i}(v) = K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}$. Define an estimator of $p_d(x)$ as

\begin{equation}
\hat{p}_d(x) \equiv \frac{\tilde{\lambda}_d(x)}{\lambda_1(x) + \tilde{\lambda}_0(x)} = \frac{\tilde{\lambda}_d(x)}{f(v, w)},
\end{equation}

\begin{equation}
\tilde{g}_{d,w,i} = \frac{p_{d,w} \tilde{1}_{n,i} g(X_i)}{n_{d,w} \tilde{p}_{d,i}(X_i)}.
\end{equation}

Similarly, we construct a sample analogue estimator of $\tau_{tet}(p)$ as follows:
where we define
\[
\tilde{\lambda}_d(x) = \frac{1}{n} \sum_{i=1}^{n} \tilde{L}_{d,w,i} K_{h,i}(v),
\]
Also define \( \mu_d(v,w) \equiv \beta_d(v,w) f(v,w) \). We construct the an estimator of \( \mu_d(v,w) \) :
\[
\tilde{\mu}_d(v,w) \equiv \frac{1}{\tilde{p}_d(v,w) n_d,w} \sum_{i \in S_{d,w}} Y_i K_{h,i}(v).
\]
Confining the summation to the indices in \( S_{d,w} \) and multiplying \( p_{d,w}/n_{d,w} \) has an effect of modifying the sample average under \( Q \) into that under \( P \). The estimators \( \tilde{\mu}_d(v,w) \) and \( \hat{\mu}_d(v,w) \) are consistent estimators of \( p_d(v,w) \) and \( \mu_d(v,w) \) under regularity conditions.

From now on, we assume that \( V_i \) is a vector of continuous random variables. The only change that is needed when \( V_i \) contains a discrete component \( V_2 \) is that we regard integration for the discrete component as summation over \( v_2 \in V_2 \).

We define the FI estimator of \( \tau_{ate}(p) \) as follows:
\[
\tilde{\tau}_{FI}^{ate}(p) = \frac{\sum_{w \in W} \int g(v,w) \{ \tilde{\mu}_1(v,w) - \tilde{\mu}_0(v,w) \} dv}{\sum_{w \in W} \int g(v,w) f(v,w) dv}.
\]
In practice, the integral can be computed using a usual numerical integration method.

Let us turn to the FI estimator of \( \tau_{tet}(p) \). Define for \( d,s \in \{0,1\} \),
\[
\tilde{\mu}_{d,s}(v,w) = \frac{1}{p_s \tilde{p}_d(v,w) n_d,w} \sum_{i \in S_{d,w}} Y_i K_{h,i}(v),
\]
which is a consistent estimator of \( \beta_{d}(v,w) f(v,w|s) \). Here \( f(v|s,w) \) denotes the conditional density of \( V_i \) given \( (D_i,W_i) = (s,w) \). Then we construct the FI estimator of \( \tau_{tet}(p) \):
\[
\tilde{\tau}_{FI}^{tet}(p) = \frac{\sum_{w \in W} \int g(v,w) \{ \tilde{\mu}_{1,1}(v,w) - \tilde{\mu}_{0,1}(v,w) \} dv}{\sum_{w \in W} \int g(v,w) f(v,w|1) dv},
\]
where \( \tilde{f}(v,w|d) \equiv \frac{p_{d,w}}{n_{d,w} p_d} \sum_{i \in S_{d,w}} K_{h,i}(v) \).

### 3.3. Confidence Intervals.

Construction of a uniform confidence interval can be proceeded as follows. Let \( \tilde{\tau}_{ate}(p) \) be either \( \tilde{\tau}_{ate}^{SA}(p) \) or \( \tilde{\tau}_{ate}^{FI}(p) \), and similarly for \( \tilde{\tau}_{tet}(p) \). First, for each \((t,p) \in \mathbb{R} \times A\), define
\[
T_{ate}(t) = \sqrt{n} \inf_{p \in A} \frac{|\tilde{\tau}_{ate}(p) - t|}{\tilde{\sigma}_{ate}(p)} \quad \text{and} \quad T_{tet}(t) = \sqrt{n} \inf_{p \in A} \frac{|\tilde{\tau}_{tet}(p) - t|}{\tilde{\sigma}_{tet}(p)},
\]
where \( \hat{\sigma}_{\text{ate}}(p) \) and \( \hat{\sigma}_{\text{tet}}(p) \) are consistent estimators of \( \sigma_{\text{ate}}(p) \) and \( \sigma_{\text{tet}}(p) \) such that
\[
\sqrt{n}\left(\hat{\tau}_{\text{ate}}(p) - \tau_{\text{ate}}(p)\right) \overset{d}{\to} N(0, \sigma^2_{\text{ate}}(p)) \quad \text{and} \\
\sqrt{n}\left(\hat{\tau}_{\text{tet}}(p) - \tau_{\text{tet}}(p)\right) \overset{d}{\to} N(0, \sigma^2_{\text{tet}}(p)).
\]

We construct confidence sets for \( \tau_{\text{ate}} \) and \( \tau_{\text{tet}} \):
\[
C_{\text{ate}} = \left\{ t \in \mathbb{R} : T_{\text{ate}}(t) \leq c_{1-\alpha/2} \right\} \quad \text{and} \\
C_{\text{tet}} = \left\{ t \in \mathbb{R} : T_{\text{tet}}(t) \leq c_{1-\alpha/2} \right\},
\]
where \( c_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2) \) and \( \Phi \) is the CDF of the standard normal distribution. Then it is shown in the paper that the confidence sets are asymptotically valid, i.e.,
\[
\liminf_{n \to \infty} P\{\tau_{\text{ate}} \in C_{\text{ate}}\} \geq 1 - \alpha, \quad \text{and} \\
\liminf_{n \to \infty} P\{\tau_{\text{tet}} \in C_{\text{tet}}\} \geq 1 - \alpha.
\]

To see how this choice of confidence sets works, first note that for any \( p_0 \in A \) such that \( \tau_{\text{ate}}(p_0) = \tau_{\text{ate}} \),
\[
\sqrt{n} \inf_{p \in A} \frac{|\hat{\tau}_{\text{ate}}(p) - \tau_{\text{ate}}(p_0)|}{\hat{\sigma}_{\text{ate}}(p)} \leq \sqrt{n} \left| \frac{|\hat{\tau}_{\text{ate}}(p_0) - \tau_{\text{ate}}(p_0)|}{\hat{\sigma}_{\text{ate}}(p_0)} \right| \overset{d}{\to} |Z|,
\]
where \( Z \) is a random variable with a standard normal distribution. Therefore, reading the critical value from the distribution of \( |Z| \) maintains the validity of the confidence sets. When \( p_0 \in A \) such that \( \tau_{\text{ate}}(p_0) = \tau_{\text{ate}} \) is unique, the confidence sets attain exactly the coverage probability in large samples. This arises when the sampling is pure treatment-based sampling.

We may consider an alternative critical value that leads to less conservative confidence sets. For example, this can be achieved by considering a tuning parameter that focuses on a subset of \( A \) on which \( \tilde{\tau}_{\text{ate}}(p) \) is close to \( t \), when we construct a critical value corresponding to \( T_{\text{ate}}(t) \). This paper does not pursue this possibility for two reasons. First, the alternative critical value introduces an additional choice of tuning parameter that can be cumbersome for practitioners. Second, the computation of critical values becomes substantially more complex than the current choice of \( c_{1-\alpha/2} \).

### 3.4. Asymptotic Properties of the Efficient Estimators

In this section, we establish that the semiparametric efficiency of the estimated functions \( \tilde{\tau}_{\text{ate}}(\cdot) \) and \( \tilde{\tau}_{\text{tet}}(\cdot) \), where \( \tilde{\tau}_{\text{ate}}(\cdot) \) is either \( \tilde{\tau}_{\text{ate}}^{SA}(\cdot) \) or \( \tilde{\tau}_{\text{ate}}^{FI}(\cdot) \), and \( \tilde{\tau}_{\text{ate}}(\cdot) \) is either \( \tilde{\tau}_{\text{tet}}^{SA}(\cdot) \) or \( \tilde{\tau}_{\text{tet}}^{FI}(\cdot) \). For simplicity of the exposition, we assume from now on that \( V_i \) is a continuous random vector. It is not hard to extend the result to the case where \( V_i \) contains a discrete component. Let \( \mathcal{V} \) be the support of \( V_i \) and define \( \mathcal{V}(w) \equiv \{ v \in \mathcal{V} : g(v, w) > 0 \} \).
Assumption 1: For each \((d, w) \in \{0, 1\} \times \mathcal{W}\), the following conditions hold.

(i) \(f(v|d, w), \beta_d(v, w)\), and \(g(v, w)\) are bounded and \(L_1 + 1\) times continuously differentiable in \(v\) with bounded derivatives on \(\mathbb{R}^{L_1}\) and uniformly continuous \((L_1 + 1)\)-th derivatives.

(ii) \(\sup_{x \in S_X} \mathbb{E}_{d, w} ||Y'_1||^r + |Y_0| |X = x| < \infty\) and \(\mathbb{E}_{d, w} ||V||^r < \infty\), for some \(r \geq 4\).

(iii) For some \(\varepsilon > 0\), \(\min_{(d, w) \in \{0, 1\} \times \mathcal{W}} \inf_{v \in \nu(w)} f(v|d, w) > \varepsilon\).

(iv) For some \(\delta \geq 4\), it is satisfied that for all \((d_1, d_2, w) \in \{0, 1\}^2 \times \mathcal{W}\),

\[
\sup_{a \in [0, \delta]} \mathbb{E}_{d_1, w} \left[ \sup_{p \in A} \left( \frac{f(V'_1|d_2, w) p_{d_2, w}}{\sum_{d_3 \in \{0, 1\}} f(V'_1|d_3, w) p_{d_3, w}} \right)^{a} \right] < \infty.
\]

Assumption 1 is the condition of sample overlap needed for the identification of \(\tau_{ate}\) and \(\tau_{tet}\). This is violated when part of \(X\) is only observed among the treated or untreated subsamples. (See Heckman, Ichimura, and Todd (1997) for a discussion on this issue.) See Khan and Tamer (2009) for an analysis of situations where Assumption 1 is violated with \(p_1(x)\) being arbitrarily close to 0 or 1. Assumption 2 requires that \(f(\cdot|d, w)\) is smooth on \(\mathbb{R}^{L_1}\). Assumption 1(i) is introduced to deal with the boundary problem of kernel estimators. In general, the performance of kernel estimators is unstable near the boundary of the support of \(V_i\). In this case, it is reasonable to trim part of the samples such that the realizations of \(V_i\) appear to be "outliers." For example, see Heckman, Ichimura and Todd (1998) for application of such trimming schemes. We view that the trimming is already incorporated in the definition of the treatment effects parameters by choosing appropriate \(g(v, w)\). For example, one may consider \(g(v, w) = 1\{(v, w) \in \mathcal{S}\}\) for some set \(\mathcal{S}\) that is in the interior of the support of \((V_i, W_i)\).

Assumption 2: (i) \(K\) is zero outside an interior of a bounded set, \(L_1 + 1\) times continuously differentiable with bounded derivatives, \(\int K(s) ds = 1\), \(\int s_1^{l_1} \ldots s_{d_1}^{l_{d_1}} K(s) ds = 0\) and \(\int |s_1^{l_1} \ldots s_{d_1}^{l_{d_1}} K(s)| ds < \infty\) for all nonnegative integers \(l_1, \ldots, l_{d_1}\) such that \(l_1 + \ldots + l_{d_1} \leq L_1\), where \(d_1\) denotes the dimension of \(V_i\).

(ii) \(n^{-1/4} h^{-d_1/2} \sqrt{\log n} + (n^{1/2} h^{d/2}) h^{L_1+1} \to 0\), as \(n \to \infty\).

(iii) The trimming sequence \(\delta_n\) in \((3.4)\) satisfies that \(\delta_n \to 0\) and \(\sqrt{n} \delta_n^\gamma \to 0\), for some \(0 < \gamma \leq \bar{a}\), where \(\bar{a}\) is as in Assumption 1(iv).

Assumption 2(i) is a standard assumption for higher order kernels. Assumption 2(ii) includes the condition for bandwidth \(h\). The following theorem establishes the asymptotic distribution of \(\hat{\tau}_{ate}^{SA}\) and \(\hat{\tau}_{ate}^{FI}\).
Theorem 2: Suppose that the Conditions C1-C5 and Assumptions 1-2 hold. Then
\[ \sqrt{n}(\hat{\tau}_{ate}(\cdot) - \tau_{ate}(\cdot)) \overset{a}{\rightarrow} \zeta_{ate}^*(\cdot) \]  
\[ \sqrt{n}(\hat{\tau}_{tet}(\cdot) - \tau_{tet}(\cdot)) \overset{a}{\rightarrow} \zeta_{tet}^*(\cdot), \]

where \( \zeta_{ate}^* \) and \( \zeta_{tet}^* \) are mean zero Gaussian processes with continuous sample paths that have covariance kernels \( I_{ate}^{-1}(\cdot, \cdot) \) and \( I_{tet}^{-1}(\cdot, \cdot) \) given in Theorem 1.

Recall that \( \hat{\tau}_{ate}(\cdot) \) denotes both \( \hat{\tau}_{ate}^{SA}(\cdot) \) and \( \hat{\tau}_{ate}^{FI}(\cdot) \). The result of Theorem 2 immediately yields the following result for the confidence sets \( C_{ate} \) and \( C_{tet} \) for \( \tau_{ate} \) and \( \tau_{tet} \) that were introduced in Section 3.1. Recall that the confidence sets were constructed by inverting the tests \( T_{ate}(t) \leq \Phi^{-1}(1-\alpha/2) \) and \( T_{tet}(t) \leq \Phi^{-1}(1-\alpha/2) \), and that \( \tau_{ate} \) and \( \tau_{tet} \) are the average treatment effects \( \tau_{ate}(p^*) \) and \( \tau_{tet}(p^*) \) with the true (potentially unknown) population share \( p^* \).

Corollary 2: Suppose that the Conditions C1-C5 and Assumptions 1-2 hold. Then
\[ \liminf_{n \to \infty} P\{\tau_{ate} \in C_{ate}\} \geq 1 - \alpha \quad \text{and} \]
\[ \liminf_{n \to \infty} P\{\tau_{tet} \in C_{tet}\} \geq 1 - \alpha. \]

As mentioned previously, one can show that when there is a unique \( p_0 \in A \) such that \( \tau_{ate}(p_0) = \tau_{ate} \), the coverage probabilities in Corollary 2 for \( \tau_{ate} \) are not conservative, i.e.,
\[ \liminf_{n \to \infty} P\{\tau_{ate} \in C_{ate}\} = 1 - \alpha. \]

The same remark also applies to \( \tau_{tet} \).

4. Monte Carlo Simulation Studies

This section presents and discusses the results from Monte Carlo simulation studies. The data generating process is as follows. Let \( \varepsilon_{0,i}, u_{1i}, \) and \( u_{2i} \) be independent random variables drawn from \( N(0,1) \). Then we generate \( V_i = (V_{1i}, V_{2i}) \) in two different ways:

Spec A: \[ \begin{cases} 
V_{1i} = 1\{u_{1i} + \varepsilon_{0,i} \geq 0\} \\
V_{2i} = 1\{u_{2i} + \varepsilon_{0,i} \geq 0\}
\end{cases} \]

Spec B: \[ \begin{cases} 
V_{1i} = u_{1i} + \varepsilon_{0,i} \\
V_{2i} = 1\{u_{2i} + \varepsilon_{0,i} \geq 0\}\]

Hence in Spec A, both \( V_{1i} \) and \( V_{2i} \) are discrete random variables, while in Spec B, only \( V_{2i} \) is discrete. Then we define an index that determines the participation of individuals in the
program:

\[ U_i = a(V_{1i} + V_{2i}) + r_i + b(W_i - 1/2), \]

where \( r_i \sim N(0, 1) \). The variable \( W_i \) constitutes the strata for the treatment-based sampling. We set \( W_i = 0.5 \) in the case of pure treatment-based sampling and \( W_i \in \{0, 1\} \) with \( P\{W_i = 1\} = 0.2 \) in the case of nonpure treatment-based sampling. Both random variables \( r_i \) and \( W_i \) are drawn independently from each other and from \((V_{1i}, V_{2i})\). Then, the participation indicator is simply defined to be

\[ D_i = 1\{U_i > 0.5\}. \]

As for the outcomes, they are specified as follows:

\[
Y_{1i} = e_{0i} + (c_{1i} + 1/2)(V_{1i} + V_{2i})/2 + bW_i + 5(e_{1i} + 1), \quad \text{and} \quad Y_{0i} = e_{0i} + (c_{0i} + 1/2)(V_{1i} + V_{2i})/2 + bW_i,
\]

where \( e_{0i}, e_{1i}, c_{0i}, \) and \( c_{1i} \), are drawn from \( N(0, 1) \). The random variables \( e_{0i}, e_{1i}, c_{0i}, \) and \( c_{1i} \) are independent from each other and independent from \((V_{1i}, V_{2i}, W_i)\). The component in \( 5(e_{1i} + 1) \) represents heterogeneous outcome differences. The Monte Carlo simulation number was 10,000. The sample size considered was 500 and 1000. The design probability \( q_1 \) was set to be 0.5, and the trimming level was 0.001.

Figure 1 illustrates an example of identified sets using a data generating process from Spec A. Each panel in Figure 1 presents an identified set for the pair \((\tau_{ate}(p), \tau_{tet}(p))\) with \( p_1 \) running in a given range. Here \( p_1 \) represents the population proportion of the people eligible for treatment. The data generating process takes \( g(X) = 1 \), and hence each identified set takes the form of a line segment, as expected. When we take \( p_1 \) to run in a longer range, the identified set becomes larger. The treatment effects on the treated remains the same for different values of \( p_1 \) (represented by the horizontal line segments) because the simulation design adopted independence between \( D_i \) and \( W_i \), i.e., \( p_{1, w} = p_1 p_w \).

First, we consider the testing problem for a fixed population share \( p \):

\[
H_0 : \ \tau_{ate}(p) = t,
\]

\[
H_1 : \ \tau_{ate}(p) \neq t,
\]

for some \( t \in \mathbb{R} \). For this we construct a test statistic

\[
T_{ate}(t; p) = \sqrt{n} \left| \frac{\bar{\tau}_{ate}(p) - t}{\hat{\sigma}_{ate}(p)} \right|.
\]

The rejection rule is

\[
\text{Reject } H_0 \text{ if } T_{ate}(p) > c_{1 - \alpha/2},
\]
where $c_{1-\alpha/2} = 1 - \Phi^{-1}(1 - \alpha/2)$, $\Phi$ is the CDF of the standard normal distribution. As for $\tau_{tet}(p)$, we construct a statistic

$$T_{tet}(t; p) = \sqrt{n} \frac{\tau_{tet}(p) - t}{\sigma_{tet}(p)},$$

and the critical value is $c_{1-\alpha/2}$.

In Spec B, we choose the bandwidth by cross-validation using propensity score under $Q$. In particular, let

$$CV(h; w) \equiv \sum_{j=1}^{n} \left[ \left( D_j - \frac{\sum_{i=1, i\neq j}^{n} D_i 1\{W_i = w\} K_{h,j}(V_i)}{\sum_{i=1, i\neq j}^{n} 1\{W_i = w\} K_{h,j}(V_i)} \right)^2 1\{W_j = w\} \right]$$

The bandwidth (depending on $w$ for the nonpure sampling) is chosen by minimizing $CV(h; w)$, that is, $h^*(w) = \arg\min_h CV(h; w)$. 

---

**Figure 1.** Identified set for $(\tau_{ate}, \tau_{tet})$ with different range of $p_1$ with non-pure treatment-based sampling
TABLE 2. Rejection Frequencies for Tests Using $T_{ate}(t; p)$ and $T_{tet}(t; p)$ Under $H_0$ at Nominal Size 5%

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$T_{ate}(p)$</th>
<th>$T_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0541 0.0537 0.0530 0.0526 0.0526</td>
<td>0.0502</td>
</tr>
<tr>
<td>Sampling</td>
<td>$n = 1000$</td>
<td>0.0510 0.0512 0.0515 0.0516 0.0510</td>
<td></td>
</tr>
<tr>
<td>Nonpure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0540 0.0537 0.0538 0.0544 0.0535</td>
<td>0.0535</td>
</tr>
<tr>
<td>Sampling</td>
<td>$n = 1000$</td>
<td>0.0572 0.0573 0.0579 0.0576 0.0514</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Under Spec A, the sample average test and the fixed integration test are the same.

Part of the simulation results are presented in Tables 2-10. Tables 2-4 report results from Spec A and Tables 5-10 those from Spec B. The SA and FI estimators coincide numerically in Spec A but differ in Spec B. In Table 2, finite sample rejection frequencies for both tests of $T_{ate}(t; p)$ and $T_{tet}(t; p)$ are reported with varying population shares $p_1$. The rejection probability for $T_{ate}(t; p)$ is not sensitive to the variation of the population shares $p_1$ for sample size 500 and 1000. Overall, at the sample sizes of 500 and 1000, both the tests perform reasonably well. The performance of $T_{tet}(t; p)$ is similar to that of $T_{ate}(t; p)$ except that the rejection probability turned out to be almost the same across different $p_1$. (Hence the rejection probabilities for $T_{tet}(t; p)$ in Table 2 are presented in a single column for brevity.)

There is no reason this should be a priori so, because although the independence of $D_i$ and $W_i$ under $P$ renders the estimator $\tilde{\tau}_{tet}(p)$ invariant to the choice of $p_1$, the asymptotic variance $\tilde{\sigma}_{tet}^2(p)$ can still vary with the choice of $p_1$. Nevertheless, the rejection probabilities for $T_{tet}(t; p)$ have turned out to be the same (up to the numerical precision allowed in the simulation) across different population shares $p_1$, perhaps because $\tilde{\sigma}_{tet}^2(p)$ does not change much when we vary $p_1$.

Tables 3-4 present finite sample performances of estimators $\tilde{\tau}_{ate}(p)$ and $\tilde{\tau}_{tet}(p)$ in terms of mean absolute deviation, mean squared error and bias. The mean absolute deviation and the mean squared error for $\tilde{\tau}_{ate}(p)$ are similar across different choices of $p_1$ except the performance seems to be slightly better as we increase $p_1$. While this is not necessarily true for the bias, the magnitude of the bias is overall very small. The finite sample performances of the estimator $\tilde{\tau}_{tet}(p)$ is the same across $p_1$’s because the estimator $\tilde{\tau}_{tet}(p)$ itself is numerically the same. For both estimators $\tilde{\tau}_{ate}(p)$ and $\tilde{\tau}_{tet}(p)$, the increase in the sample size reduces the mean absolute deviation and the mean squared error of the estimators, which is expected as the estimators are consistent.

Tables 5-10 report results based on the data generating process from Spec B, where $V_{1i}$ is continuous and $V_{2i}$ is binary, and the bandwidth is chosen through cross-validation described
Table 3. Finite Sample Performances for Point Estimates $\hat{\tau}_{ate}(p)$ and $\hat{\tau}_{tet}(p)$ (Spec A under Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$\hat{\tau}_{ate}(p)$</th>
<th>$\hat{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>mean absolute deviation</td>
<td>0.3217</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.1625</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>-0.0273</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean absolute deviation</td>
<td>0.2248</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.0798</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

Notes: Under Spec A, the sample average estimator and the fixed integration estimator are numerically the same.

Table 4. Finite Sample Performances for Point Estimates $\hat{\tau}_{ate}(p)$ and $\hat{\tau}_{tet}(p)$ (Spec A under Non-Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$\hat{\tau}_{ate}(p)$</th>
<th>$\hat{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>mean absolute deviation</td>
<td>0.5041</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.4017</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>0.0463</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean absolute deviation</td>
<td>0.2726</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.1169</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>-0.0776</td>
</tr>
</tbody>
</table>

Notes: Under Spec A, the sample average estimator and the fixed integration estimator are numerically the same.

in (4.1). Tables 5-6 present the rejection frequencies under $H_0$ at nominal size 5%. (Table 5 focuses on the sample average estimator and Table 6 the fixed integration estimator.) The empirical size properties are quite stable across different choices of population shares $p_1$. The rejection frequencies for the test of the average treatment effect are more stable across different choices of $p_1$, and become closer to the nominal size, when the sample size is increased from 500 to 1000. As for the average treatment effect on the treated, the rejection frequencies do not vary much across different $p_1$’s, almost within 0.0005. For brevity, the rejection frequencies for $\hat{\tau}_{tet}(p)$ are reported in a single column, by taking the average of the four numbers corresponding to different $p_1$’s. All the above properties are shared by both the sample average estimator and the fixed integration estimator.
Table 5. Rejection Frequencies for Tests $T^{SA}_{ate}(t; p)$ and $T^{SA}_{tet}(t; p)$ under $H_0$ at
5% with Spec B.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$T_{ate}(p)$</th>
<th>$T_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Pure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0865</td>
<td>0.0852</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
<td>0.0703</td>
<td>0.0701</td>
</tr>
<tr>
<td>Nonpure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0874</td>
<td>0.0866</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
<td>0.0713</td>
<td>0.0709</td>
</tr>
</tbody>
</table>

Notes: Bandwidths were chosen by the cross-validation based on the propensity scores.

Table 6. Rejection Frequencies for Tests $T^{FI}_{ate}(t; p)$ and $T^{FI}_{tet}(t; p)$ under $H_0$ at
5% with Spec B.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$T_{ate}(p)$</th>
<th>$T_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Pure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0679</td>
<td>0.0681</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
<td>0.0571</td>
<td>0.0574</td>
</tr>
<tr>
<td>Nonpure Treatment-Based</td>
<td>$n = 500$</td>
<td>0.0691</td>
<td>0.0699</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
<td>0.0530</td>
<td>0.0524</td>
</tr>
</tbody>
</table>

Notes: Bandwidths were chosen by the cross-validation based on the propensity scores.

Table 7. Finite Sample Performances for Point Estimates $\tilde{\tau}^{SA}_{ate}(p)$ and $\tilde{\tau}^{SA}_{tet}(p)$
(Spec B under Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$\tilde{\tau}_{ate}(p)$</th>
<th>$\tilde{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>$n = 500$ mean absolute deviation</td>
<td></td>
<td>0.4335</td>
<td>0.4231</td>
</tr>
<tr>
<td>mean squared error</td>
<td></td>
<td>0.2986</td>
<td>0.2843</td>
</tr>
<tr>
<td>bias</td>
<td></td>
<td>0.0066</td>
<td>0.0040</td>
</tr>
<tr>
<td>$n = 1000$ mean absolute deviation</td>
<td></td>
<td>0.3055</td>
<td>0.2985</td>
</tr>
<tr>
<td>mean squared error</td>
<td></td>
<td>0.1478</td>
<td>0.1409</td>
</tr>
<tr>
<td>bias</td>
<td></td>
<td>-0.0201</td>
<td>-0.0227</td>
</tr>
</tbody>
</table>

Finally, Tables 7-10 report results that show the mean absolute deviations of the point
estimators $\tilde{\tau}_{ate}(p)$ and $\tilde{\tau}_{tet}(p)$. (Tables 7-8 focus on the sample average estimator and Tables
9–10 the fixed integration estimator.) Their performance is stable over different choices of
Table 8. Finite Sample Performances for Point Estimates $\hat{\tau}_{ate}(p)$ and $\hat{\tau}_{tet}(p)$ (Spec B under Non-Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$\hat{\tau}_{ate}(p)$</th>
<th>$\hat{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01 0.05 0.10 0.20</td>
<td></td>
</tr>
<tr>
<td>$n = 500$</td>
<td>mean absolute deviation</td>
<td>0.5068 0.4951 0.4810 0.4544 0.3546</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.4087 0.3894 0.3668 0.3262 0.1968</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>−0.0396 −0.0440 −0.0489 −0.0580 −0.0950</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean absolute deviation</td>
<td>0.3561 0.3478 0.3377 0.3187 0.2465</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.2017 0.1922 0.1811 0.1610 0.0955</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>−0.0360 −0.0380 −0.0403 −0.0443 −0.0468</td>
</tr>
</tbody>
</table>

Table 9. Finite Sample Performances for Point Estimates $\hat{\tau}_{ate}(p)$ and $\hat{\tau}_{tet}(p)$ (Spec B under Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$\hat{\tau}_{ate}(p)$</th>
<th>$\hat{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01 0.05 0.10 0.20</td>
<td></td>
</tr>
<tr>
<td>$n = 500$</td>
<td>mean absolute deviation</td>
<td>0.4028 0.3946 0.3846 0.3655 0.2867</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.2585 0.2479 0.2352 0.2121 0.1299</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>−0.0067 −0.0076 −0.0086 −0.0108 −0.0278</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean absolute deviation</td>
<td>0.2881 0.2823 0.2752 0.2618 0.2090</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.1314 0.1260 0.1196 0.1079 0.0686</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>−0.0302 −0.0316 −0.0334 −0.0370 −0.0657</td>
</tr>
</tbody>
</table>

$p_1$’s although it slightly improves when we increase $p_1$. Increasing the sample size from 500 to 1000 makes the performance conspicuously better as expected.

The summary of the findings from the simulation study is as follows. First, the rejection frequencies of the tests are reasonably stable over different choices of $p_1$. Second, the finite sample performances of the estimator $\hat{\tau}_{ate}(p)$ are also stable over different choices of $p_1$, although the performance slightly improves when we increase $p_1$ toward 0.2. Third, the performances of the estimators of $\tau_{tet}(p)$ across different choices of $p_1$ are relatively much more stable than those of $\tau_{ate}(p)$.

5. Conclusion

This paper has established semiparametric efficiency bounds for certain average treatment effect parameters under treatment-based sampling. This paper also proposes efficient estimators for the parameters. An optimal design of treatment-based sampling is also derived.
Table 10. Finite Sample Performances for Point Estimates $\tilde{\tau}_{ate}(p)$ and $\tilde{\tau}_{tet}(p)$
(Spec B under Non-Pure Treatment-Based Sampling)

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$\tilde{\tau}_{ate}(p)$</th>
<th></th>
<th>$\tilde{\tau}_{tet}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>mean absolute deviation</td>
<td>0.4700</td>
<td>0.4608</td>
<td>0.4497</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.3508</td>
<td>0.3368</td>
<td>0.3202</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>-0.0615</td>
<td>-0.0629</td>
<td>-0.0646</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean absolute deviation</td>
<td>0.3326</td>
<td>0.3259</td>
<td>0.3177</td>
</tr>
<tr>
<td></td>
<td>mean squared error</td>
<td>0.1750</td>
<td>0.1678</td>
<td>0.1593</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>-0.0497</td>
<td>-0.0497</td>
<td>-0.0497</td>
</tr>
</tbody>
</table>

The theory of optimal design illuminates the role of treatment-based sampling in improving the quality of inference.

6. Appendix

6.1. Appendix A: Estimators of Variances. As we saw in Theorem 2, both SA and FI estimators have the same variance, which reaches the efficiency bounds derived in Theorem 1. Therefore, the estimation of variances of two types of estimators can be treated in the same way, except for the estimator of $\tau_{ate}(p)$ ($\tau_{tet}(p)$) appearing in the variance formula. Let $\tilde{\tau}_{ate}(p)$ and $\tilde{\tau}_{tet}(p)$ be either sample average estimators or fixed integration estimators.

Consistent estimation of $\sigma^2_{ate}(p)$ can be proceeded as follows. First, let

$$\check{\beta}_{d,p}(X_i) = \frac{\check{\mu}_d(X_i)}{\check{f}(X_i)} \quad \text{and} \quad \check{e}_{d,i}(p) = \frac{g(X_i)(Y_i - \check{\beta}_d(X_i))}{\check{g} \cdot \check{p}_d(X_i)},\ d \in \{0, 1\},$$

where $\check{f}(X_i) \equiv \lambda_0(X_i) + \lambda_1(X_i)$,

$$\check{\mu}_d(X_i) \equiv \frac{1}{\check{p}_d(X_i)} \frac{p_{d,w}}{n_{d,w}} \sum_{i \in S_{1,w}} Y_i K_{h,i}(X_i), \text{ and}$$

$$\check{g} \equiv \sum_{w \in W} \left\{ \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(X_i) + \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(X_i) \right\}.$$

We also define

$$\check{\zeta}_{d,ate,i}(p) = \frac{1}{\check{g}} g(X_i) \left[ \check{\mu}_1(X_i) - \check{\mu}_0(X_i) - \check{\tau}_{ate}(p) \right]$$

$$- \frac{1}{\check{g} m_{d,w}} \sum_{i \in S_{d,w}} g(X_i) \left[ \check{\mu}_1(X_i) - \check{\mu}_0(X_i) - \check{\tau}_{ate}(p) \right].$$
Then we construct
\[ \tilde{\sigma}^2_{ate}(p) = \sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w}^2}{q_{1,w}n_{1,w}} \sum_{i \in S_{1,w}} \left[ \tilde{\epsilon}_{1,i}^2(p) + \tilde{\zeta}_{1,ate,i}^2(p) \right] \right\} 
+ \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w}^2}{q_{0,w}n_{0,w}} \sum_{i \in S_{0,w}} \left[ \tilde{\epsilon}_{0,i}^2(p) + \tilde{\zeta}_{0,ate,i}^2(p) \right] \right\} . \]

Let us consider the asymptotic variance of \( \tilde{\tau}_{tet}(p) \) (and \( \hat{\tau}_{tet}(p) \)) and its estimator. Similarly as before, one might consider an optimal choice of \( q_{d,w} \) which is developed later.

As for the estimator of \( \sigma^2_{tet}(p) \), we let
\[ \tilde{\zeta}_{1,tet,i}^2(p) = \frac{1}{\tilde{g}_1} g(X_i) \left[ \frac{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}{\hat{f}(X_i)} - \tilde{\tau}_{tet}(p) \right] 
- \frac{1}{\tilde{g}_1 n_{1,w}} \sum_{i \in S_{1,w}} g(X_i) \left[ \frac{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}{\hat{f}(X_i)} - \tilde{\tau}_{tet}(p) \right] , \]
and
\[ \tilde{\epsilon}_{s,d,i}(p) \equiv g(X_i) \tilde{\beta}_d(X_i)(Y_i - \hat{\beta}_d(X_i))/(\tilde{g}_1 \cdot \tilde{p}_s(X_i)) , \]
where \( \tilde{g}_1 \equiv (1/p_1) \sum_{w \in \mathcal{W}} \{(p_{1,w}/n_{1,w}) \sum_{i \in S_{1,w}} g(X_i) \}. \) Then an asymptotic variance estimator is given by:
\[ \tilde{\sigma}^2_{tet}(p) = \frac{1}{p_1^2} \sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w}^2}{q_{1,w}n_{1,w}} \sum_{i \in S_{1,w}} \left[ \tilde{\epsilon}_{1,i}^2(p) + \tilde{\zeta}_{1,ate,i}^2(p) \right] \right\} 
+ \frac{1}{p_1^2} \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w}^2}{q_{0,w}n_{0,w}} \sum_{i \in S_{0,w}} \left[ \tilde{\epsilon}_{0,i}^2(p) + \tilde{\zeta}_{0,ate,i}^2(p) \right] \right\} . \]

6.2. Appendix B: Semiparametric Efficiency Bounds and Proofs. Suppose that \( \mathbf{P} \) is a model (a collection of probability measures \( P \) having a density function with respect to a common \( \sigma \)-finite measure \( \mu \)). After identifying each probability in \( \mathbf{P} \) as the square root of its density, we view \( \mathbf{P} \) as a subset of \( L_2(\mu) \). Let \( \mathcal{C}_b(A) \) be the collection of bounded and continuous real functions defined on \( A \subset \mathbb{R}^{2 \times |\mathcal{W}|} \) and \( \| \cdot \| \) be the supremum norm on \( \mathcal{C}_b(A) \). The following definitions are from Bickel et al.(1993).

**Definition B1 [Curve]:** \( \mathbf{V} \) is a curve in \( L_2(\mu) \) if it can be represented as the image of the open interval \((-1, 1)\) under a continuously Fréchet differentiable map. That is, we can write
\[ \mathbf{V} = \{ v(t) \in L_2(\mu) : |t| < 1 \}, \]
where there exists a $\mathbf{v} \in L_2(\mu)$ such that $\mathbf{v}(t + \Delta) = \mathbf{v}(t) + \Delta \mathbf{v}(t) + o(|\Delta|)$, as $|\Delta| \to 0$, for each $t \in (-1, 1)$.

**Definition B2 [Tangent Set]:** The tangent set at $\mathbf{v}_0 \in \mathbf{P}$, denoted as $\mathbf{v}_0^\circ$, is the union of all $\mathbf{v}$ of curves $\mathbf{V} \subset \mathbf{P}$ passing through $\mathbf{v}_0$, where $\mathbf{v}_0 = \mathbf{v}(0)$. The closed linear span of $\mathbf{v}_0^\circ$ is the tangent space, denoted as $\mathbf{v}_0$.

**Definition B3 [Pathwise Differentiability]:** A parameter $\tau : \mathbf{P} \to C_b(\mathbb{R})$ is pathwise differentiable at $\mathbf{v}_0$ if there exists a bounded linear function $\mathbf{v}_0^\circ(\cdot) : \mathbf{v}_0^\circ \to C_b(\mathbb{R})$ such that for any curve $\mathbf{V} \subset \mathbf{P}$ with tangent $s \in \mathbf{v}_0^\circ$, we have

$$\left\| \frac{\tau(\mathbf{v}(t)) - \tau(\mathbf{v}_0)}{t} - \mathbf{v}_0^\circ(s) \right\| = o(1),$$

as $t \to 0$.

**Proof of Theorem 1:** Let $f(y,v,d,w)$ be the density of $(Y,V,D,W)$ with respect to a $\sigma$-finite measure $\mu$ under $P \in \mathcal{P}$, where $\mathcal{P}$ is the collection of potential distributions for $(Y,V,D,W)$. Let $f(y,v|d,w)$ be the conditional density of $(Y,V)$ given $(D,W) = (d,w)$, and $\mathbf{P}_{d,w}$ denotes the collection of conditional densities $f(\cdot,\cdot|d,w)$ of $(Y,V)$ given $(D,W) = (d,w)$ with $P$ running in $\mathcal{P}$. Let $\mathbf{Q} \equiv \{ f_{Y,V|D,W}(\cdot|\cdot)q_{d,w} : f_{Y,V|D,W} \in \mathbf{P}_{d,w}, (d,w) \in \{0,1\} \times \mathbf{W} \}$. Let $\mathbf{v}_0 \in \mathbf{Q}$ be the true density and $Q$ the associated probability measure. We use subscript $Q$ for densities and expectations associated with $\mathbf{v}_0$. This subscript is not needed for the conditional densities (and conditional expectations) given $(D,W) = (d,w)$ or given $(D,W,V) = (d,w,v)$ because they remain the same both under $P$ and under $Q$. Use notations $\int \cdot d\mu(w)$, $\int \cdot d\mu(v)$, $\int \cdot d\mu(y)$, etc., to denote the integrations with respect to the marginals of $\mu$ for the coordinates of $w,v,y$, etc.

Since $A$ is compact, the space $(\mathcal{C}_b(A), \| \cdot \|)$ equipped with the supremum norm $\| \cdot \|$ is a Banach space. With abuse of notation, we view the treatment effect parameters $\tau_{ate}$ and $\tau_{tet}$ as maps from $\mathbf{Q}$ into $\mathcal{C}_b(A)$, so that, for example, $\tau_{ate}(\mathbf{v})$, $\mathbf{v} \in \mathbf{Q}$, is an element in $\mathcal{C}_b(A)$ but $\tau_{ate}(\mathbf{v})(p) \in \mathbb{R}$.

(i) First consider the semiparametric efficiency bound for $\tau_{ate}(\cdot)$. The proof is composed of three steps:

**Step 1. Calculate the tangent space.** Following Hahn (1998), under Condition C1 we write the density $f(y,v,d,w)$ as

$$f(y,v,d,w) = [f_1(y|v,w)p(v,w)]^d [f_0(y|v,w)(1-p(v,w))]^{1-d} f(v,w),$$

where

$$f(y|v,w) = \sum_{d=0}^{\infty} \sum_{w=0}^{\infty} f_{Y|V,D,W}(y|v,w,d,w) q_{d,w}.$$
where
\[ f_1(y|v, w) \equiv f(y|1, v, w), \]
\[ f_0(y|v, w) \equiv f(y|0, v, w), \]
and \( p(v, w) \equiv P\{D = 1|V = v, W = w\} \),
and \( f(y|d, v, w) \) denotes the conditional density of \( Y \) given \( (D, V, W) = (d, v, w) \). Consider a curve \( v(t) \) identified with \( f^t(y, v, d, w) \) \((|t| < 1)\), we have
\[ f^t(y, v, d, w) = \left[ f_1^t(y|v, w)p^t(v, w)\right]^d \left[ f_0^t(y|v, w) (1 - p^t(v, w))\right]^{1-d} f^t(v, w), \]
such that \( f^0(y, v, d, w) = f(y, v, d, w) \). Since \( f_Q(y, v, d, w) = f(y, v, d, w)q_{d,w}/p_{d,w} \), the density under \( Q \), \( f_Q(y, v, d, w) \) can be written as
\[ f_Q(y, v, d, w) = \left[ f_1(y|v, w)p(v, w)\right]^d \left[ f_0(y|v, w) (1 - p(v, w))\right]^{1-d} f(v, w)q_{d,w}/p_{d,w}, \]
and consider a curve \( Q^t \) identified with \( f^t(y, v, d, w)q_{d,w}/p_{d,w} \). The score of the above curve \( (P^t \text{ and } Q^t) \) is
\[ s^t(y, v, d, w) = ds_1(y|v, w) + (1 - d)s_0(y|v, w) + \frac{\partial p^t(v, w)/\partial t}{p^t(v, w)(1 - p^t(v, w))} [d - p(v, w)] + s^t(v, w), \]
where \( s_1^t(y|v, w) \), \( s_0^t(y|v, w) \) and \( s^t(v, w) \) are the score of \( f_1^t(y|v, w) \), \( f_0^t(y|v, w) \) and \( f^t(v, w) \) respectively. Also let \( s(y, v, d, w) \equiv s^0(y, v, d, w) \) (the score evaluated at the \( t = 0 \)). Now we can calculate the tangent set at \( v_0 \in Q \) as
\[ \dot{Q}^0 = \left\{ d h_1(y|v, w) + (1 - d) h_0(y|v, w) + a(v, w)(d - p(v, w)) + h(v, w) : h_1, h_0, a, h \in L_2(Q), \int h_1(y|v, w)f_1(y|v, w) = 0, \right\}
\[ \int h_0(y|v, w)f_0(y|v, w) = 0, \text{ and } \int h(v, w)f(v, w) = 0 \],
where we recall that \( Q \) in \( L_2(Q) \) is the probability measure associated with \( v_0 \). Observe that \( \dot{Q}^0 \) is linear and closed, so it is the tangent space which we denote by \( \dot{Q} \).

**Step 2.** Prove the pathwise differentiability of \( \tau_{ate} \) and compute its derivative. As for the pathwise differentiability, for given \( v_0 \in Q \), let \( V \subset Q \) be a curve passing through \( v_0 \), parametrized by \( t \in (-1, 1) \). Then the weighted average treatment effect under a point in this curve \( v(t) \), say, \( \tau_{ate}(v(t)) \) at \( p \in A \) is written as
\[ \sum_{w \in W} \int g(v, w) \left\{ f^t(y|1, v, w) - f^t(y|0, v, w) \right\} d\mu(y)f^t(v, w)d\mu(v) \]
\[ \sum_{(d,w) \in \{0,1\} \times W} p_{d,w} \int g(v, w)f^t(v|d, w)d\mu(v) \]
\[ = \sum_{(d,w) \in \{0,1\} \times W} \int g(v, w) \left\{ \int yf_1^t(y|v, w)d\mu(y) - \int yf_0^t(y|v, w)d\mu(y) \right\} p_{d,w}f^t(v|d, w)d\mu(v) \]
\[ \sum_{(d,w) \in \{0,1\} \times W} p_{d,w}f^t(v|d, w)d\mu(v) \]
for \( p \in A \). The first order derivative of \( \tau_{ate}(v(t))(p) \) with respect to \( t \) at \( t = 0 \) is equal to
\[
\frac{1}{E_p[g(X)]} E_p \left[ g(X) \left( E[Y s_1(Y | X)] - E[Y s_0(Y | X)] \right) \right] \\
+ \frac{1}{E_p[g(X)]} E_p \left[ s(V | D, W) g(X) \left\{ \tau(X) - \tau_{ate}(p) \right\} \right],
\]
where \( \tau(X) \equiv E_p [Y_1 - Y_0 | X] \). Let
\[
\dot{\psi}_{ate, p}(y, v, d, w) = \frac{g(v, w)}{E_p[g(X)]} \left( \frac{d(y - \beta_1(v, w))}{p_1(v, w)} - \frac{(1 - d)(y - \beta_0(v, w))}{p_0(v, w)} \right) + R_{d, ate}(v, w).
\]
(Recall \( R_{d, ate}(p)(v, w) \equiv t_{ate, p}(v, w) - E_{d, w}[t_{ate, p}(X)] \). We can write
\[
\frac{\partial \tau_{ate}(v(t))(p)}{\partial t} = \sum_{(d, w) \in (0, 1) \times \mathcal{W}} E_{d, w} \left[ \dot{\psi}_{ate, p}(Y, V, D, W) s(Y, V, D, W) \right] p_{d, w}.
\]
Define \( \dot{\psi}_{ate, Q}(y, v, d, w)(p) = \dot{\psi}_{ate, p}(y, v, d, w) p_{d, w} / q_{d, w} \) and rewrite
\[
(6.2) \quad \frac{\partial \tau_{ate}(v(t))(p)}{\partial t} = E_q \left[ \dot{\psi}_{ate, Q}(Y, V, D, W)(p) s(Y, V, D, W) \right].
\]
Define an operator \( \dot{\tau}_{ate} : \dot{Q} \rightarrow C_0(A) \) as
\[
\dot{\tau}_{ate}(s)(p) \equiv E_q \left[ \dot{\psi}_{ate, Q}(Y, V, D, W)(p) s(Y, V, D, W) \right], \quad s \in \dot{Q}, \ p \in A.
\]
Since (6.2) holds for all \( p \in A \) and \( \dot{\psi}_{ate, Q}(Y, V, D, W)(p) \) is continuous in \( p \) on the compact set \( A \), we have:
\[
(6.3) \quad \sup_{p \in A} |\tau_{ate}(v(t))(p) - \tau_{ate}(v_0)(p) - t \dot{\tau}_{ate}(s)(p)| = o(t), \quad \text{as} \ t \to 0,
\]
for all curves \( f_Q(y, v, d, w) = f_Q(y, v, d, w) + t s(y, v, d, w) + o(t) \). Under Conditions C2-C5,
\[
\sup_{p \in A} E_q [\dot{\psi}_{ate, Q}^2(Y, V, D, W)(p)] < \infty.
\]
Then there exists a finite \( M_1 \) such that
\[
\sup_{p \in A} E_q \left[ \dot{\psi}_{ate, Q}(Y, V, D, W)(p) s(Y, V, D, W) \right] \leq M_1 \sqrt{E_q [s^2(Y, V, D, W)]},
\]
for all \( s \in \dot{Q} \), which implies that \( \dot{\tau}_{ate} \) is bounded. Also obviously \( \dot{\tau}_{ate} \) is linear. Therefore \( \tau_{ate} \) is pathwise differentiable at \( v_0 \) with derivative \( \dot{\tau}_{ate} \).

**Step 3.** Calculate the efficient influence function, inverse information covariance functional and the semiparametric efficiency bound. For a generic element \( b^* \in (C_0(A))^* \) (the dual space
of $C_b(A)$, we have

$$b^* \hat{\tau}_{ate}(s) = E_Q \left[ (b^* \hat{\psi}_{ate,Q}(Y, V, D, W)) s(Y, V, D, W) \right].$$

Notice that $\hat{\psi}_{ate,Q} \in \hat{Q}$, so the linearity of expectation and the dual operator $b^*$ lead to $b^* \hat{\psi}_{ate,Q} \in \hat{Q}$. Then the projection of $b^* \hat{\psi}_{ate,Q}$ onto $\hat{Q}$ is itself and we obtain the efficient influence operator (see BKRW p.178 for its definition) of $\tau_{ate}$ as $\tilde{I}_{ate} : (C_b(A))^* \rightarrow \hat{Q}$, where $\tilde{I}_{ate}(b^*) = b^* \hat{\psi}_{ate,Q}$. In particular, for the evaluation map $b^* = \pi_p (\pi_p(b) \equiv b(p)$ for all $b \in C_b(A)$, the efficient influence operator becomes $\tilde{I}_{ate}(\pi_p) = \hat{\psi}_{ate,Q}(\cdot, \cdot, \cdot, \cdot)(p)$. Following BKRW p.184, the inverse information covariance functional for $\tau_{ate}$, $I_{ate}^{-1} : A \times A \rightarrow \mathbb{R}$ is given by

$$(6.4) \quad I_{ate}^{-1}(p, \tilde{p}) = E_Q[\hat{\psi}_{ate,Q}(Y, V, D, W)(p)\hat{\psi}_{ate,Q}(Y, V, D, W)(\tilde{p})].$$

By Theorem 5.2 BKRW(Convoluation Theorem), an efficient weakly regular estimator $\hat{\tau}_{ate}$ of $\tau_{ate}$ weakly converges to a mean zero Gaussian process $\zeta^*(\cdot)$ with the inverse information covariance functional $I_{ate}^{-1}(p, \tilde{p})$ characterized by (6.4). As a special case, the variance bound for any weakly regular estimator of the real parameter $\tau_{ate}(v_0)(p)$ can be written as:

$$\sum_{(d,w) \in \{0,1\} \times \mathcal{W}} E_{d,w}[\hat{\psi}_{ate,Q}^2(Y, V, D, W)]d_{d,w} = \sum_{(d,w) \in \{0,1\} \times \mathcal{W}} p_{d,w}^2 E_{d,w}[\hat{\psi}_{ate,P}^2(Y, V, D, W)].$$

(ii) Let us turn to the semiparametric efficiency bound for $\tau_{tet}$. The tangent space remains the same as that in (i). To establish the semiparametric efficiency bound, the only needed change is the computation of the efficient influence operator. Similarly as before, for given $v_0 \in Q$, let $V \subset Q$ be a curve passing through $v_0$, parametrized by $t \in (-1, 1)$. The weighted average treatment effect on the treated under a point in this curve $v(t)$, say, $\tau_{tet}(v(t))$ at $p \in A$ is written as

$$\tau_{tet}(v(t))(p) = \frac{\sum_{w \in \mathcal{W}} \int \int g(v, w)g \{f^0(y|v, w, 1) - f^0(y|v, w, 0)\} d\mu(y)f^0(v|w, 1)p_{w|1}d\mu(v)}{\sum_{w \in \mathcal{W}} \int \int g(v, w)f^0(y|v, w, 1)d\mu(y)f^0(v|w, 1)p_{w|1}d\mu(v)},$$

where $p_{w|1} = p_{1,w}/\{\sum_{w \in \mathcal{W}} p_{1,w}\}$. The first order derivative of $\tau_{tet}(v(t))(p)$ with respect to $t$ is equal to

$$\frac{1}{\hat{g}_{1,p}} \left( E_{1,p} [g(X)\{E[Ys_1(Y|X)|X, D = 1] - E[Ys_0(Y|X)|X, D = 0]\}] + E_{1,p} [s(V|D, W)g(X)\{\tau(X) - \tau_{tet}\}] \right),$$

where $\hat{g}_{1,p} = \frac{1}{\{\sum_{w \in \mathcal{W}} p_{1,w}\}}$.
where \( \tilde{g}_{1,p} = E_{1,p}[g(X)] \). We take

\[
\dot{\psi}_{tet,p}(y, v, d, w) = \frac{1}{\tilde{g}_1} \left\{ \begin{array}{c} \frac{dg(v, w)(y - \beta_1(v, w))}{p_1} \\
-p_1(v, w)(1 - d)g(v, w)(y - \beta_0(v, w))/\{p_0(v, w)p_1\} \\
dR_{1,tet}(p)(v, w)/p_1 \end{array} \right\}
\]

(Recall \( R_{1,tet}(p)(v, w) \equiv t_{tet,p}(v, w) - E_{1,w}[t_{tet,p}(X)] \).) The remainder of the proof follows the argument in the proof of (i): we construct \( \dot{\psi}_{tet,Q}(y, v, d, w)(p) \equiv \dot{\psi}_{tet,p}(y, v, d, w)p_{d,w}/q_{d,w}. \)

Under Conditions C1-C5, we can verify the pathwise differentiability of \( \tau_{tet} : Q \rightarrow C_{b}(A) \) at \( v_0 \). Write the efficient influence operator as \( I_{tet}(b^*) = b^* \dot{\psi}_{tet,Q} \) and compute the inverse information covariance functional as

\[
I_{tet}^{-1}(p, \bar{p}) = E_Q[\dot{\psi}_{tet,Q}(Y, V, D, W)(p)\dot{\psi}_{tet,Q}(Y, V, D, W)(\bar{p})].
\]

Let us turn to the situation with pure treatment-based sampling, where parameter \( \tau_{tet}(p) \) does not depend on \( p \), thus for each \( v \in Q, \tau_{tet}(v) \) is a constant real map on \( A \). We simply write \( \tau_{tet} \) suppressing the argument \( p \). In this special case of pure treatment-based sampling, the functional \( I_{tet}^{-1}(p, \bar{p}) \) and find out that it no longer depends on \( (p, \bar{p}) \). In particular, write

\[
\tau_{tet}(v(t))(p) = \int \int yg(x) \{ f^t(y|x, 1) - f^t(y|x, 0) \} d\mu(y) f^t(x|1)d\mu(x).
\]

The first order derivative of \( \tau_{tet}(p) \) with respect to \( t \) is equal to

\[
\frac{1}{\tilde{g}_1} \begin{array}{c} E_1 [g(X)\{E [Ys_1(Y|X)|X, D = 1] - E [Ys_0(Y|X)|X, D = 0]\}] \\
+ E_1 [s(X|D)g(X)\{\tau(X) - \tau_{tet}(p)\}] \end{array},
\]

where \( \tilde{g}_1 = E_1[g(X)] \). Therefore, we take

\[
\dot{\psi}_{tet,p}(y, x, d) = \frac{1}{\tilde{g}_1} \left\{ \begin{array}{c} \frac{dg(x)(y - \beta_1(x) - \{\tau(x) - \tau_{tet}\})}{p_1} \\
p_1(x)(1 - d)(y - \beta_0(x))/p_0(x)p_1 \end{array} \right\},
\]

because \( E_{1,p} [\tau(X) - \tau_{tet}(p)] = 0 \). Let \( \dot{\psi}_{tet,Q}(y, x, d)(p) \equiv \dot{\psi}_{tet,p}(y, x, d)p_{d}/q_d. \) Now the inverse information covariance functional becomes

\[
I_{tet}^{-1}(p, \bar{p}) = \sum_{(d \in \{0,1\}}} q_d E_d \left[ \dot{\psi}_{tet,Q}(Y, X, D)(p)\dot{\psi}_{tet,Q}(Y, X, D)(\bar{p}) \right]
\]

\[
= \frac{1}{\tilde{g}_1^2 \tilde{q}_1} E_1 \left[ g(X)^2(Y_1 - \beta_1(X) - \{\tau(X) - \tau_{tet}\})^2 \right]
+ \frac{1}{\tilde{g}_1^2 \tilde{q}_0 q_1 p_1} E_0 \left[ p_1(X)p_1(X)g(X)^2(Y_0 - \beta_0(X))^2 \right] / p_0(X)p_0(X).
\]
Note that by Bayes’ rule,
\[
\frac{p_0 p_1(X)}{p_1 p_0(X)} = \frac{p_0 f(X|1)p_1}{p_1 f(X|0)p_0} = f(X|1) = \frac{\tilde{p}_0 f(X|1)\tilde{p}_1}{\tilde{p}_1 f(X|0)\tilde{p}_0} = \frac{\tilde{p}_0 \tilde{p}_1(X)}{\tilde{p}_1 \tilde{p}_0(X)}.
\]

We rewrite the last term in (6.5) as
\[
\frac{1}{g_1} \frac{1}{q_0} E \left[ \frac{f^2(X|1)}{f^2(X|0)} g(X)^2 (Y_0 - \beta_0(X))^2 \right].
\]

Thus the semiparametric efficiency bound does not depend on \( p \equiv \{p_d\} \).

REFERENCES


Supplemental Note for "Optimal Inference on Treatment Effects under Treatment-Based Sampling"

Kyungchul Song and Zhengfei Yu

The supplemental note is a continuation of the appendix in Song and Yu (2014), and contains the mathematical proofs of Theorem 2 in the paper. The supplemental note is divided into two sections. The first section presents the mathematical proofs of some auxiliary results and Theorem 2. The second section contains further auxiliary results that are used in the first section.

6.3. Appendix C: Efficient Estimation and Proofs. For the proof of Theorem 2, we first establish the asymptotic linear representations for \( \tilde{\tau}_{ate} \), \( \tilde{\tau}_{tet} \), \( \tilde{\tau}_{F1} \) and \( \tilde{\tau}_{F1} \). We introduce some notations. (Throughout this supplemental note, we suppress \( p \) in \( E_p \) and \( E_{d,p} \) from the notation and write \( E \) and \( E_d \) simply.) First, define mean-deviated quantities:

\[
\begin{align*}
\xi_{d,ate}(V_i, w) & \equiv g(V_i, w) [\tau(V_i, w) - \tau_{ate}(p)] - E_{d,w}[g(V_i, w)(\tau(V_i, w) - \tau_{ate}(p))], \\
\xi_{1,tet}(V_i, w) & = g(V_i, w) [\tau(V_i, w) - \tau_{tet}(p)] - E_{1,w}[g(V_i, w)(\tau(V_i, w) - \tau_{tet}(p))],
\end{align*}
\]

where

\[
\tau(X) \equiv E [Y_i | X] - E [Y_0 | X].
\]

Also, define

\[
\varepsilon_{d,w,i} \equiv Y_{di} - \beta_d(V_i, w) \text{ and } \varepsilon_{d,1,w,i} = Y_{di} - E_{1,p}[\beta_d(X_i)].
\]

Lemmas A1 and A2 below establish asymptotic linear representations for \( \tilde{\tau}_{ate} \), \( \tilde{\tau}_{tet} \), \( \tilde{\tau}_{F1} \) and \( \tilde{\tau}_{F1} \). For the asymptotic linear representation, we define

\[
Z_i(p) = \frac{1}{E g(X_i)} \sum_{w \in W} \left\{ \frac{g(V_i, w)L_{1,w,i}(p)\varepsilon_{1,w,i}}{p_1(V_i, w)} - \frac{g(V_i, w)L_{0,w,i}(p)\varepsilon_{0,w,i}}{p_0(V_i, w)} \right\} \\
+ \frac{1}{E g(X_i)} \sum_{w \in W} (\xi_{ate}(V_i, w)L_{1,w,i}(p) + \xi_{ate}(V_i, w)L_{0,w,i}(p)).
\]
and
\[
\tilde{Z}_i(p) = \frac{1}{p_1 E_1 p g(X_i)} \sum_{w \in W} \left\{ g(V_i, w) L_{1, w, i}(p) \varepsilon_{1, 1, w, i} - \frac{g(V_i, w) L_{0, w, i}(p) p_1(V_i, w) \varepsilon_{0, 1, w, i}}{p_{0, w}(V_i, w)} \right\} + \frac{1}{p_1 E_1 p g(X_i)} \sum_{w \in W} \xi_{1, tet}(V_i, w) L_{1, w, i}(p).
\]

From here on, we suppress the argument notation and write \(L_{d, w, i}(p)\) simply as \(L_{d, w, i}(p)\).

**Lemma A1:** Suppose that Condition (C1) and Assumptions 1-2 hold. Then uniformly over \(p \in A\),
\[
\sqrt{n} \left( \tau_{ate}^{SA}(p) - \tau_{ate}(p) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(p) + o_P(1), \quad \text{and}
\]
\[
\sqrt{n} \left( \tau_{tet}^{SA}(p) - \tau_{tet}(p) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{Z}_i(p) + o_P(1).
\]

**Lemma A2:** Suppose that the Condition (C1) and Assumptions 1-2 hold. Then uniformly for \(p \in A\),
\[
\sqrt{n} \left( \tau_{ate}^{FI}(p) - \tau_{ate}(p) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(p) + o_P(1), \quad \text{and}
\]
\[
\sqrt{n} \left( \tau_{tet}^{FI}(p) - \tau_{tet}(p) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{Z}_i(p) + o_P(1).
\]

**Proof of Theorem 2:** We focus on \(\tau_{ate}^{SA}(\cdot)\) and \(\tau_{ate}^{FI}(\cdot)\) only. The proof for the cases of \(\tau_{ate}^{FI}(\cdot)\) and \(\tau_{tet}^{FI}(\cdot)\) are similar. By Lemma A1, it suffices to prove that
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(\cdot) \sim \zeta_{ate}(\cdot).
\]

Since \(E_Q[Z_i(p)] = 0\) and \(E_Q[Z_i^2(p)] < \infty\) for all \(p\), for every finite subset \(\{p_1, \ldots, p_K\} \subseteq A\), the Central Limit Theorem yields that \((Z_i(p_1), \ldots, Z_i(p_K))\) converges in distribution to a normal distribution with mean zero and covariance matrix \(\Sigma \equiv [\sigma_{kl}]\), where
\[
\sigma_{kl} = \sum_{(d, w) \in \{0, 1\} \times W} \frac{p_{k, d, w} p_{l, d, w}}{q_{d, w}} E_{d, w} \left[ e_d(p_k) e_d(p_l) + \zeta_{ate}(p_k) \zeta_{ate}(p_l) \right].
\]
Now we verify the stochastic equicontinuity of the process \((1/\sqrt{n}) \sum_{i=1}^{n} Z_i(p)\). Note that \(Z_i(p)\) is differentiable with respect to \(p\). By the mean-value theorem,

\[
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(p) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(\bar{p}) \right| \leq \left( \sup_{p \in A} \sum_{(d,w) \in \{0,1\} \times \mathcal{W}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial Z_i(p)}{\partial p_{d,w}} \right| \right) \| p - \bar{p} \|,
\]

for any pair of \(p, \bar{p} \in A\). Therefore, the stochastic equicontinuity follows once we show that

\[
\sup_{p \in A} \sum_{(d,w) \in \{0,1\} \times \mathcal{W}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial Z_i(p)}{\partial p_{d,w}} \right| = O_p(1).
\]


We write

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial Z_i(p)}{\partial p_{d,w}} = \frac{\mathbf{E}[g(V_i, w)]}{(\mathbf{E}[g(X_i)])^2} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial Q_{i,d,w}(p)}{\partial p_{d,w}} \right\} - \frac{\mathbf{E}_{d,w}[g(V_i, w)]}{(\mathbf{E}[g(X_i)])^2} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i(p) \right\},
\]

where \(Q_i(p) \equiv Z_i(p)\mathbf{E}g(X_i)\) and \(Q_{i,d,w}(p) \equiv \partial Q_i(p)/\partial p_{d,w}\). Since \(\mathbf{E}[g(V_i, w)]/(\mathbf{E}[g(X_i)])^2\) and \(\mathbf{E}_{d,w}[g(V_i, w)]/(\mathbf{E}[g(X_i)])^2\) are bounded uniformly over \(p \in A\), it suffices for (6.11) to show that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_{i,d,w}(\cdot) \quad \text{and} \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i(\cdot)
\]

weakly converge in \(l_\infty(A)\). This can be shown by establishing the convergence of the finite dimensional distributions using the Central Limit Theorem, and stochastic equicontinuity of the processes which follows by showing the first order derivatives of the two processes in (6.12) are stochastically bounded uniformly over \(p \in A\). Details are omitted. □

6.4. Appendix D: Further Auxiliary Results. In this section, we present the proofs of Lemmas A1 and A2. We begin with auxiliary results. Let us introduce some definitions: for \(d = 0, 1\),

\[
\hat{p}_{d,i}(V_i, w) \equiv \frac{\hat{\lambda}_{d,i}(V_i, w)}{\hat{\lambda}_{1,i}(v, w) + \hat{\lambda}_{0,i}(V_i, w)},
\]

where \(\hat{\lambda}_{d,i}(V_i, w) \equiv \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{d,w,j} K_h (V_{ij} - V_{ii}) 1\{V_{2j} = V_{2i}\}\). Also, define

\[
\hat{i}_{n,i} \equiv 1 \left\{ \hat{\lambda}_{1,i}(V_i, w) \wedge \hat{\lambda}_{0,i}(V_i, w) \geq \delta_n : d \in \{0, 1\} \right\},
\]

where \(\delta_n\) is a sequence that appears in Assumption 2(iii).
**Lemma B1:** Suppose that Assumptions 1-4 hold. Then, for each \( w \in \mathcal{W} \), uniformly over \( p \in A \),

\[
\max_{1 \leq i \leq n} |\hat{p}_{1,i}(V_i, w) - \hat{p}_{1,i}(V_i, w)| = O_P(\varepsilon_n) \quad \text{and} \quad \max_{1 \leq i \leq n} \tilde{1}_{n,i} |p_1(V_i, w) - \tilde{p}_{1,i}(V_i, w)| = O_P(\varepsilon_n),
\]

where \( \varepsilon_n = h^{-d_1/2} \sqrt{\log n} + h^{L_1+1} \).

**Proof:** Consider the first statement. For simplicity, we assume that \( V = V_1 \) and define \( \mathbf{E}_{Q,w,i}[L_{1,w,i}] = \mathbf{E}_Q[L_{1,w,i}|V, W_i = w] \) and \( \mathbf{E}_{Q,w,i}[L_{w,i}] = \mathbf{E}_Q[L_{w,i}|V, W_i = w] \). Recall that \( q_1(v, w) \) is the propensity score under \( Q \), i.e., \( q_1(v, w) = Q \{ D_i = 1 | (V_i, W_i) = (v, w) \} \). By Bayes’ rule,

\[
f(V_i|1, w) = q_{1,w}(V_i)f_Q(V_i)/q_{1,w} = q_1(V_i, w)q_w(V_i)f_Q(V_i)/q_{1,w},
\]

where \( q_{1,w}(V_i) = \mathbf{E}_Q[1\{(D_i, W_i) = (d, w)\}|V_i] \), \( q_w(V_i) = \mathbf{E}_Q[1\{W_i = w\}|V_i] \) and \( f_Q(\cdot) \) is the density of \( V_i \) under \( Q \). Hence

\[
p_1(V_i, w) = \frac{f(V_i|1, w)p_{1,w}}{f(V_i|1, w)p_{1,w} + f(V_i|0, w)p_{0,w}} = \frac{(q_1(V_i, w)/q_{1,w})p_{1,w}}{\sum_{d \in \mathcal{D}}(q_d(V_i, w)/q_{d,w})p_{d,w}} = \frac{\mathbf{E}_{Q,w,i}[L_{1,w,i}]}{\mathbf{E}_{Q,w,i}[L_{w,i}]}.
\]

Let \( K_{ji} = K_h(V_{ij} - V_{ii})1\{V_{2j} = V_{2i}\} \) for brevity. Also let

\[
\hat{\mathbf{E}}_{Q,w,i}[L_{1,w,i}] = \frac{1}{n-1} \sum_{j=1, j \neq i}^n L_{1,w,j}K_{ji} \quad \text{and} \quad \hat{\mathbf{E}}_{Q,w,i}[L_{w,i}] = \frac{1}{n-1} \sum_{j=1, j \neq i}^n L_{w,j}K_{ji}.
\]

By applying Theorem 6 of Hansen (2008), we find that uniformly over \( i \in \{1, ..., n\} \),

\[
\mathbf{E}_{Q,w,i}[L_{1,w,i}] - \hat{\mathbf{E}}_{Q,w,i}[L_{1,w,i}] = O_P(\varepsilon_n),
\]

\[
\mathbf{E}_{Q,w,i}[L_{w,i}] - \hat{\mathbf{E}}_{Q,w,i}[L_{w,i}] = O_P(\varepsilon_n).
\]

Furthermore, \( (6.16) \) holds uniformly for all \( p \in A \), because

\[
\mathbf{E}_{Q,w,i}[L_{1,w,i}] - \hat{\mathbf{E}}_{Q,w,i}[L_{1,w,i}]
= \frac{p_{1,w}}{q_{1,w}} \left\{ \sum_{j=1, j \neq i}^n 1\{(D_i, W_i) = (1, w)\}K_{ji}/\sum_{j=1, j \neq i}^n 1\{W_j = w\}K_{ji} - \mathbf{E}_{Q,w,i}[1\{(D_i, W_i) = (1, w)\}] \right\}.
\]
The term in the bracket is $O_P(\varepsilon_n)$ by Theorem 6 of Hansen (2008), and this convergence is uniformly for all $p$ since it does not depend on $p$. Observe that

$$
\hat{\tau}_{n,i} \left[p_1(V_i, w) - \hat{p}_1(V_i, w)\right] = \frac{\hat{p}_{Q,i}[L_{1,w,i}] - \hat{E}_{Q,w,i}[L_{1,w,i}]}{\hat{E}_{Q,w,i}[L_{w,i}]} + \hat{\tau}_{n,i} \frac{\hat{E}_{Q,w,i}[L_{w,i}] \left(\hat{E}_{Q,w,i}[L_{w,i}] - \hat{E}_{Q,w,i}[L_{w,i}]\right)}{(\hat{E}_{Q,w,i}[L_{w,i}])^2} + o_P(\varepsilon_n).
$$

(6.18)

Using Bayes’ rule, we deduce that

$$
E_{Q,w,i}[L_{w,i}] = \frac{P_{Q}(D_i = 1|V_i, W_i = w)}{q_{1,w}} + \frac{P_{Q}(D_i = 0|V_i, W_i = w)}{q_{0,w}}
$$

$$
= p_{1,w} f_Q(V_i|w, 1) q_{1,w} + p_{0,w} f_Q(V_i|w, 0) q_{0,w}
$$

$$
= p_{1,w} f_P(V_i|w, 1) f_Q(V_i|w) + p_{0,w} f_P(V_i|w, 0) f_Q(V_i|w),
$$

$$
= \frac{f_P(V_i, w)}{f_Q(V_i, w)} q_{1,w} f_P(V_i|w, 1) + q_{0,w} f_P(V_i|w, 0).
$$

Therefore,

$$
E_Q \left[ \sup_{p \in A} (E_{Q,w,i}[L_{w,i}])^{-a} \right] = E_Q \left[ \sup_{p \in A} \left( \frac{q_{1,w} f_P(V_i|w, 1) + q_{0,w} f_P(V_i|w, 0)}{f_P(V_i, w)} \right)^a \right] \leq 2^{a-1} \sum_{d,w} q_{d,w} \left[ E_{d,w} \left[ \sup_{p \in A} \left( \frac{f_P(V_i|w, 1)}{f_P(V_i, w)} \right)^a \right] \right] + E_{d,w} \left[ \sup_{p \in A} \left( \frac{f_P(V_i|w, 0)}{f_P(V_i, w)} \right)^a \right] < \infty,
$$

(6.19) for $a \in [1, \bar{a}]$, where $\bar{a} \geq 4$. Combining this with (6.16) and (6.18), we have

$$
\hat{\tau}_{n,i} \{p_1(V_i, w) - \hat{p}_1(V_i, w)\} = O_P(\varepsilon_n),
$$

uniformly over $p \in A$ and over $1 \leq i \leq n$. Hence we obtain the first statement of (6.13).

For the second statement of (6.13), let

$$
\hat{E}_{Q,w,i}[\hat{L}_{1,w,i}] = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji}
$$

and

$$
\hat{E}_{Q,w,i}[\hat{L}_{w,i}] = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \hat{L}_{w,j} K_{ji}.
$$
Observe that
\[ \left| \hat{E}_{Q,w,i}[L_{1,w,i}] - \hat{E}_{Q,w,i}[L_{1,w,i}] \right| \leq \frac{p_{d,w}}{q_{d,w}} - \frac{p_{d,w}}{q_{d,w}} \cdot \frac{\sum_{j \in S_{d,w} \setminus \{i\}} K_{ji}}{\sum_{j \in S_{w} \setminus \{i\}} K_{ji}} = o_P(\varepsilon_n). \]

Hence the argument in the proof of first statement can be applied to prove the second statement of (6.13). □

**Lemma B2 :** Suppose that \( S_i = \varphi(Y_i, X_i, D_i) \), for a given real-valued map \( \varphi \) such that for each \( w \in \mathcal{W} \), \( \sup_{v \in \mathcal{V}(w)} E_Q[|S_i|^2](V_i, W_i) = (v, w) < \infty \), and \( E_Q[S_i|V_i = \cdot](V_2, W) = (v_2, w) \) is \( L_1 + 1 \) times continuously differentiable with bounded derivatives and uniformly continuous \( (L_1 + 1) \)-th derivatives.

(i) Suppose that Condition (C1) and Assumptions 1-2 hold. Then, for \( d = 0, 1 \),
\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} (p_d(V_i, w) - \hat{p}_{d,i}(V_i, w)) = -\frac{1}{n} \sum_{i=1}^{n} \frac{E_{Q,w,i}[S_i]}{E_{Q,w,i}[L_{w,i}]} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{E_{Q,w,i}[S_i]P_d(V_i, w)J_{w,i}}{E_{Q,w,i}[L_{w,i}]} + o_P(n^{-1/2}),
\]
uniformly for \( p \in A \), where \( J_{d,w,i} \equiv L_{d,w,i} - E_{Q,w,i}[L_{d,w,i}] \) and \( J_{w,i} = J_{1,w,i} + J_{0,w,i} \).

(ii) Suppose that Condition (C1) and Assumptions 1-2 hold. Then, for \( d = 0, 1 \),
\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} (\hat{p}_{d,i}(V_i, w) - \hat{p}_{d,i}(V_i, w)) = E_{Q,w}[p_{1-d}(V_i, w)p_d(V_i, w)S_i] \left( \frac{\hat{q}_{d,w} - q_{d,w}}{q_{d,w}} - \frac{\hat{q}_{1-d,w} - q_{1-d,w}}{q_{1-d,w}} \right) + o_P(n^{-1/2}),
\]
uniformly for \( p \in A \).

**Proof:** (i) By adding and subtracting the sum:
\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} \left( \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji} \right) \frac{E_{Q,w,i}[L_{w,i}]f_Q(V_i, w)}{E_{Q,w,i}[L_{w,i}]f_Q(V_i, w)},
\]
and noting (6.15), we write

\begin{equation}
\frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left( \hat{p}_1(V_i, w) - \hat{p}_1(V_i, w) \right)
\end{equation}

\begin{align}
&= \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\mathbb{E}_{Q,w,i}[L_{1,w,i}]}{\mathbb{E}_{Q,w,i}[L_{w,i}]} - \sum_{j=1, j \neq i}^{n} \frac{L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} - \sum_{j=1, j \neq i}^{n} \frac{L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} - \sum_{j=1, j \neq i}^{n} \frac{L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} \right\}.
\end{align}

We write the last sum as

\begin{align}
&= \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} - \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \right\} \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} \right\} + o_P(n^{-1/2}) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} - \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \right\} + o_P(n^{-1/2}).
\end{align}

uniformly for all \( p \in A \). The first equality uses Lemma B1 and the second uses (6.15). Let

\[ K_{n,i} \equiv \mathbb{E}_{Q,w,i}[L_{w,i}] - \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji}, \]

and write the last sum as

\begin{align}
&= -\frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \mathbb{E}_{Q,w,i}[L_{w,i}] \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \hat{K}_{n,i} \left\{ \frac{\sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{ji}}{f_{Q}(V_i, w)} - \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \right\} + o_P(n^{-1/2}).
\end{align}

Observe that

\begin{equation}
1 - \hat{K}_{n,i} \leq 1 \left\{ \hat{\lambda}_{1,i}(V_i, w) < \delta_n \right\} + 1 \left\{ \hat{\lambda}_{0,i}(V_i, w) < \delta_n \right\}.
\end{equation}

We write the first indicator on the right hand side as

\begin{equation}
1 \left\{ \frac{\hat{E}_{Q,w,i}[L_{1,w,i}]}{n-1} \sum_{j=1, j \neq i}^{n} 1\{W_j = w\} K_{h,j,i} < \delta_n \right\} \leq 1 \left\{ \mathbb{E}_{Q,w,i}[L_{1,w,i}] < \kappa_n \right\},
\end{equation}
where \( \kappa_n = (\delta_n + R_{1n}) / c \) (with \( c > 0 \) such that \( \min_{w \in W} \inf_{v \in \mathcal{V}(w)} f_Q(v, w) > c \) (see Assumption 1(iii)) and

\[
R_{1n} \equiv \max_{1 \leq i \leq n} \left| \frac{\mathbb{E}_{Q,w,i}[L_{1,w,i}]}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,j,i} - \mathbb{E}_{Q,w,i}[L_{1,w,i}] \cdot f_Q(V_i, w) \right|.
\]

Note that from (6.16), we have \( R_{1n} = O_P(\varepsilon_n) \). Thus we can take a nonstochastic sequence \( \kappa'_n \) and \( \eta > 0 \) such that \( \kappa''_n = o(n^{-1/2}) \) and \( \max\{\gamma, 2\} \leq \eta \leq \bar{a} \), using Assumptions 2(ii) and (iii). (Here \( \gamma \) and \( \bar{a} \) are constants in Assumptions 2(iii) and 1(iv).) Replacing \( \kappa_n \) in (6.22) by this \( \kappa'_n \), we find that with probability approaching one, we have

\[
\sup_{p \in A} \left| \frac{1}{n} \sum_{i=1}^{n} S_i \left\{ 1 - \hat{1}_{n,i} \right\} p_1(V_i, w) K_n \right| \leq \sup_{p \in A} \left\{ \frac{K_n}{n} \sum_{i=1}^{n} S_i p_1(V_i, w) \right\} \left( 1 \left\{ \mathbb{E}_{Q,w,i}[L_{1,w,i}] \leq \kappa'_n \right\} + 1 \left\{ \mathbb{E}_{Q,w,i}[L_{0,w,i}] \leq \kappa'_n \right\} \right),
\]

where \( K_n = \max_{1 \leq i \leq n} |K_{n,i}|. \) It is not hard to see that \( \sup_{p \in A} K_n = O_P(1) \), because

\[
\sup \max_{1 \leq i \leq n} |K_{n,i}| \leq \sup_{p \in A} \sup_{w \in W} \sup_{v \in \mathcal{V}(w)} \frac{2f_P(v, w)}{f_Q(v, w)} + O_P(\varepsilon_n) = O_P(1)
\]

and \( \min_{w \in W} \inf_{v \in \mathcal{V}(w)} f_Q(v, w) > c \) for some positive constant \( c > 0 \), using Assumption 1(iii). Note that the expectation \( \mathbb{E}_Q \) of (6.24) is bounded by (for some \( C > 0 \))

\[
C \kappa''_n \mathbb{E}_Q \left[ \mathbb{E}^{\eta}_{Q,w,i}[L_{w,i}] \right] = O \left( \kappa''_n \right) = o(n^{-1/2}),
\]

uniformly over \( p \in A \). Hence we conclude that

\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} p_1(V_i, w) K_{n,i} = \frac{1}{n} \sum_{i=1}^{n} S_i p_1(V_i, w) K_{n,i} + o_P(n^{-1/2}),
\]

uniformly over \( p \in A \). Applying the similar argument to the second to the last sum of (6.20) to eliminate \( \hat{1}_{n,i} \), we finally write

\[
= \frac{1}{n} \sum_{i=1}^{n} S_i \left( p_1(V_i, w) - \hat{p}_i(V_i, w) \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{S_i}{\mathbb{E}_{Q,w,i}[L_{w,i}]} \left\{ \mathbb{E}_{Q,w,i}[L_{1,w,i}] - \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{1,w,j} K_{j,i} \right\} \right\}
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} S_i p_1(V_i, w) \left\{ \mathbb{E}_{Q,w,i}[L_{w,i}] - \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{j,i} \right\} + o_P(n^{-1/2}),
\]
uniformly over \( p \in A \). By Lemma D1 below, the difference of the last two terms is asymptotically equivalent to (up to \( o_P(n^{-1/2}) \), uniformly over \( p \in A \)).

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ \frac{\mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{1,w,i}]}{\mathbb{E}_{Q,w,i}[L_{w,i}]} \right] \right. \\
- \left. \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ \frac{\mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{w,i}]}{\mathbb{E}_{Q,w,i}[L_{w,i}]} \right] \right. \\
- \left. \mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{w,i}] \right\} \right.
\]

\[
= -\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ \mathbb{E}_{Q,w,i}[S_i]p_1(V_i, w)\mathbb{E}_{Q,w,i}[L_{w,i}] \right] \right. \\
- \left. \mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{w,i}] \right\} \right.
\]

\[
= 1 \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ \mathbb{E}_{Q,w,i}[S_i]p_1(V_i, w) \right] \right. \\
- \left. \mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{w,i}] \right\} \right.
\]

\[
= 1 \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ \mathbb{E}_{Q,w,i}[S_i]p_1(V_i, w) \right] \right. \\
- \left. \mathbb{E}_{Q,w,i}[S_i]\mathbb{E}_{Q,w,i}[L_{w,i}] \right\} \right.
\]

\[
= 0,
\]

using (6.15).

(ii) We focus on the case of \( d = 1 \). The case for \( d = 0 \) can be dealt with precisely in the same way. First, we let \( 1_{n,i} = 1 \{ \mathbb{E}_{Q,w,i}[L_{w,i}] \geq \delta_n \} \), and write

\[
(6.26) \quad \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w))
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w))
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{n,i} (1 - 1_{n,i}) (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w)).
\]

By Lemma B1, \( \max_{1 \leq i \leq n} |\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w)| \hat{1}_{n,i} = O_P(\varepsilon_n) \) uniformly over \( p \in A \). Furthermore, since \( S_i \)'s are i.i.d. under \( Q \), and the \( r \)-th conditional moment given \( (V_i, W_i) =
(v, w) is bounded uniformly over v ¥ V(w) and over w ¥ W, we find that

\[(6.27) \quad E_Q \left[ \frac{1}{n} \sum_{i=1}^{n} |S_i| (1 - 1_{n,i}) \right] \leq C E_Q \left[ 1 \{ E_{Q,w,i} [L_{w,i}] \leq \delta_n \} \right] \]
\[\leq C \delta_n E_Q \left[ E_{Q,w,i}^\alpha [L_{w,i}] \right],\]

by Markov’s inequality, for some \( \gamma \leq \alpha \leq \bar{\alpha} \). By (6.19), the last expectation is finite. Since \( \delta_n = o(n^{-1/2}) \) (Assumption 2(iii)), we conclude that

\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w)) = \frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w)) + o_P(n^{-1/2}),
\]

uniformly for \( p \in A \).

As for the leading sum on the right hand side (6.26), note that

\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} (\hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w)) = \frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \left\{ \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji} - \sum_{j=1, j \neq i}^{n} \hat{L}_{w,j} K_{ji} \right\}
\]
\[= \frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \left\{ \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} - \hat{L}_{w,j} \right\} K_{ji} + \frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji} \left\{ \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \right\} - \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \text{.}
\]

Now, note that as for the second term,

\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji} \left\{ \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \right\} - \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \text{.}
\]
\[= \frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji} \left\{ \sum_{j=1, j \neq i}^{n} \hat{L}_{w,j} - L_{w,j} \right\} K_{ji} \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} \text{.}
\]

Using Lemma B1, we can write the last sum as

\[
\frac{1}{n} \sum_{i=1}^{n} S_i \hat{I}_{n,i} \sum_{j=1, j \neq i}^{n} \hat{L}_{1,w,j} K_{ji} \left\{ \sum_{j=1, j \neq i}^{n} \hat{L}_{w,j} - L_{w,j} \right\} K_{ji} \frac{1}{n} \sum_{j=1, j \neq i}^{n} L_{w,j} K_{ji} + o_P(n^{-1/2}),
\]
uniformly for $p \in A$. Therefore, we can write

$$
\begin{align*}
(6.28) \quad & \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{i,n,1} \{ \hat{p}_{1,i}(V_i, w) - \hat{p}_{1,i}(V_i, w) \} \\
\quad & = -\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{i,n,1} \sum_{j=1,j \neq i}^{n} L_{0,w,j} K_{ji} \sum_{j=1,j \neq i}^{n} \left\{ \hat{L}_{1,w,j} - L_{1,w,j} \right\} K_{ji} \\
\quad & + \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{i,n,1} \sum_{j=1,j \neq i}^{n} L_{0,w,j} K_{ji} \sum_{j=1,j \neq i}^{n} \left\{ \hat{L}_{1,w,j} - L_{1,w,j} \right\} K_{ji} + o_P(n^{-1/2}) \\
\quad & = -\frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{i,n,1} \frac{p_0(V_i, w) \sum_{j=1,j \neq i}^{n} \left\{ \hat{L}_{1,w,j} - L_{1,w,j} \right\} K_{ji}}{\sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji}} \\
\quad & + \frac{1}{n} \sum_{i=1}^{n} S_i \hat{1}_{i,n,1} \frac{p_1(V_i, w) \sum_{j=1,j \neq i}^{n} \left\{ \hat{L}_{0,w,j} - L_{0,w,j} \right\} K_{ji}}{\sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji}} + o_P(n^{-1/2}),
\end{align*}
$$

uniformly over $p \in A$. Here the uniformity over $p \in A$ follows from

$$
\frac{\sum_{j=1,j \neq i}^{n} L_{0,w,j} K_{ji}}{\sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji}}
$$

$$
= \left( 1 + \frac{p_1,w q_0,w \sum_{j=1,j \neq i}^{n} 1 \{(D_j, W_j) = (1, w)\} K_{ji}}{p_0,w q_1,w \sum_{j=1,j \neq i}^{n} 1 \{(D_j, W_j) = (0, w)\} K_{ji}} \right)^{-1},
$$

where $\sum_{j=1,j \neq i}^{n} 1 \{(D_j, W_j) = (d, w)\} K_{ji} / \sum_{j=1,j \neq i}^{n} 1 \{W_i = w\} K_{ji}$ converges to $q_d(V_i, w)$ and does not depend on $p$.

We write

$$
(6.29) \quad \hat{1}_{i,n,1} \sum_{j=1,j \neq i}^{n} \frac{\hat{L}_{1,w,j} - L_{1,w,j}}{L_{w,j} K_{ji}}
$$

$$
= \left( \frac{p_1,w}{q_1,w} - \frac{p_1,w}{q_1,w} \right) \frac{\hat{1}_{i,n,1} \sum_{j=1,j \neq i}^{n} 1 \{(D_j, W_j) = (1, w)\} K_{ji}}{\sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji}}.
$$

As for the last term, we note that

$$
\hat{1}_{i,n,1} \sum_{j=1,j \neq i}^{n} 1 \{(D_i, W_i) = (1, w)\} K_{ji} / \sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji}
$$

$$
= \hat{1}_{i,n,1} \frac{\sum_{j=1,j \neq i}^{n} 1 \{(D_i, W_i) = (1, w)\} K_{ji}}{\sum_{j=1,j \neq i}^{n} L_{w,j} K_{ji} / \sum_{j=1,j \neq i}^{n} 1 \{W_i = w\} K_{ji}}
$$

$$
= \frac{q_1(V_i, w)}{q_1(V_i, w)p_1,w/q_1,w + q_0(V_i, w)p_0,w/q_0,w} + o_P(n^{-1/4}),
$$
uniformly over \( p \in A \), \((\text{using the fact that } O_p(\varepsilon_n) = o_P(n^{-1/4}) \text{ by Assumption 2(ii)})\). Hence the first term in \((6.29)\) is written as

\[
\frac{p_{1,w}}{q_{1,w}} - \frac{p_{1,w}}{q_{1,w}} \frac{q_1(V_i, w)}{q_1(V_i, w)p_{1,w}/q_{1,w} + q_0(V_i, w)p_{0_w}/q_{0,w}} + o_P(n^{-1/2})
\]

\[
= \left( \frac{q_1(V_i, w)p_{1,w}/q_{1,w}}{q_1(V_i, w)p_{1,w}/q_{1,w} + q_0(V_i, w)p_{0_w}/q_{0,w}} + o_P(n^{-1/2}) \right)
\]

\[
= \left( \frac{q_1(V_i, w)}{q_{1,w}} \right) p_1(V_i, w) + o_P(n^{-1/2}),
\]

where we used \((6.15)\) for the last equality.

Similarly, we find that

\[
\hat{1}_{n,i} \hat{1}_{n,i} \frac{p_1(V_i, w) \sum_{j=1}^n \left\{ \hat{L}_{0,w,j} - L_{0,w,j} \right\} K_{ji}}{L_{0,j} K_{ji}} = \left( \frac{q_1(V_i, w)}{q_{1,w}} \right) p_1(V_i, w) + o_P(n^{-1/2}),
\]

uniformly over \( p \in A \). Applying these results back to the last two sums in \((6.28)\), we conclude that

\[
\frac{1}{n} \sum_{i=1}^n S_i \hat{1}_{n,i} \{ \hat{p}_{1,i}(V_i, w) - \tilde{p}_{1,i}(V_i, w) \}
\]

\[
= \frac{1}{n} \sum_{i=1}^n S_i \hat{p}_0(V_i, w) p_1(V_i, w) \left( \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_P(n^{-1/2}),
\]

uniformly over \( p \in A \). Finally, we write the last sum as

\[
E_q \left[ p_0(V_i, w) p_1(V_i, w) S_i \right] \left( \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_P(n^{-1/2}),
\]

uniformly over \( p \in A \) and this completes the proof. ■

**Lemma B3**: (i) Suppose that Condition \((C1)\) and Assumptions 1-2 hold, and let \( \varepsilon_{d,w,i} = Y_{d,i} - \beta_d(V_i, w) \). Then,

\[
\frac{p_{1,w}}{q_{1,w}} n \sum_{i \in S_{1,w}} \hat{1}_{n,i} g(V_i, w) Y_i - \frac{p_{0,w}}{q_{0,w}} n \sum_{i \in S_{0,w}} \hat{1}_{n,i} g(V_i, w) Y_i
\]

\[
= \frac{1}{n} \sum_{i=1}^n g(V_i, w) L_{1,w,i} + \frac{1}{n} \sum_{i=1}^n g(V_i, w) L_{0,w,i} + o_P(n^{-1/2}),
\]

uniformly over \( p \in A \).
(ii) Suppose that Condition (C1) and Assumptions 1-2, and define \( \varepsilon_{d,w,i} \) as in (i). Then,

\[
\begin{align*}
&\frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} \hat{1}_{n,i} \frac{g(V_i, w)Y_i}{\hat{p}_{1,i}(V_i, w)} - \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \hat{1}_{n,i} \frac{g(V_i, w)Y_i}{\hat{p}_{0,i}(V_i, w)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)L_{1,w,i} \hat{1}_{n,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)L_{0,w,i} \hat{1}_{n,i}}{p_0(V_i, w)} \\
&+ \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w)\tau(V_i, w) - E_{1,w} [g(V_i, w)\tau(V_i, w)] \right\} L_{1,w,i} \\
&+ \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w)\tau(V_i, w) - E_{0,w} [g(V_i, w)\tau(V_i, w)] \right\} L_{0,w,i} \\
&+ E_{1,w} [g(V_i, w)\tau(V_i, w)] p_{1,w} + E_{0,w} [g(V_i, w)\tau(V_i, w)] p_{0,w} + o_P(n^{-1/2}),
\end{align*}
\]

uniformly over \( p \in A \).

**Proof:** (i) We first write

\[
\begin{align*}
&\frac{p_{1,w}}{q_{1,w}n} \sum_{i \in S_{1,w}} \hat{1}_{n,i} \frac{g(V_i, w)Y_i}{\hat{p}_{1,i}(V_i, w)} - \frac{p_{0,w}}{q_{0,w}n} \sum_{i \in S_{0,w}} \hat{1}_{n,i} \frac{g(V_i, w)Y_i}{\hat{p}_{0,i}(V_i, w)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{1}_{n,i} \frac{g(V_i, w)Y_i L_{1,w,i}}{\hat{p}_{1,i}(V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \hat{1}_{n,i} \frac{g(V_i, w)Y_i L_{0,w,i}}{\hat{p}_{0,i}(V_i, w)} = A_{1n} + A_{2n}.
\end{align*}
\]

We first write

\[
A_{1n} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i L_{1,w,i}}{p_1(V_i, w)} + \tilde{A}_{1n},
\]

where

\[
\tilde{A}_{1n} = \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i L_{1,w,i} \hat{1}_{n,i} \left( \frac{1}{\hat{p}_{1,i}(V_i, w)} - \frac{1}{p_1(V_i, w)} \right).
\]

As for \( \tilde{A}_{1n} \), note that

\[
\begin{align*}
&\frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i L_{1,w,i} \hat{1}_{n,i} \left( \frac{p_1(V_i, w) - \hat{p}_{1,i}(V_i, w)}{\hat{p}_{1,i}(V_i, w) p_1(V_i, w)} \right) \\
= &\frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i L_{1,w,i} \hat{1}_{n,i} \left( \frac{p_1(V_i, w) - \hat{p}_{1,i}(V_i, w)}{p_1^2(V_i, w)} \right) \\
&+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i L_{1,w,i} \hat{1}_{n,i} \frac{p_1(V_i, w) - \hat{p}_{1,i}(V_i, w)}{p_1(V_i, w)} \left( \frac{1}{\hat{p}_{1,i}(V_i, w)} - \frac{1}{p_1(V_i, w)} \right).
\end{align*}
\]
The supremum (over $p$) of the absolute value of the last sum has an upper bound with leading term

$$
\frac{1}{n} \sup_{p \in A} \sum_{i=1}^{n} |g(V_i, w)Y_i L_{1,w,i}^1 \hat{1}_{n,i} (p_1(V_i, w) - \hat{p}_{1,i}(V_i, w))^2 p_1(V_i, w)^3.
$$

On the other hand, observe that from (6.15), for any $q \in [1, \bar{a}] , \bar{a} \geq 4$,

$$
E_Q \left[ \sup_{p \in A} p_1^{-q} \right] = \sum_{(d,w) \in \{0,1\} \times W} E_{d,w} \left[ \sup_{p \in A} \left\{ \frac{f(V_i|1, w)p_{1,w}}{f(V_i|1, w)p_{1,w} + f(V_i|0, w)p_{0,w}} \right\}^{-q} \right] q_{d,w}
$$

$$
\leq \sum_{(d,w) \in \{0,1\} \times W} E_{d,w} \left[ \sup_{p \in A} \left\{ \frac{f(V_i|1, w)p_{1,w} + f(V_i|0, w)p_{0,w}}{f(V_i|1, w)p_{1,w}} \right\}^q \right] q_{d,w}.
$$

(6.31)

The last term is bounded due to Assumption 1(iv). Furthermore, observe that for some $C > 0$,

$$
\sup_{v \in \mathcal{V}(w)} E_Q \left[ \sup_{p \in A} |g(V_i, w)Y_i L_{1,w,i}^1| |(V_i, W_i) = (v, w)\right]
$$

$$
\leq C \sup_{v \in \mathcal{V}(w)} E_Q [Y_i^2 |(V_i, W_i) = (v, w)].
$$

The last term is bounded due to Assumption 1(ii). Hence by Lemma B1, we find that the sum in (6.30) is $o_P(n^{-1/2})$ (by the fact that $\epsilon_n^2 = o_P(n^{-1/2})$). We conclude that

$$
\hat{A}_{1n} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i L_{1,w,i}^1 \hat{1}_{n,i} (p_1(V_i, w) - \hat{p}_{1,i}(V_i, w))}{p_1^2(V_i, w)} + o_P(n^{-1/2}),
$$

(6.32)

uniformly over $p \in A$. Let $S_i = g(V_i, w)Y_i L_{1,w,i}^1 \hat{1}_{n,i}$. Then, for some $C > 0$,

$$
\sup_{v \in \mathcal{V}(w)} E_Q [S_i^2 |(V_i, W_i) = (v, w)] \leq C \sup_{v \in \mathcal{V}(w)} E_Q [Y_i^2 |(V_i, W_i) = (v, w)].
$$

The last term is bounded by Condition C2. As we saw in (6.31), the last term is bounded. We apply Lemma B2(i) to obtain that the leading sum in (6.32) is asymptotically equivalent to (up to $o_P(n^{-1/2})$)

$$
- \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)E_Q,w,i[Y_i L_{1,w,i}^1]J_{1,w,i}}{p_1^2(V_i, w)E_Q,w,i[L_{w,i}]} + \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)E_Q,w,i[Y_i L_{1,w,i}^1]J_{w,i}}{p_1(V_i, w)E_Q,w,i[L_{w,i}]},
$$

where $J_{1,w,i}$ and $J_{w,i}$ are as defined in Lemma B2. Using the fact that

$$
E_{Q,w,i}[Y_i L_{1,w,i}] = E[Y_i | V_i, (D_i, W_i) = (1, w)] q_1(V_i, w) p_{1,w}/q_{1,w}
$$

$$
= \beta_1(V_i, w) q_1(V_i, w) p_{1,w}/q_{1,w}.
$$
and \(q_1(V_i, w)p_{1,w}/\{E_{Q,w,i}[L_{w,i}]q_{1,w}\} = p_1(V_i, w)\) from (6.15), we write

\[
(6.34) \quad \frac{E_{Q,w,i}[Y_iL_{1,w,i}]}{E_{Q,w,i}[L_{w,i}]} = \beta_1(V_i, w)p_1(V_i, w), \quad \text{(using Condition C1.)}
\]

Using this, we write the first term in (6.33) as

\[
-\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) J_{1,w,i}
\]

and the second term as

\[
-\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) J_{0,w,i}
\]

Hence the difference in (6.33) is equal to

\[
-\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) J_{1,w,i} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) J_{0,w,i}.
\]

Therefore, we conclude that

\[
\tilde{A}_{1n} = -\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) p_0(V_i, w) \frac{J_{1,w,i}}{p_1(V_i, w)} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_1(V_i, w) J_{0,w,i} + o_P(n^{-1/2}).
\]

uniformly over \(p \in A\).

We turn to \(A_{2n}\), which can be written as

\[
A_{2n} = \frac{1}{n} \sum_{i=1}^{n} \tilde{1}_{n,i} g(V_i, w) Y_i L_{0,w,i}/p_0(V_i, w) + \tilde{A}_{2n} + o_P(n^{-1/2}),
\]

where

\[
\tilde{A}_{2n} = -\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \tilde{1}_{n,i} Y_i L_{0,w,i} \left( \frac{1}{\hat{p}_{0,i}(V_i, w)} - \frac{1}{p_0(V_i, w)} \right).
\]

Similarly as before, we write

\[
\tilde{A}_{2n} = \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_0(V_i, w) J_{1,w,i}
\]

\[
-\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \beta_0(V_i, w) p_1(V_i, w) J_{0,w,i} + o_P(n^{-1/2}),
\]
uniformly over \( p \in A \). Using the arguments employed to show (6.25) and combining the two results for \( \tilde{A}_{1n} \) and \( \tilde{A}_{2n} \), we deduce that

\[
\tilde{A}_{1n} - \tilde{A}_{2n} = -\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \frac{\beta_1(V_i, w)p_0(V_i, w)}{p_1(V_i, w)} + \beta_0(V_i, w) \right) J_{1,w,i} \\
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \beta_1(V_i, w) + \frac{\beta_0(V_i, w)p_1(V_i, w)}{p_0(V_i, w)} \right) J_{0,w,i} + o_P(n^{-1/2}) \\
= -\frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \beta_1(V_i, w) - \tau(V_i, w)p_1(V_i, w) \right) J_{1,w,i} \\
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \tau(V_i, w)p_0(V_i, w) + \beta_0(V_i, w) \right) J_{0,w,i} + o_P(n^{-1/2}),
\]

using the fact that \( \tau(X) = \beta_1(X) - \beta_0(X) \).

Therefore,

\[
\frac{p_{1,w}}{q_{1,w}n} \sum_{i \in S_{1,w}} \hat{1}_{i,w} g(V_i, w)Y_i \\
- \frac{p_{0,w}}{q_{0,w}n} \sum_{i \in S_{0,w}} \hat{1}_{i,w} g(V_i, w)Y_i \\
= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)L_{1,w,i} - \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)L_{0,w,i} \\
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\beta_1(V_i, w) - \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\beta_0(V_i, w) \\
- \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \frac{\beta_1(V_i, w) - \tau(V_i, w)p_1(V_i, w)}{p_1(V_i, w)} \right) J_{1,w,i} \\
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \frac{\tau(V_i, w)p_0(V_i, w) + \beta_0(V_i, w)}{p_0(V_i, w)} \right) J_{0,w,i} + o_P(n^{-1/2}).
\]
By rearranging the terms, we rewrite
\[
\frac{p_{1,w}}{q_{1,w}} \sum_{i \in S_{1,w}} \hat{I}_{n,i} g(V_i, w) Y_i = \frac{p_{0,w}}{q_{0,w}} \sum_{i \in S_{0,w}} \hat{I}_{n,i} g(V_i, w) Y_i
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) L_{1,w,i} \hat{L}_{1,1} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) L_{1,1} \hat{L}_{0,1}
\]
\[
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \tau(V_i, w) L_{1,1} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \tau(V_i, w) L_{0,1}
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \beta_1(V_i, w) - \tau(V_i, w) \right) \frac{p_1(V_i, w)}{p_1(V_i, w)} \left( \beta_0(V_i, w) + \tau(V_i, w) \right) \frac{q_1(V_i, w)}{q_1(V_i, w)}
\]
\[
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left( \beta_0(V_i, w) + \tau(V_i, w) \right) \frac{q_0(V_i, w)}{q_0(V_i, w)}.
\]

However, by Bayes' rule (see (6.14)),
\[
(6.35) \quad \frac{p_{1,w} q_1(V_i, w)}{q_{1,w}} = \frac{p_{1,w} f_1(V_i, w) f_Q(V_i, w)}{f_Q(V_i, w)} = \frac{p_{1,w} f(V_i, w)}{f_Q(V_i, w)}.
\]

Therefore,
\[
H_{n,i} = \frac{f_P(V_i, w)}{f_Q(V_i, w)} \left\{ \frac{\beta_1(V_i, w) - \tau(V_i, w)}{p_1(V_i, w)} \right\} p_1(V_i, w) - \frac{\beta_0(V_i, w) + \tau(V_i, w)}{p_0(V_i, w)} p_0(V_i, w)
\]

from which it follows that $H_{n,i} = 0$ by the definition of $\tau(V_i, w)$. Hence we obtain the wanted result.

(ii) We write
\[
(6.36) \quad \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} \hat{L}_{1,1} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} \hat{L}_{0,1}
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} L_{1,1} - \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} L_{0,1}
\]
\[
+ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} \{ \hat{L}_{1,1} - L_{1,1} \} - \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) Y_i \tilde{I}_{n,i} \{ \hat{L}_{0,1} - L_{0,1} \}.
\]
We write the first difference as
\[
\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i \hat{1}_{n,i} L_{1,w,i}}{\hat{p}_1(i, V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i \hat{1}_{n,i} L_{0,w,i}}{\hat{p}_0(i, V_i, w)} \right\} + \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i \hat{1}_{n,i} L_{1,w,i}}{p_1^2(V_i, w)} A_i - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)Y_i \hat{1}_{n,i} L_{0,w,i}}{p_0^2(V_i, w)} B_i \right\} + O_P(n^{-1/2})
\]
\[
= J_{1n} + J_{2n} + O_P(n^{-1/2}),
\]
uniformly over \( p \in A \), where
\[
A_i = \hat{p}_1(i, V_i, w) - \hat{p}_1(i, V_i, w), \quad \text{and} \quad B_i = \hat{p}_0(i, V_i, w) - \hat{p}_0(i, V_i, w).
\]
Note that the normalized sums with trimming factor \( \hat{1}_{n,i} \) can be replaced by the same sums but with \( \hat{1}_{n,i} \) (with the resulting discrepancy confined to \( O_P(n^{-1/2}) \), uniformly for \( p \in A \), because
\[
1 - \hat{1}_{n,i} = \text{op}(n^{-1/2}), \quad \text{and} \quad
1 - \hat{1}_{n,i} = \text{op}(n^{-1/2}),
\]
uniformly over \( p \in A \). The first line was shown in the proof of Lemma B2. (See arguments below (6.21).) Similar arguments apply to the second line so that
\[
1 - \hat{1}_{n,i} \leq 1 \left\{ \hat{\lambda}_{1,i}(V_i, w) < \delta_n \right\} + 1 \left\{ \hat{\lambda}_0,i(V_i, w) < \delta_n \right\}.
\]
We write the first indicator on the right hand side as
\[
1 \left\{ \frac{\hat{E}_{Q,w,i}[\tilde{L}_{1,w,i}]}{n-1} \sum_{j=1,j\neq i}^{n} 1\{W_i = w\} K_{h,ji} < \delta_n \right\} \leq 1 \left\{ \hat{E}_{Q,w,i}[L_{1,w,i}] < \kappa_{2n} \right\},
\]
where \( \kappa_{2n} = (\delta_n + R_{1n} + R_{2n})/c \) (with \( c > 0 \) such that \( \min_{w \in W} \inf_{v \in V(w)} f_Q(v, w) > c \) (see Assumption 1(iii)), \( R_{1n} \) is as defined in (6.23) and
\[
R_{2n} \equiv \max_{1 \leq i \leq n} \left| \frac{\hat{E}_{Q,w,i}[L_{1,w,i}]}{n-1} - \frac{\hat{E}_{Q,w,i}[\tilde{L}_{1,w,i}]}{n-1} \right| \sum_{j=1,j\neq i}^{n} 1\{W_i = w\} K_{ji} \leq \frac{p_{d,w} - \hat{p}_{d,w}}{q_{d,w}} \cdot \max_{1 \leq i \leq n} \left| \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} 1\{W_i = w\} K_{ji} \right| = O_P(\varepsilon_n).
\]
Recall that \( R_{1n} = O_P(\varepsilon_n) \). Thus as before, we can take a nonstochastic sequence \( \kappa_{2n}' \) and \( \eta > 0 \) such that \( \kappa_{2n}' = o(n^{-1/2}) \) and \( \max\{\gamma,2\} \leq \eta \leq \bar{\alpha} \), using Assumptions 2(ii) and (iii).
Replacing $\kappa_{2n}$ in (6.38) by this $\kappa'_{2n}$, we find that with probability approaching one,

$$|1 - \hat{I}_{n,i}| \leq 1 \{ \mathbf{E}_{Q,w,i}[L_{1,w,i}] \leq \kappa'_{2n} \} + 1 \{ \mathbf{E}_{Q,w,i}[L_{0,w,i}] \leq \kappa'_{2n} \}$$

Note that the expectation $\mathbf{E}_Q$ of the last term is bounded by (for some $C > 0$)

$$C \mathbf{E}_Q \left[ \mathbf{E}_{Q,w,i}[L_{w,i}] \right] = O \left( \kappa'_{2n} \right) = o(n^{-1/2}),$$

uniformly over $p \in A$. Thus we obtain the second convergence in (6.37).

As for $J_{2n}$, by applying Lemma B2(ii), we have

$$J_{2n} = \mathbf{E}_Q \left[ \frac{g(V_i, w)Y_iL_{1,w,i}p_0(V_i, w)}{p_1(V_i, w)} \right] \left( \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_P(n^{-1/2})$$

$$= \mathbf{E}_Q \left[ g(V_i, w)Y_i \left\{ \frac{L_{1,w,i}p_0(V_i, w)}{p_1(V_i, w)} + \frac{L_{0,w,i}p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}}$$

$$+ o_P(n^{-1/2}),$$

uniformly for $p \in A$. On the other hand, as for the last difference in (6.36), we have

$$\frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i \hat{I}_{n,i} \left\{ \hat{L}_{1,w,i} - L_{1,w,i} \right\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i \hat{L}_{1,w,i} - L_{1,w,i} + o_P(n^{-1/2})$$

$$= - \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_iL_{1,w,i} \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} + o_P(n^{-1/2})$$

$$= - \mathbf{E}_Q \left[ \frac{g(V_i, w)Y_iL_{1,w,i}}{p_1(V_i, w)} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} + o_P(n^{-1/2}),$$

uniformly for $p \in A$. Here uniformity again follows from the fact that $p_{1,w}$ and $p_{0,w}$ can be factored out from the converging random sequence. In particular,

$$\frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_iL_{1,w,i}$$

$$= \frac{p_{1,w}}{q_{1,w}} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i\mathcal{I}_{1,w,i} + \frac{p_{0,w}}{q_{1,w}} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)Y_i\mathcal{I}_{1,w,i} \frac{f(V_i|0, w)}{f(V_i|1, w)},$$
where \( I_{1,w} \equiv 1\{ (D_i, W_i) = (1, w) \} \). The CLT can be applied to terms that do not depend on \( p \). Similarly,

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) Y_i \tilde{I}_{n,i} \{ \hat{L}_{0,w,i} - L_{0,w,i} \}}{\hat{p}_{0,i}(V_i, w)} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) Y_i \{ \hat{L}_{0,w,i} - L_{0,w,i} \}}{p_0(V_i, w)} + o_P(n^{-1/2})
\]

uniformly for \( p \in A \). Combining these results, we conclude that

\[
(6.39) \quad \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) Y_i \tilde{I}_{n,i} \{ \hat{L}_{1,w,i} - L_{1,w,i} \}}{\hat{p}_{1,i}(V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) Y_i \tilde{I}_{n,i} \{ L_{0,w,i} - \hat{L}_{0,w,i} \}}{\hat{p}_{0,i}(V_i, w)}
\]

\[
+ \mathbb{E}_Q \left[ g(V_i, w) Y_i \left\{ -L_{1,w,i} + \frac{L_{0,w,i} p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}}
\]

\[
- \mathbb{E}_Q \left[ g(V_i, w) Y_i \left\{ \frac{L_{1,w,i} p_0(V_i, w)}{p_1(V_i, w)} - L_{0,w,i} \right\} \right] \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} + o_P(n^{-1/2}).
\]

uniformly for \( p \in A \). The last difference is written as

\[
\mathbb{E}_Q \left[ g(V_i, w) \left\{ -Y_{i} L_{1,w,i} + Y_{0,i} \frac{L_{0,w,i} p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}}
\]

\[
- \mathbb{E}_Q \left[ g(V_i, w) \left\{ Y_{i} L_{1,w,i} \frac{p_0(V_i, w)}{p_1(V_i, w)} - Y_{0,i} L_{0,w,i} \right\} \right] \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}}
\]

\[
= \mathbb{E}_Q \left[ g(V_i, w) \left\{ -Y_{i} - Y_{0,i} \right\} L_{1,w,i} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}}
\]

\[
+ \mathbb{E}_Q \left[ g(V_i, w) Y_{0,i} \left\{ \frac{L_{0,w,i} p_1(V_i, w)}{p_0(V_i, w)} - L_{1,w,i} \right\} \right] \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}}
\]

\[
- \mathbb{E}_Q \left[ g(V_i, w) \left\{ Y_{i} - Y_{0,i} \right\} L_{0,w,i} \right] \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}}
\]

\[
+ \mathbb{E}_Q \left[ g(V_i, w) Y_{i} \left\{ \frac{L_{0,w,i} p_0(V_i, w)}{p_1(V_i, w)} - L_{1,w,i} \right\} \right] \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}}.
\]
The second and the fourth expectations vanish because
\[
\mathbb{E}_Q \left[ g(V_i, w) Y_{0,i} \left\{ -L_{1,w,i} + \frac{L_{0,w,i} p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] = \mathbb{E} \left[ g(V_i, w) \beta_0(V_i, w) \left\{ -1\{(D_i, W_i) = (1, w)\} + \frac{1\{(D_i, W_i) = (0, w)\} p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] = 0,
\]
and similarly,
\[
\mathbb{E}_Q \left[ g(V_i, w) y_i \left\{ L_{0,w,i} - \frac{L_{1,w,i} p_0(V_i, w)}{p_1(V_i, w)} \right\} \right] = \mathbb{E} \left[ g(V_i, w) \beta_1(V_i, w) \left\{ p_0(V_i, w) - p_0(V_i, w) \right\} \right] = 0.
\]
Furthermore, observe that
\[
\mathbb{E}_Q \left[ g(V_i, w) \{ -\{Y_{1,i} - Y_{0,i}\} L_{1,w,i}\} \right] = -\mathbb{E} \left[ g(V_i, w)\{Y_{1,i} - Y_{0,i}\}\{1\{(D_i, W_i) = (1, w)\}\} \right] = -\mathbb{E} \left[ g(V_i, w)\{\beta_1(V_i, w) - \beta_0(V_i, w)\} p_1(V_i, w) \right] = -\mathbb{E} \left[ g(V_i, w)\tau(V_i, w) p_1(V_i, w) \right],
\]
and similarly,
\[
-\mathbb{E}_Q \left[ g(V_i, w)\{Y_{1,i} - Y_{0,i}\} L_{0,w,i}\right] = -\mathbb{E} \left[ g(V_i, w)\tau(V_i, w) p_0(V_i, w) \right].
\]
Hence, as for the last two terms in (6.39), we find that
\[
\mathbb{E}_Q \left[ g(V_i, w) Y_i \left\{ -L_{1,w,i} + \frac{L_{0,w,i} p_1(V_i, w)}{p_0(V_i, w)} \right\} \right] \frac{q_{1,w} - q_{1,w}}{q_{1,w}} = -\mathbb{E} \left[ g(V_i, w)\tau(V_i, w) p_0(V_i, w) \right].
\]
and

\[- \mathbb{E}_Q \left[ g(V_i, w) Y_i \left\{ \frac{L_{1,w,i} p_0(V_i, w)}{p_1(V_i, w)} - L_{0,w,i} \right\} \right] \tilde{q}_{0,w} - q_{0,w} \]

\[= - \mathbb{E}_Q [g(V_i, w) \tau(V_i, w) L_{0,w,i} \tilde{q}_{0,w} - q_{0,w}] \]

\[= - p_{0,w} \mathbb{E}_{0,w} [g(V_i, w) \tau(V_i, w)] \frac{\tilde{q}_{0,w} - q_{0,w}}{q_{0,w}} \]

\[= - \mathbb{E}_{0,w} [g(V_i, w) \tau(V_i, w)] \frac{1}{n} \sum_{i=1}^{n} (L_{0,w,i} - p_{0,w}). \]

Applying the result of (i) of this lemma to the first difference of (6.39), we conclude that the difference in (ii) in this lemma is equal to

\[\frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) L_{1,w,i} \varepsilon_{1,w,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w) L_{0,w,i} \varepsilon_{0,w,i}}{p_0(V_i, w)} + \Gamma_{n,w} + o_P(n^{-1/2}), \]

uniformly for \( p \in A \), where

\[\Gamma_{n,w} \equiv \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \tau(V_i, w) L_{1,w,i} + \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \tau(V_i, w) L_{0,w,i} \]

\[ - \mathbb{E}_{1,w} [g(V_i, w) \tau(V_i, w)] \frac{1}{n} \sum_{i=1}^{n} (L_{1,w,i} - p_{1,w}) \]

\[ - \mathbb{E}_{0,w} [g(V_i, w) \tau(V_i, w)] \frac{1}{n} \sum_{i=1}^{n} (L_{0,w,i} - p_{0,w}). \]

The proof is complete because

\[\Gamma_{n,w} = \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w) \tau(V_i, w) - \mathbb{E}_{1,w} [g(V_i, w) \tau(V_i, w)] \right\} L_{1,w,i} \]

\[+ \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w) \tau(V_i, w) - \mathbb{E}_{0,w} [g(V_i, w) \tau(V_i, w)] \right\} L_{0,w,i} \]

\[+ \mathbb{E}_{1,w} [g(V_i, w) \tau(V_i, w)] p_{1,w} + \mathbb{E}_{0,w} [g(V_i, w) \tau(V_i, w)] p_{0,w}. \]

\[\square\]

**Proof of Lemma A1:** Let us consider the first statement in (6.8). We write \( \tilde{\tau}_{ate}(p) - \tau_{ate}(p) \) as

\[(6.40) \quad \frac{1}{\mathbb{E}g(X_i)} \sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} \tilde{R}_{n,i} \frac{g(V_i, w) Y_i}{p_{1,i}(V_i, w)} - \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{R}_{n,i} \frac{g(V_i, w) Y_i}{p_{0,i}(V_i, w)} \right\} + \tilde{R}_n - \tau_{ate}(p), \]

where \( \tilde{R}_n \) and \( \tilde{R}_{n,i} \) are the left-hand side of (6.39) and (6.40).
where

\[(6.41) \quad \hat{E}_d g(X_i) = \sum_{w \in W} \frac{p_d, w}{n_{d, w}} \sum_{i \in S_{d, w}} \tilde{1}_{n, i} \frac{g(V, w)}{\tilde{p}_{d, i}(V, w)}, \]

\[
\tau_d(p) = \frac{E [g(X_i) \beta_d(X_i)]}{E g(X_i)},
\]

and

\[
\tilde{R}_n \equiv \left\{ \frac{1}{E_1 g(X_i)} - \frac{1}{E g(X_i)} \right\} \sum_{w \in W} \frac{p_{1, w}}{n_{1, w}} \sum_{i \in S_{1, w}} \tilde{1}_{n, i} \frac{g(V, w)Y_i}{\tilde{p}_{1, i}(V, w)} - \left\{ \frac{1}{E_0 g(X_i)} - \frac{1}{E g(X_i)} \right\} \sum_{w \in W} \frac{p_{0, w}}{n_{0, w}} \sum_{i \in S_{0, w}} \tilde{1}_{n, i} \frac{g(V, w)Y_i}{\tilde{p}_{0, i}(V, w)}.
\]

It is not hard to show that

\[(6.42) \quad \sum_{w \in W} \frac{p_{1, w}}{n_{1, w}} \sum_{i \in S_{1, w}} \tilde{1}_{n, i} \frac{g(V, w)Y_i}{\tilde{p}_{1, i}(V, w)} = \tau_1(p) E g(X_i) + O_P(\varepsilon_n) \quad \text{and} \quad \sum_{w \in W} \frac{p_{0, w}}{n_{0, w}} \sum_{i \in S_{0, w}} \tilde{1}_{n, i} \frac{g(V, w)Y_i}{\tilde{p}_{0, i}(V, w)} = \tau_0(p) E g(X_i) + O_P(\varepsilon_n),\]

uniformly for \( p \in A \). This is because

\[
\frac{1}{n_{1, w}} \sum_{i \in S_{1, w}} \tilde{1}_{n, i} \frac{g(V, w)Y_i}{\tilde{p}_{1, i}(V, w)} = \frac{1}{n_{1, w}} \sum_{i \in S_{1, w}} g(V, w)Y_i + \frac{1}{n_{1, w}} \sum_{i \in S_{1, w}} g(V, w)Y_i \tilde{1}_{n, i} - 1 \frac{1}{\tilde{p}_{1, i}(V, w)} \tilde{1}_{n, i}.
\]

Note that the second term is \( o_P(n^{-1/2}) \) (see (6.37)) and the third term \( O_P(\varepsilon_n) \) (by Lemma B1). As for the first term, we have

\[
\frac{1}{n_{1, w}} \sum_{i \in S_{1, w}} g(V, w)Y_i = E_{1, w} \left[ \frac{g(V, w)\beta_1(V, w)}{p_1(V, w)} \right] + O_P(n^{-1/2}),
\]

uniformly for \( p \in A \). Furthermore,

\[
\sum_{w \in W} p_{1, w} E_{1, w} \left[ \frac{g(V, w)\beta_1(V, w)}{p_1(V, w)} \right] = E \left[ \frac{g(V, W_i)\beta_1(V, W_i)}{p_1(V, W_i)} \right] p_1 = E \left[ \frac{g(V, W_i)\beta_1(V, W_i)D_i}{p_1(V, W_i)} \right].
\]

In view of the definition of \( \tau_d \) in (6.41), the last expectation is equal to \( \tau_1(p) E g(X_i) \). Collecting above results leads to (6.42).
Apply the similar argument to \( \tilde{E}_d g(X_i) \), we have
\[
\tilde{E}_d g(X_i) = \sum_{w} \frac{p_{d, w}}{n_{d, w}} \sum_{i \in S_{d, w}} \tilde{I}_{n, i} \cdot \frac{g(V_i, w)}{\tilde{p}_{d, i}(V_i, w)} = E g(X_i) + O_P(\varepsilon_n).
\]
Hence
\[
\hat{R}_a = \left\{ \frac{1}{E_1 g(X_i)} - \frac{1}{E g(X_i)} \right\} \tau_1(p) E g(X_i) - \frac{1}{E_0 g(X_i)} \tau_0(p) E g(X_i) + o_P(n^{-1/2})
\]
\[
= - \frac{\tilde{E}_1 g(X_i) - E g(X_i)}{E g(X_i)} \tau_1(p) + \frac{\tilde{E}_0 g(X_i) - E g(X_i)}{E g(X_i)} \tau_0(p) + o_P(n^{-1/2})
\]
uniformly for \( p \in A \), where \( o_P(n^{-1/2}) \) comes from \( O_P(\varepsilon_n^2) = o_P(n^{-1/2}) \). The last equality follows because \( \tau_{ate}(p) = \tau_1(p) - \tau_0(p) \). Observe that
\[
(6.43) \quad - \frac{\tilde{E}_1 g(X_i) \tau_1(p) - \tilde{E}_0 g(X_i) \tau_0(p)}{E g(X_i)}
\]
\[
= - \frac{1}{E g(X_i)} \left\{ \sum_{w \in W} \left\{ \frac{p_{1, w}}{n_{1, w}} \sum_{i \in S_{1, w}} \tilde{I}_{n, i} \cdot \frac{g(V_i, w) \tau_1(p)}{\tilde{p}_{1, i}(V_i, w)} - \frac{p_{0, w}}{n_{0, w}} \sum_{i \in S_{0, w}} \tilde{I}_{n, i} \cdot \frac{g(V_i, w) \tau_0(p)}{\tilde{p}_{0, i}(V_i, w)} \right\} \right\}.
\]
By replacing \( Y_i \{ (D_i, W_i) = (1, w) \} \) by \( \tau_1 \{ (D_i, W_i) = (1, w) \} \) and \( Y_i \{ (D_i, W_i) = (0, w) \} \) by \( \tau_0 \{ (D_i, W_i) = (0, w) \} \) in Lemma B3(ii) and noting that \( \tau_{ate}(p) = \tau_1(p) - \tau_0(p) \), we find that the last term in (6.43) is equal to
\[
- \frac{\tau_{ate}(p)}{E g(X_i)} \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w) - E_{1, w} \left[ g(V_i, w) \right] \right\} L_{1, w, i}
\]
\[
- \frac{\tau_{ate}(p)}{E g(X_i)} \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w) - E_{0, w} \left[ g(V_i, w) \right] \right\} L_{0, w, i}
\]
\[
- \frac{\tau_1(p) - \tau_0(p)}{E g(X_i)} \sum_{w \in W} \{ E_{1, w}[g(V_i, w)]p_{1, w} + E_{0, w}[g(V_i, w)]p_{0, w} \} + o_P(n^{-1/2}),
\]
uniformly for \( p \in A \). Observe that for the last term,
\[
\frac{\tau_1(p) - \tau_0(p)}{E g(X_i)} \sum_{w \in W} \{ E_{1, w}[g(V_i, w)]p_{1, w} + E_{0, w}[g(V_i, w)]p_{0, w} \} = \tau_{ate}(p).
\]
Therefore, by applying Lemma B3(ii) to the leading term of (6.40), and recalling the definitions in (6.6), we conclude that \( \tilde{\tau}_{ate}^{SA}(p) - \tau_{ate}(p) \) is asymptotically equivalent to (up to
\[ \alpha_p(n^{-1/2}) \text{ uniformly over all } p \in A \]

\[
\frac{1}{n} \mathbb{E}_g(X_i) \sum_{w \in W} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)L_{1,w,i \in 1,w,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i, w)L_{0,w,i \in 0,w,i}}{p_0,w(V_i, w)} \right\}
\]

\[
+ \frac{1}{n} \mathbb{E}_g(X_i) \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} (\xi_{1,ate}(V_i, w)L_{1,w,i} + \xi_{0,ate}(V_i, w)L_{0,w,i})
\]

\[
+ \frac{1}{n} \mathbb{E}_g(X_i) \sum_{w \in W} \{ \mathbb{E}_{1,w}[g(V_i, w)\tau(V_i, w)]p_{1,w} + \mathbb{E}_{0,w}[g(V_i, w)\tau(V_i, w)]p_{0,w} \} - \tau_{ate}(p).
\]

The second to the last term is actually \( \tau_{ate}(p) \) cancelling the last \( \tau_{ate}(p) \). This gives the first statement of Lemma A1.

Now, we prove the second statement in (6.8). Let

\[ \mathbb{E}_1 [\beta_0(X_i)] = \mathbb{E} [\beta_0(X_i)|D_i = 1] \]

and write \( \tilde{\tau}_{tet}(p) - \tau_{tet}(p) \) as

\[
\frac{1}{p_1 \mathbb{E}_1 [g(X_i)]} \sum_{w \in W} \left\{ \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w)Y_i - \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{I}_{n,i}g(V_i, w)\tilde{p}_{1,i}(V_i, w)Y_i \right\}
\]

(6.44) \[ + \tilde{R}_n - \tau_{tet}(p), \]

where

\[
\tilde{R}_n \equiv M_{1n} \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w)Y_i - M_{2n} \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{I}_{n,i}g(V_i, w)\tilde{p}_{1,i}(V_i, w)Y_i,
\]

with

\[
M_{1n} = \left( \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) \right)^{-1} - \frac{1}{p_1 \mathbb{E}_1 [g(X_i)]}, \quad \text{and}
\]

\[
M_{2n} = \left( \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{I}_{n,i}g(V_i, w)\tilde{p}_{1,i}(V_i, w)/\tilde{p}_{0,i}(V_i, w) \right)^{-1} - \frac{1}{p_1 \mathbb{E}_1 [g(X_i)]}.
\]

Note that

\[
\sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) = \sum_{w \in W} p_{1,w} \mathbb{E}_1 [g(X_i)] + O_p(n^{-1/2}) = \mathbb{E}_1 [g(X_i)] p_1 + O_p(n^{-1/2}).
\]
This holds uniformly over \( p \in A \). And
\[
\sum_{w \in \mathcal{W}} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) Y_i = \sum_{w \in \mathcal{W}} p_{1,w} \mathbf{E}_{1,w} [g(X_i) Y_i] + O_P(n^{-1/2})
\]
\[
= \sum_{w \in \mathcal{W}} p_{1,w} \mathbf{E}_{1,w} [g(X_i) \beta_1(X_i)] + O_P(n^{-1/2})
\]
\[
= \mathbf{E}_1 [g(X_i) \beta_1(X_i)] p_1 + O_P(n^{-1/2}).
\]
This holds uniformly for all \( p \in A \). In addition,
\[
\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{I}_{n,i} g(V_i, w) \frac{\hat{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)}
\]
\[
= \sum_{w \in \mathcal{W}} p_{0,w} \mathbf{E}_{0,w} \left[ g(V_i, w) \frac{p_1(V_i, w)}{p_0(V_i, w)} \right] + O_P(\varepsilon_n)
\]
\[
= \mathbf{E}_0 \left[ g(X_i) \frac{p_1(X_i)}{p_0(X_i)} \right] p_0 + O_P(\varepsilon_n)
\]
\[
= \mathbf{E} \left[ g(X_i) \frac{p_1(X_i)(1 - D_i)}{p_0(X_i)} \right] + O_P(\varepsilon_n)
\]
\[
= \mathbf{E} \left[ g(X_i) p_1(X_i) \right] + O_P(\varepsilon_n) = \mathbf{E}_1 \left[ g(X_i) \right] p_1 + O_P(\varepsilon_n),
\]
uniformly for all \( p \in A \). The convergence rate follows from similar arguments in deriving (6.42). The uniformity comes from the fact that
\[
\frac{1}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \frac{p_1(V_i, w)}{p_0(V_i, w)} = \mathbf{E}_{0,w} \left[ g(V_i, w) \frac{p_1(V_i, w)}{p_0(V_i, w)} \right] + O_P(n^{-1/2}),
\]
uniformly for \( p \in A \). Also,
\[
\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{I}_{n,i} g(V_i, w) \frac{\hat{p}_{1,i}(V_i, w) Y_i}{\tilde{p}_{0,i}(V_i, w)}
\]
\[
= \sum_{w \in \mathcal{W}} p_{0,w} \mathbf{E}_{0,w} \left[ g(V_i, w) \frac{p_1(V_i, w) Y_i}{p_0(V_i, w)} \right] + O_P(\varepsilon_n)
\]
\[
= \sum_{w \in \mathcal{W}} p_{0,w} \mathbf{E}_{0,w} \left[ g(V_i, w) \frac{p_1(V_i, w) \beta_0(V_i, w)}{p_0(V_i, w)} \right] + O_P(\varepsilon_n).
\]
\[
= \mathbf{E}_0 \left[ g(X_i) \frac{p_1(X_i) \beta_0(X_i)}{p_0(X_i)} \right] p_0 + O_P(\varepsilon_n),
\]
uniformly for all \( p \in A \). We can rewrite the leading term as
\[
\mathbf{E} \left[ g(X_i) \frac{p_1(X_i) \beta_0(X_i)(1 - D_i)}{p_0(X_i)} \right] = \mathbf{E} \left[ g(X_i) p_1(X_i) \beta_0(X_i) \right] = \mathbf{E}_1 \left[ g(X_i) \beta_0(X_i) \right] p_1.
Hence we can write \( \hat{R}_n \) as (up to \( o_P(n^{-1/2}) \) uniformly over \( p \in A \))

\[
\frac{1}{p_1 E_1^2 [g(X_i)]} \left\{ p_1 E_1 [g(X_i)] - \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) \right\} E_1 [g(X_i)\beta_0(X_i)] \\
- \frac{1}{p_1 E_1^2 [g(X_i)]} \left\{ p_1 E_1 [g(X_i)] - \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{t}_{ni} g(V_i, w) \frac{\tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} \right\} E_1 [g(X_i)\beta_0(X_i)].
\]

(Note that \( o_P(n^{-1/2}) \) comes from \( O_P(\varepsilon_n^2) = o_P(n^{-1/2}) \). We rewrite the difference as

\[
\tau_{tet}(p) - \frac{1}{p_1 E_1^2 [g(X_i)]} \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) E_1 [g(X_i)\beta_1(X_i)] \\
+ \frac{1}{p_1 E_1^2 [g(X_i)]} \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{t}_{ni} \frac{\tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} E_1 [g(X_i)\beta_0(X_i)].
\]

Plugging this result into (6.44) and defining

\[
\tilde{\varepsilon}_{d,i} = Y_{di} - \frac{E_1 [g(X_i)\beta_d(X_i)]}{E_1 [g(X_i)]},
\]

we write \( \tilde{\tau}_{tet}(p) - \tau_{tet}(p) \) as (up to \( o_P(n^{-1/2}) \) uniformly over \( p \in A \))

\[
\frac{1}{p_1 E_1 [g(X_i)]} \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) \tilde{\varepsilon}_{1,i} \\
- \frac{1}{p_1 E_1 [g(X_i)]} \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \frac{\tilde{t}_{ni} g(V_i, w) \tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} \tilde{\varepsilon}_{0,i}
\]

(6.45)

\[
= \frac{1}{p_1 E_1 [g(X_i)]} (B_n - C_n - D_n),
\]

where

\[
B_n = \sum_{w \in W} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) \tilde{\varepsilon}_{1,i} - \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \frac{p_1(V_i, w) \tilde{\varepsilon}_{0,i}}{p_0(V_i, w)},
\]

\[
C_n = \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{\varepsilon}_{0,i} \left\{ \frac{\tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} - \frac{\tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} \right\},
\]

and

\[
D_n = \sum_{w \in W} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{\varepsilon}_{0,i} \left\{ \frac{\tilde{p}_{1,i}(V_i, w)}{\tilde{p}_{0,i}(V_i, w)} - \frac{p_1(V_i, w)}{p_0(V_i, w)} \right\}.
\]
We consider $D_n$ first. Write it as (up to $o_P(n^{-1/2})$ uniformly over $p \in A$)

$$\sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{\varepsilon}_{0,i} \left( \frac{\hat{p}_{0,i}(V_i, w)p_0(V_i, w) - p_1(V_i, w)\hat{p}_{0,i}(V_i, w)}{p_0^2(V_i, w)} \right) \right\}$$

$$= \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w}}{q_{0,w} n} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{\varepsilon}_{0,i} \left( \frac{\hat{p}_{1,i}(V_i, w) - p_1(V_i, w)}{p_0(V_i, w)} \right) \right\}$$

$$+ \sum_{w \in \mathcal{W}} \left\{ \frac{p_{0,w}}{q_{0,w} n} \sum_{i \in S_{0,w}} g(V_i, w) \tilde{\varepsilon}_{0,i} \left( \frac{p_1(V_i, w)\{p_0(V_i, w) - \hat{p}_{0,i}(V_i, w)}{p_0^2(V_i, w)} \right) \right\}$$

$$= D_{1n} + D_{2n}.$$ 

The uniformity comes from the fact that the convergence rate of $\hat{p}_{0,i}(V_i, w)$ to $p_0(V_i, w)$ is uniform for $p$. Apply Lemma B2(i) to write $D_{1n}$ as (up to $o_P(n^{-1/2})$ uniformly for all $p \in A$.)

$$\sum_{w \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \mathbb{E}_{Q_{w,i}} \left[ \tilde{\varepsilon}_{0,i} L_{0,w,i} \right] \mathcal{J}_{i,w} \right\}$$

$$- \sum_{w \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \mathbb{E}_{Q_{w,i}} \left[ \tilde{\varepsilon}_{0,i} L_{0,w,i} \right] p_1(V_i, w) \mathcal{J}_{w,i} \right\}.$$ 

Defining

$$\Delta_{d,w,i} \equiv \beta_d(V_i, w) - \frac{\mathbb{E}_1 \left[ g(X_i) \beta_d(X_i) \right]}{\mathbb{E}_1 \left[ g(X_i) \right]},$$

we write the last difference as

$$\sum_{w \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0,w,i} \mathcal{J}_{i,w} \right\} - \sum_{w \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)p_1(V_i, w) \Delta_{0,w,i} \mathcal{J}_{w,i} \right\},$$

because (using (6.15))

$$\frac{\mathbb{E}_{Q_{w,i}} \left[ \tilde{\varepsilon}_{0,i} L_{0,w,i} \right]}{\mathbb{E}_{Q_{w,i}} \left[ L_{w,i} \right]} = p_0(V_i, w) \left\{ \beta_0(V_i, w) - \frac{\mathbb{E}_1 \left[ g(X_i) \beta_0(X_i) \right]}{\mathbb{E}_1 \left[ g(X_i) \right]} \right\} = p_0(V_i, w) \Delta_{0,w,i},$$

and

$$\frac{\mathbb{E}_{Q_{w,i}} \left[ \tilde{\varepsilon}_{0,i} L_{0,w,i} \right] p_1(V_i, w)}{\mathbb{E}_{Q_{w,i}} \left[ L_{w,i} \right] p_0(V_i, w)}$$

$$= p_1(V_i, w) \left\{ \beta_0(V_i, w) - \frac{\mathbb{E}_1 \left[ g(X_i) \beta_0(X_i) \right]}{\mathbb{E}_1 \left[ g(X_i) \right]} \right\} = p_1(V_i, w) \Delta_{0,w,i}.$$
Applying Lemma B2(i), we write $D_{2n}$ as (up to $o_P(n^{-1/2})$ uniformly for all $p \in A$)

$$
- \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) E_{Q, w, i} [g(V_i, w) \xi_{0, i} L_{0, w, i}]}{p_0^2(V_i, w) E_{Q, w, i} [L_{w, i}]} J_{0, w, i}
+ \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) E_{Q, w, i} [g(V_i, w) \xi_{0, i} L_{0, w, i}]}{p_0(V_i, w) E_{Q, w, i} [L_{w, i}]} J_{w, i}
= - \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) g(V_i, w) \Delta_{0, w, i}}{p_0(V_i, w)} J_{0, w, i} + \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} p_1(V_i, w) g(V_i, w) \Delta_{0, w, i} J_{w, i}.
$$

Therefore, $D_{1n} + D_{2n}$ is equal to

$$
\sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \{ g(V_i, w) \Delta_{0, w, i} J_{1, w, i} - p_1(V_i, w) g(V_i, w) \Delta_{0, w, i} J_{w, i} \}
- \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) g(V_i, w) \Delta_{0, w, i} J_{0, w, i}}{p_0(V_i, w)}
+ \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} p_1(V_i, w) g(V_i, w) \Delta_{0, w, i} J_{w, i} + o_P(n^{-1/2})
= \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0, w, i} J_{1, w, i} - \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} p_1(V_i, w) g(V_i, w) \Delta_{0, w, i} J_{0, w, i} + o_P(n^{-1/2}),
$$

uniformly for all $p \in A$. As for the last difference, recall the definition $J_{d, w, i} \equiv L_{d, w, i} - E_{Q, w, i} [L_{d, w, i}]$ and write it as

$$
\sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0, w, i} L_{1, w, i}
- \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) g(V_i, w) \Delta_{0, w, i}}{p_0(V_i, w)} L_{0, w, i}
- \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0, w, i} E_{Q, w, i} [L_{1, w, i}]
+ \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) g(V_i, w) \Delta_{0, w, i}}{p_0(V_i, w)} E_{Q, w, i} [L_{0, w, i}].
$$
Therefore,

\[
\begin{align*}
D_n &= D_{1n} + D_{2n} \\
&= \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0,w,i} L_{1,w,i} \\
&\quad - \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w)g(V_i, w)\Delta_{0,w,i}}{p_0(V_i, w)} L_{0,w,i} + o_P(n^{-1/2}),
\end{align*}
\]

uniformly for all \( p \in A \). Now, we turn to \( C_n \) (in (6.45)) which we write as

\[
\begin{align*}
&\sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\varepsilon_{0,i}L_{0,w,i} \left\{ \hat{1}_{n,i} \frac{\hat{p}_{1,i}(V_i, w)}{\hat{p}_{0,i}(V_i, w)} - \hat{1}_{n,i} \frac{\hat{p}_{1,i}(V_i, w)}{\hat{p}_{0,i}(V_i, w)} \right\} + o_P(n^{-1/2}) \\
&= \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} H_i \frac{\hat{p}_{1,i}(V_i, w)\{\hat{p}_{0,i}(V_i, w) - \hat{p}_{0,i}(V_i, w)\}}{p_0^2(V_i, w)} \\
&\quad + \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} H_i \frac{\{\hat{p}_{1,i}(V_i, w) - \hat{p}_{1,i}(V_i, w)\}\hat{p}_{0,i}(V_i, w)}{p_0^2(V_i, w)} + o_P(n^{-1/2}) \\
&= \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} H_i \frac{\hat{p}_{1,i}(V_i, w)\{\hat{p}_{0,i}(V_i, w) - \hat{p}_{0,i}(V_i, w)\}}{p_0^2(V_i, w)} \\
&\quad + \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} H_i \frac{\{\hat{p}_{0,i}(V_i, w) - \hat{p}_{0,i}(V_i, w)\}\hat{p}_{0,i}(V_i, w)}{p_0^2(V_i, w)} + o_P(n^{-1/2}) \\
&= \sum_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} H_i \frac{\hat{p}_{0,i}(V_i, w) - \hat{p}_{0,i}(V_i, w)}{p_0^2(V_i, w)} + o_P(n^{-1/2}),
\end{align*}
\]

uniformly for all \( p \in A \), where \( H_i = g(V_i, w)\varepsilon_{0,i}L_{0,w,i} \). The uniformity comes from the fact that the convergence rate of \( \hat{p}_{0,i}(V_i, w) \) and \( \hat{p}_{0,i}(V_i, w) \) to \( p_0(V_i, w) \) is uniform for \( p \). The second equality follows from Lemma B2(ii). As for the last term, we apply Lemma B2(ii) to
write it as (up to $o_P(n^{-1/2})$, uniformly for all $p \in A$)

$$\sum_{w \in W} E_Q \left[ \frac{g(V_i, w)p_1(V_i, w)\tilde{z}_{0,i}L_{0,w,i}}{p_0(V_i, w)} \right] \left( \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} - \frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} \right)$$

because

$$E_Q \left[ \frac{p_1(V_i, w)\tilde{z}_{0,i}L_{0,w,i}}{p_0(V_i, w)} \right] = E \left[ g(V_i, w)\frac{p_1(V_i, w)}{p_0(V_i, w)}\tilde{z}_{0,i}1\{(D_i, W_i) = (0, w)\} \right]$$

$$= E \left[ g(V_i, w)p_1(V_i, w)\tilde{z}_{0,i} \right]$$

$$= E \left[ g(V_i, w)p_1(V_i, w)\Delta_{0,w,i} \right].$$

Now, let us turn to $B_n$ (in (6.45)), which can be written as

$$\sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\tilde{z}_{1,i}L_{1,w,i} - \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w)g(V_i, w)\tilde{z}_{0,i}L_{0,w,i}}{p_0(V_i, w)} + E_n,$$

where

$$E_n = \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\tilde{z}_{1,i}(\hat{L}_{1,w,i} - L_{1,w,i})$$

$$- \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w)g(V_i, w)\tilde{z}_{0,i}(\hat{L}_{0,w,i} - L_{0,w,i})}{p_0(V_i, w)}.$$

Now, we focus on $E_n$. Observe that

$$\frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\tilde{z}_{1,i}(\hat{L}_{1,w,i} - L_{1,w,i})$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\tilde{z}_{1,i}p_1,w \left( \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}^2} \right) 1\{(D_i, W_i) = (1, w)\} + o_P(n^{-1/2})$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(V_i, w)\tilde{z}_{1,i}L_{1,w,i} \left( \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} \right) + o_P(n^{-1/2})$$

$$= E_Q \left[ g(V_i, w)\tilde{z}_{1,i}L_{1,w,i} \right] \left( \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} \right) + o_P(n^{-1/2}),$$
uniformly for all $p \in A$. (Here the uniformity comes from that the convergence of $\hat{q}_{1,w}$ to $q_{1,w}$ does not depend on $p$). As for the last expectation,

$$
\mathbb{E}_Q [g(V_i, w) \tilde{\varepsilon}_{1,i} L_{1,w,i}]
= (p_{1,w}/q_{1,w}) \mathbb{E}_Q [g(V_i, w) \tilde{\varepsilon}_{1,i} 1\{(D_i, W_i) = (1, w)\}]
= \mathbb{E} [g(V_i, w) \tilde{\varepsilon}_{1,i} 1\{(D_i, W_i) = (1, w)\}]
= \mathbb{E} [p_1(V_i, w)g(V_i, w) (\beta_0(V_i, w) - \mathbb{E}_1 [g(X_i)\beta_0(X_i)] / \mathbb{E}_1 [g(X_i)]|D_i = 1)]
= \mathbb{E} [g(V_i, w)p_1(V_i, w)\Delta_{1,w,i}].
$$

Hence

$$
\frac{1}{n} \sum_{i=1}^n g(V_i, w) \tilde{\varepsilon}_{1,i} (\hat{L}_{1,w,i} - L_{1,w,i}) = \mathbb{E} [p_1(V_i, w)g(V_i, w)\Delta_{1,w,i}] \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} + o_P(n^{-1/2}),
$$

uniformly for all $p \in A$. Also,

$$
\frac{1}{n} \sum_{i=1}^n \frac{p_1(V_i, w)g(V_i, w) \tilde{\varepsilon}_{0,i} (\hat{L}_{0,w,i} - L_{0,w,i})}{p_0(V_i, w)}
= \frac{1}{n} \sum_{i=1}^n \frac{p_1(V_i, w)g(V_i, w) \tilde{\varepsilon}_{0,i}}{p_0(V_i, w)} \left( \frac{q_{0,w} - \hat{q}_{0,w}}{q_{0,w}} \right) L_{0,w,i}
= \mathbb{E} [p_1(V_i, w)g(V_i, w)\Delta_{0,w,i}] \frac{q_{0,w} - \hat{q}_{0,w}}{q_{0,w}} + o_P(n^{-1/2}),
$$

uniformly for all $p \in A$. Therefore, we write $E_n$ as

$$
\sum_{w \in W} \mathbb{E} [p_1(V_i, w)g(V_i, w)\Delta_{1,w,i}] \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}}
- \sum_{w \in W} \mathbb{E} [p_1(V_i, w)g(V_i, w)\Delta_{0,w,i}] \frac{q_{0,w} - \hat{q}_{0,w}}{q_{0,w}} + o_P(n^{-1/2}),
$$

uniformly for all $p \in A$.

Now, we collect all the results for $B_n$, $C_n$, and $D_n$ and plug these into (6.45) and to deduce that (up to $o_P(n^{-1/2})$ uniformly for all $p \in A$)

$$
\tilde{\tau}_{tet}^{SA}(p) - \tau_{tet}(p) = \frac{1}{p_1 \mathbb{E}_1 [g(X_i)]} \sum_{j=1}^6 G_{jn} + o_P(n^{-1/2}), \quad \text{uniformly over } p \in A,
$$
where

\[
G_{1n} = \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \left\{ g(V_i, w) \bar{\varepsilon}_{1,i} L_{1,w,i} \right\} - \frac{p_1(V_i, w) g(V_i, w) \bar{\varepsilon}_{0,i} L_{0,w,i}}{p_0(V_i, w)},
\]

\[
G_{2n} = \sum_{w \in W} E \left[ p_1(V_i, w) g(V_i, w) \Delta_{1,w,i} \right] q_{1,w} - \hat{q}_{1,w},
\]

\[
G_{3n} = -\sum_{w \in W} E \left[ p_1(V_i, w) g(V_i, w) \Delta_{0,w,i} \right] \frac{q_{0,w} - \hat{q}_{0,w}}{q_{0,w}},
\]

\[
G_{4n} = -\sum_{w \in W} E \left[ p_1(V_i, w) g(V_i, w) \Delta_{0,w,i} \right] \left( \frac{q_{0,w} - \hat{q}_{0,w}}{q_{0,w}} - \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} \right),
\]

\[
G_{5n} = -\sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \Delta_{0,w,i} L_{1,w,i}, \text{ and}
\]

\[
G_{6n} = \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) g(V_i, w) \Delta_{0,w,i}}{p_0(V_i, w)} L_{0,w,i}.
\]

We rewrite \(G_{2n} + G_{3n} + G_{4n}\) as

\[
\sum_{w \in W} E \left[ p_1(V_i, w) g(V_i, w) \left( \tau(V_i, w) - \frac{E_1[g(X_i) \tau(X_i)]]}{E_1[g(X_i)]} \right) \right] \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}},
\]

uniformly for all \(p \in A\).

By writing

\[\varepsilon_{d,i} = Y_{di} - \beta_d(V_i, w) + \beta_d(V_i, w) - \frac{E_1[g(X_i) \beta_d(X_i)]}{E_1[g(X_i)]} = \varepsilon_{d,w,i} + \Delta_{d,w,i},\]

and splitting the sums, we rewrite \(\tilde{\tau}_{tet}^S(p) - \tau_{tet}(p)\) as

\[
\tilde{\tau}_{tet}^S(p) - \tau_{tet}(p) = \frac{1}{p_1 E_1[g(X_i)]} \sum_{j=5}^{9} G_{jn} + o_P(n^{-1/2}),
\]

uniformly for all \(p \in A\), where

\[
G_{7n} = \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left\{ \varepsilon_{1,w,i} L_{1,w,i} - \frac{p_1(V_i, w) \varepsilon_{0,w,i} L_{0,w,i}}{p_0(V_i, w)} \right\},
\]

\[
G_{8n} = \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \left\{ \Delta_{1,w,i} L_{1,w,i} - \frac{p_1(V_i, w) \Delta_{0,w,i} L_{0,w,i}}{p_0(V_i, w)} \right\},
\]

\[
G_{9n} = \sum_{w \in W} E \left[ p_1(V_i, w) g(V_i, w) \left( \tau(V_i, w) - \frac{E_1[g(X_i) \tau(X_i)]]}{E_1[g(X_i)]} \right) \right] \frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}}.
\]
Note that the component in $G_{9n}$,

$$
E \left[ p_1(V_i, w)g(V_i, w) \left( \tau(V_i, w) - \frac{E_1[g(X_i)\tau(X_i)]}{E_1[g(X_i)]} \right) \right] = E_{1,w} \left[ g(V_i, w) \left( \tau(V_i, w) - \frac{E_1[g(X_i)\tau(X_i)]}{E_1[g(X_i)]} \right) \right] p_{1,w}.
$$

Using this and noting that

$$
\tau_{tet}(p) = \frac{E_1[g(X_i)\tau(X_i)]}{E_1[g(X_i)]}
$$

by the definition of $\tau_{tet}(p)$, we rewrite $G_{9n}$ as

$$
\sum_{w \in \mathcal{W}} E \left[ p_1(V_i, w)g(V_i, w) \left( \tau(V_i, w) - \tau_{tet}(p) \right) (q_{1,w} - \hat{q}_{1,w}) \right] = G_{10n}, \text{ say.}
$$

As for $G_{5n} + G_{6n} + G_{8n}$, we note that the part $G_{8n}$ that contains $p_1(V_i, w)\Delta_{0,w,i}L_{0,w,i}/p_0(V_i, w)$ cancels with $G_{6n}$, yielding that $G_{5n} + G_{6n} + G_{8n}$ is equal to

$$
\sum_{w \in \mathcal{W}} \sum_{i=1}^{n} \frac{1}{n} g(V_i, w) (\Delta_{1,w,i} - \Delta_{0,w,i}) L_{1,w,i}
$$

$$
= \sum_{w \in \mathcal{W}} \sum_{i=1}^{n} \frac{1}{n} g(V_i, w) \left( \tau(V_i, w) - \frac{E_1[g(X_i)\tau(X_i)]}{E_1[g(X_i)]} \right) L_{1,w,i}
$$

$$
= \sum_{w \in \mathcal{W}} \sum_{i=1}^{n} \frac{1}{n} g(V_i, w) (\tau(V_i, w) - \tau_{tet}(p)) L_{1,w,i} = G_{11n}, \text{ say,}
$$

using (6.47) again. Thus, we can rewrite $\tau_{tet}^S(p) - \tau_{tet}(p)$ as

$$
\frac{1}{p_1 E_1[g(X_i)]} \left\{ G_{7n} + G_{10n} + G_{11n} \right\}.
$$

However, as for $G_{10n}$, note that

$$
\frac{G_{10n}}{p_1 E_1[g(X_i)]} = \frac{1}{p_1 E_1[g(X_i)]} \sum_{w \in \mathcal{W}} p_{1,w} E_{1,w} \left[ g(V_i, w) (\tau(V_i, w) - \tau_{tet}(p)) \right]
$$

$$
= \frac{1}{p_1 E_1[g(X_i)]} E [g(X_i) (\tau(X_i) - \tau_{tet}(p)) 1\{D_i = 1\}]
$$

$$
= E_1 [g(X_i) (\tau(X_i) - \tau_{tet}(p))] = 0,
$$

as for $G_{10n}$, note that
from (6.47). Therefore, we conclude that
\[
\frac{1}{p_1} E_1 [g(X_i)] \{ G_{7n} + G_{11n} \}
\]
\[
= \frac{1}{p_1} E_1 [g(X_i)] \sum_{w \in W} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{1,w,i} L_{1,w,i} - \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(V_i, w) \varepsilon_{0,w,i} L_{0,w,i}}{p(V_i, w)} \right\}
\]
\[
+ \frac{1}{p_1} E_1 [g(X_i)] \sum_{w \in W} \frac{1}{n} \sum_{i=1}^{n} (\tau(V_i, w) - \tau_{tet}(p) - E_{1,w} [\tau(V_i, w) - \tau_{tet}]) L_{1,w,i} + o_P(n^{-1/2}),
\]
uniformly for all \( p \in A \). Hence the wanted result follows by the CLT. ■

In the following, we define \( \mu_d(v, w) \equiv \beta_d(v, w) f(v, w) \), \( \mu_{d,s}(v, w) \equiv \beta_d(v, w) f(v, w|s) \), \( \varepsilon_{d,i}(v, w) \equiv Y_{d,i} - \beta_d(v, w) \) and \( \beta_d^2(v, w) \equiv g(v, w) \beta_d(v, w) \).

**Lemma B4:** Suppose that Condition (C1) and Assumptions 1-2 hold. Then for each \( (d, w) \in \{0, 1\} \times W \),
\[
\sqrt{n} \int (\bar{\mu}_d(v, w) - \mu_d(v, w)) g(v, w) dv
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{d,w,i} g(V_i, w) \varepsilon_{d,i}(V_i, w) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{1,w,i} \{ \beta_d^2(V_i, w) - E_{1,w} [\beta_d^2(V_i, w)] \}
\]
\[
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{0,w,i} \{ \beta_d^2(V_i, w) - E_{0,w} [\beta_d^2(V_i, w)] \} + o_P(1),
\]
uniformly over \( p \in A \).

**Proof:** (i) Write \( \bar{\mu}_d(v, w) - \mu_d(v, w) \) as
\[
\frac{1}{\bar{p}_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \frac{\beta_d(v, w) f(v, w)}{f(v, w)} \right) \hat{L}_{d,w,i} K_{h,i}(v)
\]
\[
= \frac{1}{\bar{p}_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \beta_d(v, w) \right) \hat{L}_{d,w,i} K_{h,i}(v)
\]
\[
+ \frac{\beta_d(v, w) f(v, w)}{\bar{p}_d(v, w)} \left\{ \frac{1}{f(v, w)} - \frac{1}{\bar{f}(v, w)} \right\} \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{d,w,i} K_{h,i}(v).
\]
From the definition of \( \bar{f}(v, w) \) and \( \bar{p}_d(v, w) \) (see (3.5)), the second term becomes
\[
\beta_d(v, w) \{ \bar{f}(v, w) - f(v, w) \},
\]
so that we write $\tilde{p}_d(v, w) - \mu_d(v, w)$ as

$$\frac{1}{p_d(v, w) n} \sum_{i=1}^{n} \varepsilon_{d,i}(v, w) \hat{L}_{d,w,i} K_{h,i}(v) + \beta_d(v, w) \{ \tilde{f}(v, w) - f(v, w) \}.$$  

Following the same proof as in Lemma B1, we can show that

$$\sup_{v \in V(w)} \left| \frac{\tilde{p}_d(v, w) - p_d(v, w)}{p_d(v, w) + o_P(1)} \right| = O_P(\varepsilon_n),$$

uniformly for all $p \in A$. Using this, we deduce that

$$\left| \int \left( \frac{1}{\tilde{p}_d(v, w)} - \frac{1}{p_d(v, w)} \right) \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{d,i}(v, w) \hat{L}_{d,w,i} K_{h,i}(v) g(v, w) dv \right| \leq O_P(\varepsilon_n) \times A_n,$$

where $A_n \equiv A_{1n} + A_{2n}$ with

$$A_{1n} \equiv \int \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{d,i}(v, w) g(v, w) L_{d,w,i} K_{h,i}(v) \right| dv,$$

and

$$A_{2n} \equiv \int \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{d,i}(v, w) g(v, w) \left( \hat{L}_{d,w,i} - L_{d,w,i} \right) K_{h,i}(v) \right| dv.$$

We focus on $A_{1n}$. We bound the term by

$$\int \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{d,i}(V_i, w) g(V_i, w) L_{d,w,i} K_{h,i}(v) \right| dv$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \int |(\varepsilon_{d,i}(V_i, w) g(V_i, w) - \varepsilon_{d,i}(v, w) g(v, w)) L_{d,w,i} K_{h,i}(v)| dv.$$

The leading term is $O_P(n^{-1/2}h^{-d/2})$ from Theorem 1 of Lee, Song and Whang (2013), uniformly for all $p \in A$. (The uniformity over $p \in A$ is ensured because we can pull $p_{d,w}$ in the $L_{d,w,i}$ out of the integral.) The expectation of the second term is $o(n^{-1/2}h^{-d/2})$ using the higher order kernel property of $K$ and the smoothness condition for $\beta_d(v, w) g(v, w)$ in $v$. (See Assumptions 1(i), 2(i) and (ii).)

As for $A_{2n}$, we bound the term by

$$\left| \frac{1}{q_{d,w}} - \frac{1}{Q_{d,w}} \right| \int \left| \frac{p_{d,w}}{n} \sum_{i \in S_{d,w}} \varepsilon_{d,i}(v, w) g(v, w) K_{h,i}(v) \right| dv = O_P(n^{-1}h^{-d/2}),$$

uniformly for all $p \in A$. by applying the similar arguments to the integral.
Combining the results for $A_{1n}$ and $A_{2n}$, we conclude that

$$\left| \int \left( \frac{1}{\hat{p}_d(v, w)} - \frac{1}{p_d(v, w)} \right) \frac{1}{n} \sum_{i=1}^{n} \varepsilon_d, i(v, w) \hat{L}_{d, w, i} K_{h, i}(v) g(v, w) dv \right| = O_P(n^{-1/2} h^{-d/2} \varepsilon_n) = o_P(n^{-1/2}),$$

uniformly for all $p \in A$. Since (as we saw from the treatment of $A_{2n}$)

$$\left| \int \frac{1}{n} \sum_{i=1}^{n} \varepsilon_d, i(v, w) \left( \hat{L}_{d, w, i} - L_{d, w, i} \right) K_{h, i}(v) g(v, w) dv \right| = o_P(n^{-1/2}),$$

uniformly for all $p \in A$, the leading sum in (6.48) is equal to

$$\int \frac{1}{p_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_d, i(v, w) L_{d, w, i} K_{h, i}(v) g(v, w) dv + o_P(n^{-1/2}),$$

uniformly for all $p \in A$. Using change of variables, we deduce that

$$\int \frac{g(v, w)}{p_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \varepsilon_d, i(V_i, w) L_{d, w, i} K_{h, i}(v) g(v, w) dv = \frac{1}{n} \sum_{i=1}^{n} g(V_i, w) \varepsilon_d, i(V_i, w) L_{d, w, i} K_{h, i}(v) g(v, w) dv + o_P(n^{-1/2}).$$

We consider $\int g(v, w) \beta_d(v, w) \{ \tilde{f}(v, w) - f(v, w) \} dv$, which equals

$$(6.49) \quad \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{w, i} g(V_i, w) \beta_d(V_i, w) = \int g(v, w) \beta_d(v, w) f(v, w) dv + o_P(n^{-1/2}).$$

As for the above difference, note that

$$f(v, w) = f(v, w, 1) + f(v, w, 0) = \frac{f(v, w, 1)}{p_{1, w}} + \frac{f(v, w, 0)}{p_{0, w}},$$

where the first equality follows because $\hat{q}_{1, w} = n_{1, w}/n$. However, note that

$$\frac{p_{1, w} n_{1, w}}{n \hat{q}_{1, w}} = \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{1, w, i} \text{ and } \frac{p_{0, w} n_{0, w}}{n \hat{q}_{0, w}} = \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{0, w, i}. $$
Therefore, the leading difference in (6.49) is written as

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{L}_{1,w,i} \left\{ \beta^g_d(V_i, w) - \frac{1}{p_{1,w}} \int \beta^g_d(v, w) f(v, w|1)p_1 dv \right\} 
+ \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{0,w,i} \left\{ \beta^g_d(V_i, w) - \frac{1}{p_{0,w}} \int \beta^g_d(v, w) f(v, w|0)p_0 dv \right\} 
= \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{1,w,i} \left\{ \beta^g_d(V_i, w) - E_{1,w} [\beta^g_d(V_i, w)] \right\} 
+ \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{0,w,i} \left\{ \beta^g_d(V_i, w) - E_{0,w} [\beta^g_d(V_i, w)] \right\}.
\]

It is not hard to show that both the sums are equal to those with \( \hat{L}_{1,w,i} \) and \( \hat{L}_{0,w,i} \) replaced by \( L_{1,w,i} \) and \( L_{0,w,i} \) up to an error term of \( o_P(n^{-1/2}) \) order uniformly for \( p \in A \).

**Lemma B5:** Suppose that Condition (C1) and Assumptions 1-2 hold. Then for each \((d, w) \in \{0, 1\} \times W\),

\[
\sqrt{n} \int (\tilde{\mu}_{d,1}(v, w) - \mu_{d,1}(v, w)) g(v, w) dv = \frac{1}{p_1 \sqrt{n}} \sum_{i=1}^{n} L_{d,w,i} p_1(V_i, w) g(V_i, w) \frac{\varepsilon_d(V_i, w)}{p_d(V_i, w)}
+ \frac{1}{p_1 \sqrt{n}} \sum_{i=1}^{n} L_{1,w,i} \{ \beta^g_d(V_i, w) - E_{1,w} [\beta^g_d(V_i, w)] \} + o_P(1),
\]

uniformly for \( p \in A \).

**Proof:** Similarly as Lemma B4, we write \( \tilde{\mu}_{d,1}(v, w) - \mu_{d,1}(v, w) \) as

\[
\frac{1}{p_1 \tilde{p}_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{d,w,i} K_{h,i}(v) \cdot \left( Y_i - \frac{\beta_d(v, w) f(v, w|1)}{\tilde{f}(v, w|1)} \right)
= \frac{1}{p_1 \tilde{p}_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{d,w,i} K_{h,i}(v) \cdot (Y_i - \beta_d(v, w)) + \beta_d(v, w) \left( \tilde{f}(v, w|1) - f(v, w|1) \right).
\]

As for the leading term, using similar arguments in the proof of Lemma B4, we find that

\[
\int \frac{1}{p_1 \tilde{p}_d(v, w)} \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{d,w,i} K_{h,i}(v) \cdot (Y_i - \beta_d(v, w)) g(v, w) dv
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \hat{L}_{d,w,i} \int \frac{p_1(v, w)}{p_d(v, w)} K_{h,i}(v) \cdot (Y_i - \beta_d(v, w)) g(v, w) dv + o_P(n^{-1/2})
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} L_{d,w,i} \frac{p_1(V_i, w) g(V_i, w)}{p_d(V_i, w)} (Y_i - \beta_d(V_i, w)) + o_P(n^{-1/2}),
\]
uniformly over \( p \in A \). The last equality uses the fact that the summands have mean zero conditional on \((D_i, W_i) = (1, w)\). As for the second term, we also note that

\[
\int \beta_d(v, w) \left( \tilde{f}(v, w|1) - f(v, w|1) \right) g(v, w) dv
\]

\[
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}_{1, w, i} \int \beta_d^g(V_i, w) K_{h, i}(v) dv - \int \beta_d^g(v, w) K_{h, i}(v) f(v, w, 1) dv \right)
\]

\[
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}_{1, w, i} \beta_d^g(V_i, w) - \int \beta_d^g(v, w) K_{h, i}(v) f(v, w, 1) dv \right) + o_P(n^{-1/2}),
\]

uniformly for \( p \in A \). Similarly as before, we rewrite the last sum as

\[
\frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}_{1, w, i} \beta_d^g(V_i, w) - \frac{1}{p_1} \frac{1}{w} \int \beta_d^g(v, w) K_{h, i}(v) f(v, w, 1) dv \right)
\]

\[
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}_{1, w, i} \beta_d^g(V_i, w) - \mathbb{E}_{1, w} [\beta_d^g(V_i, w)] \right) + o_P(n^{-1/2})
\]

\[
= \frac{1}{p_1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}_{1, w, i} \beta_d^g(V_i, w) - \mathbb{E}_{1, w} [\beta_d^g(V_i, w)] \right) + o_P(n^{-1/2}),
\]

uniformly for \( p \in A \), because, again, the summands have mean zero conditional on \((D_i, W_i) = (1, w)\). FIG.

\[\text{Proof of Lemma A2: We consider the first statement of (6.9). Let}
\]

\[\hat{A} = \int (\tilde{\mu}_1(v, w) - \tilde{\mu}_0(v, w)) g(v, w) dv \quad \text{and} \quad A = \int (\mu_1(v, w) - \mu_0(v, w)) g(v, w) dv.
\]

Also let \( \hat{B} = \int \tilde{f}(v, w) g(v, w) dv \) and \( B = \int f(v, w) g(v, w) dv \). By Lemma B4, \( \hat{A} - A = O_P(1/\sqrt{n}) \) and by replacing \( Y_i \) by 1 in the proof of Lemma B4, we find that \( \hat{B} - B = O_P(1/\sqrt{n}) \). Hence we conclude that

\[
\sqrt{n}\{\hat{\tau}_{ATE} - \tau_{ATE}(p)\} = \frac{\sqrt{n}(\hat{A} - A)}{B} - \frac{A}{B} \cdot \frac{\sqrt{n}(\hat{B} - B)}{B} + o_P(1).
\]
Applying Lemma B4 to $\hat{A} - A$ and $\hat{B} - B$ and combining the terms, we write $\sqrt{n}\{\tau_{\text{ate}}(p) - \tau_{\text{ate}}(p)\}$ as

$$
\frac{1}{B} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{1,w,i} g(V_i, w) \frac{\varepsilon_{1,i}(V_i, w)}{p_1(V_i, w)} \\
- \frac{1}{B} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{0,w,i} g(V_i, w) \frac{\varepsilon_{0,i}(V_i, w)}{p_0(V_i, w)} \\
+ \frac{1}{B} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{1,w,i} (g(V_i, w)(\tau(V_i, w) - \tau_{\text{ate}}(p))) - E_{1,w}[g(V_i, w)(\tau(V_i, w) - \tau_{\text{ate}}(p))] \\
+ \frac{1}{B} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{0,w,i} (g(V_i, w)(\tau(V_i, w) - \tau_{\text{ate}}(p))) - E_{0,w}[g(V_i, w)(\tau(V_i, w) - \tau_{\text{ate}}(p))] + o_P(1),
$$

uniformly for $p \in A$.

Let us turn to the second statement of (6.9). We write

$$
\tilde{\tau}_{\text{tet}}(p) - \tau_{\text{tet}}(p) = \left[ \frac{\sum_{w \in W} \int_Y g(v, w) \bar{\mu}_{1,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) \bar{f}(v, w|1) dv} - \frac{\sum_{w \in W} \int_Y g(v, w) \mu_{1,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) f(v, w|1) dv} \right] \\
- \left[ \frac{\sum_{w \in W} \int_Y g(v, w) \bar{\mu}_{0,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) \bar{f}(v, w|1) dv} - \frac{\sum_{w \in W} \int_Y g(v, w) \mu_{0,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) f(v, w|1) dv} \right].
$$

Note that

$$
\frac{\sum_{w \in W} \int_Y g(v, w) \bar{\mu}_{d,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) \bar{f}(v, w|1) dv} - \frac{\sum_{w \in W} \int_Y g(v, w) \mu_{d,1}(v, w) dv}{\sum_{w \in W} \int_Y g(v, w) f(v, w|1) dv} = \frac{1}{E_1[g(X)]} \sum_{w \in W} \int (\bar{\mu}_{d,1}(v, w) - \mu_{d,1}(v, w)) g(v, w) dv \\
- \frac{\tau_{d,1}}{E_1[g(X)]} \sum_{w \in W} \int \left( \bar{f}(v, w|1) - f(v, w|1) \right) g(v, w) dv + o_P(n^{-1/2}),
$$

where

$$
\tau_{d,1} \equiv \frac{E_1[\beta_d(X)g(X)]}{E_1[g(X)]}.
$$

As for the last sum, by putting $Y_i = 1$ in the proof of Lemma B5, we obtain the following:

$$
(6.50) \quad \sqrt{n} \int \{ \bar{f}(v, w|1) - f(v, w|1) \} g(v, w) dv \\
= \frac{1}{p_1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_{1,w,i} (g(V_i, w) - E_{1,w}[g(V_i, w)]) + o_P(1).
$$
Using Lemma B5 (recall that \( \tilde{\mu}_{d,1}(v, w) = \tilde{\mu}_d(v, w)\tilde{p}_{1,i}(V_i, w)/p_1 \) and \( 6.50 \)), we conclude that for \( d \in \{0, 1\} \),

\[
\sqrt{n} \left\{ \sum_{w \in W} \int_Y g(v, w) \tilde{\mu}_{d,1}(v, w)dv - \sum_{w \in W} \int_Y g(v, w)f(v, w)dv \right\} = \sum_{w \in W} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{L_{d,w,i}p_1(V_i, w)g(V_i, w)}{E_1 [g(X)] p_1 \cdot p_d(V_i, w)} \varepsilon_{d,i}(V_i, w) + o_P(n^{-1/2}),
\]

uniformly for all \( p \in A \), where

\[
J(V_i, w) = g(V_i, w)(\beta_d(V_i, w) - \tau_{d,1}) - E_{1,w}[g(V_i, w)(\beta_d(V_i, w) - \tau_{d,1})].
\]

Combining two terms with \( d \in \{0, 1\} \) gives the desired result. \( \square \)

The following lemma is used to prove Lemma B2(i) and may be useful for other purposes. Hence we make the notations and assumptions self-contained here. Let \((Z_i, H_i, X_i)_{i=1}^n \) be an i.i.d. sample from \( P \), where \( Z_i \) and \( H_i \) are random variables. Let \( X_i = (X_{1i}, X_{2i}) \in \mathbb{R}^{L_1+L_2} \) where \( X_{1i} \) is continuous and \( X_{2i} \) is discrete, and let \( K_{ji} = K_h(X_{1j} - X_{1i})1\{X_{2j} = X_{2i}\} \), \( K_h(·) = K(·/h)/h^{L_1} \). Let \( \mathcal{X} \) be the support of \( X_i \) and \( f(·) \) be its density with respect to a \( \sigma \)-finite measure.

**Assumption D1 :** (i) For some \( \sigma \geq 4 \), \( \sup_{x \in \mathcal{X}} ||x_1||^{L_1}E[|Z_i|^{\sigma}|X_i = (x_1, x_2)] < \infty \), \( E[|H_i|^{\sigma}] < \infty \), and \( E[||X_i||^{\sigma}] < \infty \).

(ii) \( f(·, x_2) \), \( E[Z_i|X_{1i} = ·, X_{2i} = x_2]f(·, x_2) \) and \( E[H_i|X_{1i} = ·, X_{2i} = x_2]f(·, x_2) \) are \( L_1 + 1 \) times continuously differentiable with bounded derivatives on \( \mathbb{R}^{L_1} \) and their \((L_1 + 1)\)-th derivatives are uniformly continuous.

(iii) \( f \) is bounded and bounded away from zero on \( \mathcal{X} \).

**Assumption D2 :** For the kernel \( K \) and the bandwidth \( h \), Assumption 2 holds.

**Lemma D1 :** Suppose that Assumptions D1-D2 hold. Let \( 1_{n,i} = 1\{|X_i| \geq c_n\} \). Then

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_i \left\{ E[Z_i|X_i] - \frac{1}{n} \sum_{i=1}^{n} Z_j K_{ji} \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ E[E[H_i|X_i] Z_i] - E[H_i|X_i] Z_i \} + o_P(1).
\]
Proof : For simplicity, we only prove the result for the case where $X_i = X_{1,i}$ so that $X_i$ is continuous. Write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_i \left\{ E[Z_i|X_i] - \frac{1_{n,i}}{nf(X_i)} \sum_{j=1, j \neq i}^{n} Z_j K_{ji} \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_i \left\{ E[Z_i|X_i] \hat{f}(X_i) - \frac{1_{n,i}}{nf(X_i)} \sum_{j=1, j \neq i}^{n} Z_j K_{ji} \right\} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_i \left\{ E[Z_i|X_i] \{f(X_i) - \hat{f}(X_i)\} \right\} = A_{1n} + A_{2n}.$$

It suffices to show that

$$A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[H_i|X_i] \{E[Z_i|X_i] - Z_i\} + o_P(1), \text{ and}$$

$$A_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{E[E[H_i|X_i] E[Z_i|X_i]] - E[H_i|X_i] E[Z_i|X_i]\} + o_P(1).$$

First we write

$$A_{1n} = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} q_h(S_i, S_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[q_h(S_i, S_j)|S_j] + r_{1,n},$$

where $q_h(S_i, S_j) = H_i \{E[Z_i|X_i] - Z_j\} K_{ji}/f(X_i)$ and $S_i = (X_i, Z_i, H_i)$, and

$$r_{1,n} = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \{q_h(S_i, S_j) - E[q_h(S_i, S_j)|S_j]\}.$$

Observe that

$$n^{-1} E \left( q_h(S_i, S_j)^2 \right) = n^{-1} E \left[ H_i^2 \{E[Z_i|X_i] - Z_j\}^2 K_{ji}^2 / f^2(X_i) \right]$$

$$\leq n^{-1} \sqrt{E \left[ K_{ji}^4 \right]} = O(n^{-1} h^{-2L_1} \delta_n^{-2}) = o(1)$$

by change of variables and by Assumptions D1(iii) and D2. Therefore, by Lemma 3.1 of Powell, Stock, and Stoker (1989), $r_n = o_P(1)$. As for $E[q_h(S_i, S_j)|S_j]$, we use change of variables, Taylor expansion, and deduce that

$$E \left[ |E[q_h(S_i, S_j)|S_j] - E[H_j|X_j] \{E[Z_j|X_j] - Z_j\}| \right] = o(n^{-1/2}).$$

The wanted representation follows from this.
As for \( A_{2n} \),
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{H_i \mathbb{E}[Z_i|X_i]}{f(X_i)} \left\{ f(X_i) - \hat{f}(X_i) \right\} = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} s_h(S_i, S_j),
\]
where
\[
s_h(S_i, S_j) = \frac{H_i \mathbb{E}[Z_i|X_i]}{f(X_i)} \left\{ f(X_i) - \mathbb{E}[K_{ji}|X_i] \right\}.
\]
Since we can write \( \mathbb{E}[K_{ji}|X_i] = f(X_i) + O_P(h_i^{L_i+1}) \) uniformly over \( 1 \leq i \leq n \), we find that
\[
\mathbb{E}[s_h(S_i, S_j)|S_i] = \frac{H_i \mathbb{E}[Z_i|X_i]}{f(X_i)} \left\{ f(X_i) - \mathbb{E}[K_{ji}|X_i] \right\} = o_P(n^{-1/2}),
\]
uniformly over \( 1 \leq i \leq n \). Hence we can write
\[
A_{2n} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbb{E}[s_h(S_i, S_j)|S_j] + r_{2,n} + o_P(1),
\]
where
\[
r_{2,n} = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \{ s_h(S_i, S_j) - \mathbb{E}[s_h(S_i, S_j)|S_j] \}.
\]
Note that \( n^{-1} \mathbb{E}\left[s_h(S_i, S_j)^2\right] = o(1) \) and that
\[
\mathbb{E}[s_h(S_i, S_j)|S_j] = \mathbb{E}\left[H_i \mathbb{E}[Z_i|X_i] - \frac{\mathbb{E}[H_i|X_i] \mathbb{E}[Z_i|X_i]}{f(X_i)} K_{ji}|S_j\right]
\]
\[
= \mathbb{E}[H_i \mathbb{E}[Z_i|X_i]] - \mathbb{E}[H_j|X_j] \mathbb{E}[Z_j|X_j] + o_P(n^{-1/2}),
\]
uniformly over \( 1 \leq j \leq n \), yielding the desired representation for \( A_{2n} \).

References

Vancouver School of Economics, University of British Columbia, 997-1873 East Mall, Vancouver, BC, V6T 1Z1, Canada
E-mail address: kysong@mail.ubc.ca

Vancouver School of Economics, University of British Columbia, 997-1873 East Mall, Vancouver, BC V6T 1Z1, Canada.
E-mail address: yuzhengfei@gmail.com