



Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
<http://www.econ.upenn.edu/pier>

PIER Working Paper 09-011

“Efficient Estimation of Average Treatment Effects under
Treatment-Based Sampling”

by

Kyungchul Song

<http://ssrn.com/abstract=1360961>

Efficient Estimation of Average Treatment Effects under Treatment-Based Sampling

Kyungchul Song¹

Department of Economics, University of Pennsylvania

March 6, 2009

Abstract

Nonrandom sampling schemes are often used in program evaluation settings to improve the quality of inference. This paper considers what we call treatment-based sampling, a type of standard stratified sampling where part of the strata are based on treatments. This paper first establishes semiparametric efficiency bounds for estimators of weighted average treatment effects and average treatment effects on the treated. In doing so, this paper illuminates the role of information about the aggregate shares from the original data set. This paper also develops an optimal design of treatment-based sampling that yields the best semiparametric efficiency bound. Lastly, this paper finds that adapting the efficient estimators of Hirano, Imbens, and Ridder (2003) to treatment-based sampling does not always lead to an efficient estimator. This paper proposes different estimators that are efficient in such a situation.

Key words and Phrases: treatment-based sampling, semiparametric efficiency, treatment effects.

JEL Classifications: C12, C14, C52.

1 Introduction

Program evaluation studies often adopt nonrandom sampling to improve the quality of inference. Typically, participants are oversampled to get a larger number of observations than would be obtained in a same size random sample. The rationale for treatment-based sampling

¹This paper began as a ramification from my joint work with Petra Todd. I would like to express my gratitude to her for numerous kind and valuable comments and advice. All errors are solely mine. Address correspondence to: Kyungchul Song, Department of Economics, University of Pennsylvania, 528 McNeil Building, 3718 Locust Walk, Philadelphia, Pennsylvania 19104-6297.

is to reduce data collection costs and to improve the precision of the estimated treatment effects. For example, Ashenfelter and Card (1985) analyzed data from the CETA training program using a sample constructed by combining subsamples of program participants and a sample of nonparticipants drawn from the CPS. Also, the studies of Lalonde (1986), Dehejia and Wahba (1998, 1999) and Smith and Todd (2005) investigated the NSW training program where the training group consisted of individuals eligible for the program and the comparison sample were drawn from the CPS and PSID surveys. Numerous studies focused on the JTPA job training program (e.g. Heckman, Ichimura, Smith and Todd (1998), Heckman, Ichimura and Todd (1997), Ham and Lalonde (1996)). These studies typically used data set that oversampled program participants, where the program participants represented about 50% in the study sample in comparison to 3% in the population.

This paper proposes semiparametrically efficient inference procedures under treatment-based sampling. The sampling scheme that this paper focuses on can be described as follows. Let D be a random variable that takes values in $\mathcal{D} = \{0, 1\}$, where $D = d$ denotes participating in the d -th program. Let $X = (V, W)$ be a vector of covariates, where W is a discrete random variable taking values from a finite set \mathcal{W} . Assume that initially a random sample of size N for the discrete vector (D, W) is collected and let $N_{d,w} = \sum_{i=1}^N 1\{(D_i, W_i) = (d, w)\}$, $(d, w) \in \mathcal{D} \times \mathcal{W}$. From each of the $N_{d,w}$ subsamples, a random sample $\{Z_i\}_{i=1}^{n_{d,w}}$ of predetermined size $n_{d,w}$, $(d, w) \in \mathcal{D} \times \mathcal{W}$, for a vector $Z = (Y, V)$, is collected, where $Y = \sum_{d \in \mathcal{D}} Y_d 1\{D = d\}$ and Y_d denotes a potential outcome of a participant in the d -th program. In this paper, we call this type of sampling, *treatment-based sampling* as the strata $\mathcal{D} \times \mathcal{W}$ contain treatments.² When $W_i = 1$ for all i , so that the strata are constructed only based on the treatments, we call this sampling *pure treatment-based sampling*. (In pure treatment-based sampling, we suppress the subscript of w writing p_d instead of $p_{d,w}$, for example.) Throughout this paper, it is assumed that the aggregate shares $\{N_{d,w}/N\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ are the only available information from the original sample of size N , and we no longer have individual observations for $(D_i, W_i)_{i=1}^N$ from the original data set. Although the observations in the combined sample $\{(D_i, Z_i)\}_{i=1}^n$ are independent across different i 's, the marginals under treatment-based sampling are not identical. Hence inference based on random sampling can be misleading.

The objects of inference here are counterfactual quantities called weighted average treatment effect and average treatment effect on the treated:

$$\tau_{wate} = \frac{\mathbf{E}[g(X)\{Y_1 - Y_0\}]}{\mathbf{E}[g(X)]} \text{ and } \tau_{atet} = \mathbf{E}[Y_1 - Y_0 | D = 1]. \quad (1)$$

²The term, "treatment-based sampling", was borrowed from an anonymous reviewer's report when the idea of this paper was submitted as a research proposal to the NSF.

Efficient inference procedures were suggested for these parameters by Hahn (1998) and Hirano, Imbens, and Ridder (2003) (HIR, hereafter) under a random sampling scheme.

The identification of τ_{atet} under pure treatment-based sampling does not require knowledge of the aggregate shares p_d . However, information about the aggregate shares $p_{d,w}$ is required for the identification of τ_{wate} under treatment-based sampling and that of τ_{atet} under treatment-based sampling with additional strata \mathcal{W} . In this case, this paper assumes that we know the aggregate shares $p_{d,w}$. As Wooldridge (2001) has pointed out, this assumption is often motivated by the sampling environment where $N_{d,w}$ is very large as compared to the size of subsamples $n_{d,w}$. This method of sampling is reasonable in many situations where it is much less costly to gather information about (D, W) than the outcome Y or covariates X . In this case, a proper large sample theory would be one with $n_{d,w}/N_{d,w} \rightarrow_P 0$. At the level of treatment-based samples, the asymptotic theory is equivalent to assuming that we know the aggregate shares $p_{d,w} = P\{(D, W) = (d, w)\}$. When the aggregate shares are estimated using other data sources, the inference procedure in this paper should be viewed as one conditioned on the external information. However, when the object of interest is τ_{atet} under pure treatment-based sampling, we do not assume that the aggregate shares p_d are known.

The sampling scheme of this paper is a kind of standard stratified (SS) sampling (Imbens and Lancaster (1996)), and is one of various nonrandom sampling schemes studied in the literature. Early literatures on nonrandom sampling have assumed that the conditional probability of Z 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 (2008) investigated the problem of efficient estimation under nonrandom sampling schemes. Tripathi (2008) considered moment-based models under various nonrandom 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 (2008) 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. Unlike variable probability sampling, we cannot identify the joint distribution of observations from the standard stratified sampling without full knowledge of $p_{d,w}$ or some other data sources that ensure the identification of $p_{d,w}$.

In the vast literature of program evaluations, there are surprisingly few researches that

deal with inference under treatment-based sampling. Chen, Hong, and Tarozi (2008) established semiparametric efficiency bounds and proposed efficient estimation in a broader context where one has outcome observations with missing values and has auxiliary data that aid identification. The approach of Chen, Hong, and Tarozi (2008) applies to some stratified sampling schemes. However, their framework does not apply here because the elimination of an observation from the treatment-based sample is based not only on W in the covariate X but also on the treatment decision D . Hence the unconfoundedness condition assumed in their paper fails for observations from treatment-based sampling. A paper by Heckman and Todd (2008) offers a nice, simple idea to estimate τ_{atet} under pure treatment-based sampling without assuming knowledge of aggregate shares. However, their paper does not focus on efficient procedures.

This paper first establishes efficiency bounds for τ_{wate} and τ_{atet} under treatment-based sampling. As byproduct, we also obtain a necessary and sufficient condition for the sampling design under which a pure treatment-based sampling scheme offers a better semiparametric efficiency bound than a random sampling scheme. Furthermore, we characterize optimal treatment-based sampling that leads to a best semiparametric efficiency bound.

The result of an optimal design of treatment-based sampling is related to a recent paper by Hahn, Hirano, and Karlan (2008) who suggested an optimal design of social experiments that is conducive to program evaluations of improved quality. Their paper proposes a two-stage design of social experiments in which individuals are assigned to treatment based on their propensity scores and these propensity scores are designed to attain low asymptotic variance of average treatment effects. However, their paper's framework is different from this paper as it considers observations drawn from the population by random sampling in both stages. More importantly, unlike their paper, this paper is not concerned with a design of treatment decision for each individual as it is assumed that the population proportions $p_{d,w}$ are already given. The primary focus of this paper is on nonrandom sampling of observations related to treatment programs and its effect upon the inference quality of treatment effects estimators.

The main challenge in the development of optimal inference in this situation is that it is not clear *a priori* how we can obtain an efficient estimator from the computed semiparametric efficiency bounds. Obviously the usual approach of the sample analogue principle does not work because the observations are not from random sampling. One might think that one can apply the efficient estimator of Hahn (1998) or Hirano, Imbens, and Ridder (2008) to this situation of treatment-based sampling by employing appropriate change of measure as in Tripathi (2008). However, this paper demonstrates that this approach of naive adaptation does not work in general. Indeed, it is shown that the adapted version of the weighted average

treatment effects is inefficient. The main reason is that in this case, the knowledge of the aggregate shares is not ancillary. This paper proposes a different estimator that achieves the semiparametric efficiency bound in this situation. As this paper shows, it turns out that the weighting by propensity scores in HIR should be modified to achieve efficiency.

The situation of primary relevance in practice appears to be the case where the object of interest is τ_{atet} and the sampling is pure treatment-based sampling. This is the situation that was of focus in Heckman and Todd (2008). In this situation, we do not require knowledge of the aggregate shares, for τ_{atet} is identified without it. It is shown that in this case, the knowledge of the aggregate shares p_d is ancillary. Hence one might consider an estimator that is obtained by adapting the estimator of HIR to the pure treatment-based sampling scheme. It is shown that indeed such an estimator is efficient. However, this estimator requires knowledge of the aggregate shares for its construction. This paper proposes an alternative estimator that is efficient and does not require knowledge of the aggregate shares.

This paper proceeds as follows. Section two discusses a general method to find semi-parametric efficiency bounds under treatment-based sampling data designs. Section three establishes semiparametric efficiency bounds for weighted average treatment effects and average treated effects on the treated. Section four investigates efficient estimation. Section five concludes and the proofs are relegated to the appendix.

2 Treatment-Based Sampling and Semiparametric Efficiency

2.1 Treatment-Based Sampling

In this section, we discuss a general method of computing semiparametric efficiency bounds under treatment-based sampling and an optimal design of treatment-based sampling. Suppose we are interested in a certain parameter $\psi(P)$ from the distribution P of a random vector (Z, D, W) , where (D, W) is a discrete random variable taking values from a finite set $\mathcal{D} \times \mathcal{W}$, and $Z = (Y, V)$ denotes a vector of outcome variable Y and a covariate vector V .

First, note that a likelihood for observations generated from standard stratified sampling can be viewed as a conditional likelihood from multinomial sampling given $\{n_{d,w}\}_{d,w \in (\mathcal{D} \times \mathcal{W})}$. As pointed out by Imbens and Lancaster (1996) (see also Tripathi (2008)), (D, W) is ancillary in both stratified sampling and multinomial sampling, and hence it suffices for semiparametric efficiency to consider only multinomial sampling with probabilities $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$. Furthermore, $\{n_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ is a sufficient statistic for multinomial distributions, and hence as long as semiparametric efficiency is concerned, we can assume that the marginal design

probabilities $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ in multinomial sampling are known. As it will turn out later, however, we do not require full knowledge of $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ for the actual construction of efficient estimators.

Let the observations $\{(Z_i, D_i, W_i)\}_{i=1}^n$ for (Z, D, W) be generated by the multinomial sampling scheme using known probabilities $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$. In other words, we draw a stratum (d, w) from $\mathcal{D} \times \mathcal{W}$ using the multinomial distribution with known probabilities $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$, and then draw Z conditional on $(D, W) = (d, w)$ until the total sample size becomes n . Unless $q_{d,w} = p_{d,w}$ for all $(d, w) \in \mathcal{D} \times \mathcal{W}$, the observations $\{(Z_i, D_i, W_i)\}_{i=1}^n$ are not i.i.d. draws from P . The observations $\{(Z_i, D_i, W_i)\}_{i=1}^n$ are i.i.d., however, under the probability Q with density $q_{d,w} f_{Z|D,W}(z|d, w)$, where $f_{Z|D,W}(z|d, w)$ is the conditional density of Z given $(D, W) = (d, w)$. Therefore, the situation of treatment-based sampling is that we have observations that are i.i.d. from the probability Q but the parameter of interest is a functional of the probability P . The notations of expectation and variance without subscripts are assumed to be the expectation and variance under P . Expectation $\mathbf{E}_{d,w}$ denotes the conditional expectation given $(D, W) = (d, w)$.

2.2 Semiparametric Efficiency under Treatment-Based Sampling

In this section, we explain how we can compute the semiparametric efficiency bound for the parameter $\psi(P)$. The standard theory of efficiency in semiparametric models and methods to compute efficiency bounds are fairly well established and expounded in the literature. (See Newey (1990) and Bickel, Klaassen, Ritov, and Wellner (1993) for a review.) Closely related to this paper, Bickel and Kwon (2001) showed how we can adapt the results based on i.i.d. sampling to a multinomial sampling environment. (See Example 1 there.) To save the space, we assume basic terminologies and concepts in Bickel, Klassen, Ritov, and Wellner (1993) and highlight how the standard method can be adapted to observations from treatment-based sampling.

Since we know the marginal probabilities $q_{d,w}$, we consider the following form of regular parametric submodels:

$$f_t(z, d, w) = f_{Z|D,W}^t(z|d, w)q_{d,w}, \quad t \in [0, \varepsilon), \quad \varepsilon > 0, \quad (2)$$

where $\{f_{Z|D,W}^t(\cdot|d, w) : t \in [0, \varepsilon)\}$ denotes a regular parametric submodel passing through $f_{Z|D,W}(\cdot|d, w)$, the conditional density of Z given $(D, W) = (d, w)$. Then, the parametric submodel $\{f_t : t \in [0, \varepsilon)\}$ is associated with a score, $s(z, d, w) = s_{d,w}(z) \in L_2(Q)$, where $s_{d,w} = \frac{\partial}{\partial t} \log f_{Z|D,W}^t(\cdot|d, w)|_{t=0}$ denotes the score associated with $\{f_{Z|D,W}^t(\cdot|d, w) : t \in [0, \varepsilon)\}$. Let \mathcal{T} denote the tangent space, i.e., the closed linear span of all such scores s for all regular

parametric submodels in the form of (2).

There are two situations for the identification of $\psi(P)$ that this paper considers. The first situation is where we can identify $\psi(P)$ only using the conditional distribution of Z given (D, W) . The second situation is where we have knowledge of the aggregate shares $p_{d,w}$ which is needed to identify $\psi(P)$. In both cases, the relevant tangent space is the same \mathcal{T} and $\psi(P)$ is identified from the knowledge of Q and $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$. Hence, we can write

$$\psi(P) = \psi_Q(Q),$$

for some functional ψ_Q . The parameter of interest $\psi_Q(Q)$ is assumed to be Fréchet differentiable and to have $\dot{\psi}_Q \in L_2(Q)$ such that for all regular parametric submodels of the form in (2),

$$\frac{\partial \psi_Q(Q_t)}{\partial t} \Big|_{t=0} = \mathbf{E}_Q \left[\dot{\psi}_Q(Z, D, W) s(Z, D, W) \right].$$

When $\dot{\psi}_Q \in \mathcal{T}$, we call it an efficient influence function and denote it by $\dot{\psi}_Q^e$. Then, the semiparametric efficiency bound is given by the inverse of

$$V_{TS} \equiv \text{Var}_Q(\dot{\psi}_Q^e(Z, D, W)) = \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} q_{d,w} \mathbf{E}_{d,w} \left[\dot{\psi}_Q^e(Z, D, W)^2 \right]. \quad (3)$$

In this paper, we find $\dot{\psi}_Q^e(Z, D, W)$ in the following way. First, note that \mathcal{T} can be also viewed as the tangent space at P with parametric submodels P_t having density $f_{Z|D,W}^t(z|d, w) p_{d,w}$. We first find $\dot{\psi}_P \in L_2(P)$ such that for all regular parametric submodels with density $f_{Z|D,W}^t(z|d, w) p_{d,w}$,

$$\frac{\partial \psi(P_t)}{\partial t} \Big|_{t=0} = \mathbf{E} \left[\dot{\psi}_P(Z, D, W) s(Z, D, W) \right], \quad (4)$$

for some $s \in \mathcal{T}$. Then, observe that

$$\mathbf{E} \left[\dot{\psi}_P(Z, D, W) s(Z, D, W) \right] = \mathbf{E}_Q \left[\dot{\psi}_Q(Z, D, W) s(Z, D, W) \right],$$

if we take $\dot{\psi}_Q(z, d, w) = \dot{\psi}_P(z, d, w) p_{d,w} / q_{d,w}$. Hence we find an influence function $\dot{\psi}_P^e$ under P such that $\dot{\psi}_Q^e(z, d, w) = \dot{\psi}_P^e(z, d, w) p_{d,w} / q_{d,w}$ falls into \mathcal{T} . Thus, $\dot{\psi}_Q^e(z, d, w)$ constructed in this way is an efficient influence function.

2.3 Optimal Design of Treatment-Based Sampling

The conventional wisdom tells us that when the proportion of a subsample in the population is small, sampling relatively more from the subsample may 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. (See Hahn, Hirano, and Karlan (2008) for a similar observation.) Based on the variance bound in (3), we can develop a theory of an optimal design of treatment-based sampling.

Once we identify $\dot{\psi}_P^e(z, d, w) = \dot{\psi}_Q^e(z, d, w)q_{d,w}/p_{d,w}$, we can design an optimal treatment-based sampling as follows. Let

$$J_{d,w} = p_{d,w}^2 \mathbf{E}_{d,w} \left[\dot{\psi}_P^e(Z, D, W)^2 \right].$$

Then, we can write the variance bound as

$$V_{TS} = \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \frac{J_{d,w}}{q_{d,w}}.$$

We can view $J_{d,w}/q_{d,w}$ as the contribution of the (d, w) -subsample to the variance bound. The contribution decreases in $q_{d,w}$ which naturally captures the fact that by sampling more from the (d, w) -th subsample, we can reduce the sampling variability to the estimator that is contributed by the subsample. Then, the natural question is concerned with the optimal design of treatment-based sampling. We define the optimal design to be those $\{q_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ such that minimize V_{TS} under the constraint that $q_{d,w} \geq 0$ and $\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} q_{d,w} = 1$. It is easy to see that the optimal design is given by

$$q_{d,w}^* = \frac{\sqrt{J_{d,w}}}{\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \sqrt{J_{d,w}}}. \quad (5)$$

Therefore, the optimal design of treatment-based sampling suggests that we sample from the (d, w) -subsample precisely according to the "noise" proportion of $\sqrt{J_{d,w}}$ in $\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \sqrt{J_{d,w}}$. 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 guidance for optimally choosing the size of the sampling fractions $q_{d,w}$.³

Using the optimal design of treatment-based sampling $q_{d,w}^*$ yields *the minimum semipara-*

³The results here are predicated on the assumption that it is equally costly to sample from the (d, w) subsample for each $(d, w) \in \mathcal{D} \times \mathcal{W}$. However, sometimes, it may be less costly to sample from a specific subsample from others. In this case, we can incorporate an appropriate differential cost consideration into the optimal design by turning the optimization problem into one subject to certain inequality constraints.

metric efficiency bound for $\psi(P)$ as

$$\left\{ \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \sqrt{J_{d,w}} \right\}^2. \quad (6)$$

The variance in (6) is the minimum variance bound over all the choices of the sampling probabilities $q_{d,w}$. The variance (6) 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 yield improved inference than random sampling. Let V_{RS} be the variance bound 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

$$\min \left\{ p_1, \frac{J_1}{J_1 + J_0} \right\} \leq q_1 \leq \max \left\{ p_1, \frac{J_1}{J_1 + J_0} \right\}. \quad (7)$$

Therefore, it is not always true that sampling more from a subsample of low population proportion leads to a better result. Improvement by treatment-based sampling 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 strictly upon random sampling.

While the optimal design of treatment-based sampling clarifies the role of treatment-based sampling in improving the quality of inference, there is a caveat obvious yet worth mentioning. The results of the optimal design are justified only for large samples, where we can obtain a reliable estimate of $J_{d,w}$ for each $(d, w) \in \mathcal{D} \times \mathcal{W}$. When the suggested optimal proportion of the (d, w) -subsample is too small to the extent that the asymptotic justification of the estimate of $J_{d,w}$ is cast in doubt, the optimal design suggested here can lead to a grossly suboptimal choice.

3 Semiparametric Efficiency Bounds

3.1 Preliminary Discussion

We turn back to the set-up of program evaluations introduced in the beginning of this paper. This paper assumes the unconfoundedness assumption:

$$(Y_0, Y_1) \perp\!\!\!\perp D | X. \quad (8)$$

Under the unconfoundedness condition, a variety of treatment effect parameters are identified. For example, consider the weighted average treatment effect:

$$\tau_{wate} = \frac{\mathbf{E}[g(X)\{Y(1) - Y(0)\}]}{\mathbf{E}[g(X)]} = \frac{\mathbf{E}[g(X)Y|D = 1] - \mathbf{E}[g(X)Y|D = 0]}{\mathbf{E}[g(X)]}. \quad (9)$$

As pointed out by HIR, the weighted average treatment effect is reduced to τ_{atet} when we take $g(X) = p_1(X)$.

In treatment-based sampling, the weighted average treatment effect is not identified without knowledge of the marginal probabilities $p_{d,w}$, because the marginal distribution of X is not identified from the data in this case. However, when the sampling is pure treatment-based sampling, we can identify the average treatment effect on the treated:

$$\tau_{atet} = \mathbf{E}[Y_1 - Y_0|D = 1] = \mathbf{E}[\mathbf{E}[Y_1|X, D = 1] - \mathbf{E}[Y_0|X, D = 0]|D = 1], \quad (10)$$

without knowledge of p_d . In fact, one can show that under (8), 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 .

When the treatment-based sampling has an additional dimension for the strata, i.e., \mathcal{W} , here, the knowledge of $p_{d,w}$ is required for identification of τ_{atet} . This is because the conditional distribution of X given $D = 1$ is not identified from the observations from treatment-based sampling without knowledge of $p_{d,w}$.

3.2 Efficiency Bound for Weighted Average Treatment Effect

In this section, we establish the semiparametric efficiency bound for τ_{wate} under treatment-based sampling. The efficiency bound under treatment-based sampling can be different from that of Hahn (1998) or HIR for three reasons. First, the observations are from treatment-based sampling, not from random sampling. Second, the procedure in this paper assumes that we know marginal probabilities $p_{d,w}$ (except for τ_{atet} under pure treatment-based sampling) while Hahn (1998) or HIR do not assume it. Third, the unconfoundedness condition (8) is imposed on the original data set, not on the observations from treatment-based sampling. The unconfoundedness condition affects the semiparametric efficiency bound here, but not in the same way when the conditions are imposed directly on the observations.

In the computation of the efficiency bound, we do not assume that the propensity scores, either $p_1(X)$ under P or $q_1(X)$ under Q , are known, as this is not plausible in practice. While τ_{atet} can be identified as τ_{wate} with $g(X) = p_1(X)$ as noted by HIR, we need to identify it separately because for τ_{wate} , $g(X)$ is assumed to be known, while for τ_{atet} , $g(X) = p_1(X)$ is

not. Furthermore, in the case of pure treatment-based sampling, the identification of τ_{atet} does not require knowledge of the aggregate shares p_d , in contrast to τ_{wate} . Hence we present the results for τ_{ate} in a separate subsection that follows. We introduce some notations:

$$\begin{aligned}\beta_d(X) &\equiv \mathbf{E}[Y_d|X] \\ \sigma_d^2(X) &\equiv \mathbf{E}[(Y_d - \beta_d(X))^2|X], \text{ and} \\ \tau(X) &\equiv \mathbf{E}[Y_1|X] - \mathbf{E}[Y_0|X].\end{aligned}$$

Theorem 1 : *Under (8), the semiparametric efficiency bound for τ_{wate} under treatment-based sampling is equal to $V_{TS}^{-1}(\tau_{wate})$, where*

$$V_{TS}(\tau_{wate}) \equiv \frac{1}{\{\mathbf{E}[g(X)]\}^2} \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \frac{p_{d,w}^2}{q_{d,w}} \mathbf{E}_{d,w} \left[g(X)^2 \frac{\sigma_d^2(X)}{p_d^2(X)} + \zeta_d^2(X) \right],$$

and $\zeta_d(x) \equiv g(x)\{\tau(x) - \tau_{wate}\} - \mathbf{E}_{d,w}[g(X)\{\tau(X) - \tau_{wate}\}]$ with $x \equiv (v, w)$. In particular, when sampling is pure treatment-based sampling and $p_d = q_d$, $V_{TS}(\tau_{atet}) = V_{RS}(\tau_{atet})$, where

$$V_{RS}(\tau_{wate}) \equiv \frac{1}{\{\mathbf{E}[g(X)]\}^2} \mathbf{E} \left[g(X)^2 \left\{ \frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} \right\} + \sum_{d \in \mathcal{D}} \zeta_d^2(X) p_d(X) \right].$$

The results of Theorem 1 show that knowledge of $p_{d,w}$ is not ancillary in general. In the special case of pure treatment-based sampling where the sampling is random sampling, i.e., $p_d = q_d$, we can compare our results with that of HIR who found the variance bound to be

$$V_{HIR}(\tau_{wate}) \equiv \frac{1}{\{\mathbf{E}[g(X)]\}^2} \mathbf{E} \left[g(X)^2 \left\{ \frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} \right\} + g^2(X)(\tau(X) - \tau_{wate})^2 \right],$$

where $p_d(X) \equiv P\{D = d|X\}$, $\sigma_d^2(X) \equiv \mathbf{E}[(Y_d - \beta_d(X))^2|X]$, and $\beta_d(X) \equiv \mathbf{E}[Y_d|X]$. Therefore, $V_{TS}(\tau_{wate}) \leq V_{HIR}(\tau_{wate})$ and the equality holds if and only if

$$\mathbf{E}_d[g(X)\{\tau(X) - \tau_{wate}\}] = 0 \text{ for all } d \in \mathcal{D}. \quad (11)$$

This result implies that knowledge of marginal probabilities p_d is not ancillary for τ_{wate} .

The ancillarity of the aggregate shares is not merely a matter of a theoretical concern. The ancillarity is closely related to the question of whether the estimators of Hahn (1998) or HIR could serve as a guidance for efficient estimators under treatment-based sampling. The non-ancillarity of the aggregate shares in Theorem 1 suggests that the answer will be negative because the knowledge of the aggregate shares was not assumed in Hahn (1998) or

HIR. This will be confirmed later when we develop efficient estimators.

3.3 Efficiency Bound for Treatment Effects on the Treated

Let us turn to τ_{atet} . The following theorem offers the semiparametric efficiency bound under treatment-based sampling.

Theorem 2 : (i) *Suppose that (8) holds and $\{p_{d,w}\}_{(d,w) \in \mathcal{D} \times \mathcal{W}}$ are known. Then the semiparametric efficiency bound for τ_{atet} under treatment-based sampling is equal to $V_{TS}^{-1}(\tau_{atet})$, where*

$$V_{TS}(\tau_{atet}) \equiv \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \frac{p_{d,w}^2}{q_{d,w}} \mathbf{E}_{d,w} \left[\frac{d}{p_1^2} \left\{ \sigma_1^2(X) + \tilde{\zeta}_1^2(X) \right\} + \frac{1-d}{p_1^2} \frac{\sigma_0^2(X) p_1^2(X)}{p_0^2(X)} \right]$$

and $\tilde{\zeta}_d(x) \equiv \tau(x) - \tau_{atet} - \mathbf{E}_{d,w} [\tau(X) - \tau_{atet}]$ with $x \equiv (v, w)$.

(ii) *Suppose that (8) holds and the sampling is pure treatment-based sampling. Then, regardless of whether we know $\{p_d\}_{d \in \mathcal{D}}$ or not, the semiparametric efficiency bound for τ_{atet} is given by $V_{PTS}^{-1}(\tau_{atet})$, where*

$$V_{PTS}(\tau_{atet}) = \frac{1}{q_1} \mathbf{E} \left[\sigma_1^2(X) + \{\tau(X) - \tau_{atet}\}^2 | D = 1 \right] + \frac{1}{q_0} \mathbf{E} \left[\frac{f(X|1)^2 \sigma_0^2(X)}{f(X|0)^2} | D = 0 \right]. \quad (12)$$

Under random sampling where $p_{d,w} = q_{d,w}$, $V_{TS}(\tau_{atet})$ is smaller than the variance bound of Hahn (1998) that does not assume knowledge of $p_{d,w}$. Therefore, the aggregate shares are not ancillary in general. However, the situation becomes different when the sampling is pure treatment-based sampling. In this case, the aggregate shares p_d are ancillary. Indeed, in pure treatment-based sampling with $p_d = q_d$, the variance bound is reduced to

$$V_{RS}(\tau_{atet}) \equiv \mathbf{E} \left[\left\{ \frac{p_1(X) \sigma_1^2(X)}{p_1^2} + \frac{\sigma_0^2(X) p_1^2(X)}{p_0(X) p_1^2} \right\} + \frac{\{\tau(X) - \tau_{atet}\}^2 p_1(X)}{p_1^2} \right]$$

which is nothing but the variance bound of Hahn (1998) for τ_{atet} . Therefore, the variance bound in (12) can be viewed as a generalization of the variance bound of Hahn (1998) to pure treatment-based sampling.

3.4 Optimal Design of Treatment-Based Sampling

The semiparametric efficiency bounds depend on the sampling design $q_{d,w}$, and as discussed before, we can develop an optimal design of treatment-based sampling. Theorems 1 and 2

allow us to identify $J_{d,w}$ in (5) in this context of estimating average treatment effects.

Corollary 1 : *Suppose that we are under the conditions of Theorems 1 and 2. Then the optimal choice of $q_{d,w}(\tau_{wate})$ for τ_{wate} and $q_{d,w}(\tau_{atet})$ for τ_{atet} are given by*

$$q_{d,w}(\tau_{wate}) = \frac{\sqrt{J_{d,w}^*}}{\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \sqrt{J_{d,w}^*}} \text{ and } q_{d,w}(\tau_{atet}) = \frac{\sqrt{\tilde{J}_{d,w}}}{\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \sqrt{\tilde{J}_{d,w}}},$$

where

$$\begin{aligned} J_{d,w}^* &= \frac{p_{d,w}^2}{\{\mathbf{E}[g(X)]\}^2} \mathbf{E}_{d,w} \left[g(X)^2 \frac{\sigma_d^2(X)}{p_d^2(X)} + \zeta_d^2(X) \right] \text{ and} \\ \tilde{J}_{d,w} &= p_{d,w}^2 \mathbf{E}_{d,w} \left[\frac{d}{p_1^2} \left\{ \sigma_1^2(X) + \tilde{\zeta}_1^2(X) \right\} + \frac{1-d}{p_1^2} \frac{\sigma_0^2(X) p_1^2(X)}{p_0^2(X)} \right]. \end{aligned}$$

We can apply the discussions in Section 2.3 in this situation. The optimal treatment-based sampling for τ_{wate} is reduced to random sampling if $J_{d,w}$ is the same for all $(d, w) \in \mathcal{D} \times \mathcal{W}$. Even if the marginal probability $p_{d,w}$ is relatively small, we do not necessarily have to sample relatively more from the subsample $(D, W) = (d, w)$ if the information from the subsample is strong enough.

In the case of pure treatment-based sampling, the estimation of the optimal design does not require knowledge of p_d . Indeed, we define

$$\bar{J}_1 = \mathbf{E} \left[\sigma_1^2(X) + \{\tau(X) - \tau_{atet}\}^2 | D = 1 \right] \text{ and } \bar{J}_0 = \mathbf{E} \left[\frac{f(X|1)^2 \sigma_0^2(X)}{f(X|0)^2} | D = 0 \right].$$

Then, $V_{PTS}(\tau_{atet}) = \bar{J}_1/q_1 + \bar{J}_0/q_0$. The optimal design of q_1 in Corollary 1 is given by

$$q_1(\tau_{atet}) = \frac{\sqrt{\bar{J}_1}}{\sqrt{\bar{J}_1} + \sqrt{\bar{J}_0}}$$

and the necessary and sufficient condition for $V_{PTS}(\tau_{atet}) \leq V_{PRS}(\tau_{atet})$ is given by as a condition in (7) with J_1 and J_0 replaced by \bar{J}_1 and \bar{J}_0 . Note that estimation of \bar{J}_d does not require knowledge of the aggregate shares p_d .

4 Efficient Estimation of Weighted Average Treatment Effects

4.1 Propensity Score Estimation

In this section, we focus on the propensity score. Let $f(x)$ be the density of X with respect to some σ -finite measure, and $f(v|d, w)$ be the conditional density function of V given $(D, W) = (d, w)$. By the Bayes' rule, the propensity score is identified as

$$p_d(v, w) = \frac{f(v|d, w)p_{d,w}}{\sum_{d \in \mathcal{D}} f(v|d, w)p_{d,w}}. \quad (13)$$

While we can identify $f(v|d, w)$ nonparametrically using the $(D, W) = (d, w)$ subsamples in the treatment-based sample, the identification of $p_d(v, w)$ certainly requires knowledge of $p_{d,w}$.

We consider two consistent estimators of the propensity score that are based on the identification in (13). Let $X = (V_1, X_2) \in \mathbf{R}^L$ and $X_2 = (V_2, W)$, where $V_1 \in \mathbf{R}^{L_1}$ is continuous and $X_2 \in \mathbf{R}^{L_2}$ is discrete. Let $S_{d,w} = \{1 \leq i \leq n : (D_i, W_i) = (d, w)\}$. Define $\hat{f}(v_1, v_2|d, w) = \frac{1}{q_{d,w}n} \sum_{i \in S_{d,w}} K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}$, where $K_h(s_1, \dots, s_{L_1}) = K(s_1/h, \dots, s_{L_1}/h)/h^{L_1}$ and $K(\cdot)$ is a multivariate kernel function. Then, we define the estimator of the propensity score $p_d(x_1, x_2)$ as

$$\hat{p}_d(v, w) = \frac{\hat{f}(v_1, v_2|d, w)p_{d,w}}{\sum_{d \in \mathcal{D}} \hat{f}(v_1, v_2|d, w)p_{d,w}}. \quad (14)$$

Let $L_{d,w,i} \equiv \frac{p_{d,w}}{q_{d,w}} 1\{(D_i, W_i) = (d, w)\}$, and $L_{w,i} \equiv L_{1,w,i} + L_{2,w,i}$. We can rewrite the estimator as

$$\hat{p}_d(v, w) = \frac{\sum_{i=1}^n L_{d,w,i} K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}}{\sum_{i=1}^n L_{w,i} K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}}.$$

Therefore, the propensity score estimator is a weighted Nadaraya-Watson estimator. This is intuitive because the probability under treatment-based sampling is the average of conditional probabilities using different weights.

Alternatively, we can estimate the propensity score using the estimated fraction $\hat{q}_{d,w} = \frac{1}{n} \sum_{i=1}^n 1\{(D_i, W_i) = (d, w)\} = n_{d,w}/n$ in place of $q_{d,w}$. Using this, we define

$$\tilde{p}_d(v, w) \equiv \frac{\sum_{i=1}^n \hat{L}_{d,w,i} K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}}{\sum_{i=1}^n \hat{L}_{w,i} K_h(V_{1i} - v_1) 1\{V_{2i} = v_2\}}, \quad (15)$$

where $\hat{L}_{d,w,i} \equiv \frac{p_{d,w}}{\hat{q}_{d,w}} 1\{(D_i, W_i) = (d, w)\}$ and $\hat{L}_{w,i} \equiv \hat{L}_{1,w,i} + \hat{L}_{2,w,i}$.

4.2 Efficient Estimation of Weighted Average Treatment Effect

In this section, we search for an efficient estimator. The first idea in this pursuit will be that we obtain by adapting the estimator of HIR to treatment-based sampling:

$$\hat{\tau}_{wate} \equiv \frac{\sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w}}{q_{1,w}} \frac{1}{n} \sum_{i \in S_{1,w}} g(V_i, w) Y_i / \hat{p}_1(V_i, w) - \frac{p_{0,w}}{q_{0,w}} \frac{1}{n} \sum_{i \in S_{0,w}} g(V_i, w) Y_i / \hat{p}_0(V_i, w) \right\}}{\sum_{w \in \mathcal{W}} \left\{ \frac{p_{1,w}}{q_{1,w}} \frac{1}{n} \sum_{i \in S_{1,w}} g(V_i, w) + \frac{p_{0,w}}{q_{0,w}} \frac{1}{n} \sum_{i \in S_{0,w}} g(V_i, w) \right\}},$$

where $\hat{p}_d(v, w)$ is estimated by (14). Observe that when $p_{d,w} = q_{d,w}$ and \mathcal{W} is a singleton, the estimator $\hat{\tau}_{wate}$ is precisely reduced to the estimator of HIR except with a different nonparametric estimator for the propensity score. Therefore, this estimator is a generalization of the estimator of HIR to the treatment-based sampling. In Theorem 2 below, we show that this estimator is consistent and asymptotically normal, but inefficient.

Alternatively, we suggest the following estimator:

$$\tilde{\tau}_{wate} \equiv \frac{\sum_{w \in \mathcal{W}} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) Y_i / \tilde{p}_1(V_i, w)}{\sum_{w \in \mathcal{W}} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} g(V_i, w) / \tilde{p}_1(V_i, w)} - \frac{\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) Y_i / \tilde{p}_0(V_i, w)}{\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} g(V_i, w) / \tilde{p}_0(V_i, w)},$$

where $\tilde{p}_d(v, w)$ is as in (15). The estimator $\tilde{\tau}_{wate}$ involves a further weighting of $g(V_i, w)$ by $\tilde{p}_d(V_i, w)$. It is worth noting that the estimator $\hat{\tau}_{wate}$ uses the true marginal probability $q_{d,w}$ under Q while the estimator $\tilde{\tau}_{wate}$ uses its estimator $\hat{q}_d = n_d/n$.

We formalize the results that have been discussed so far. We introduce the following assumptions. Let \mathcal{X} be the support of X and $f(\cdot)$ be its density with respect to a σ -finite measure.

Assumption 1 : For $[a, b] \subset (0, 1)$, $p_1(x) \in [a, b]$ for all x in the set $\{x \in \mathcal{X} : g(x) \neq 0\}$.

Assumption 2 : (i) (a) $f(\cdot)$ is bounded away from zero on \mathcal{X} .

(b) $f(\cdot, x_2)$, $p_1(\cdot, x_2)$, $\beta_0(\cdot, x_2)$, $\beta_1(\cdot, x_2)$, and $g(\cdot, x_2)$ are $L_1 + 1$ times continuously differentiable with bounded derivatives.

(ii)(a) $\mathbf{E}Y_1^2 < \infty$ and $\mathbf{E}Y_0^2 < \infty$, (b) for either $d = 1$ or $d = 0$, $\sup_{x \in \mathcal{X}} p_d(x) \|x\|^{L_1} < \infty$, and (c) $|g(\cdot)|$ is bounded.

(iii) $p_{d,w} \in (0, 1)$ and $q_{d,w} \in (0, 1)$ for all $(d, w) \in \mathcal{D} \times \mathcal{W}$ and $\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} p_{d,w} = 1$ and $\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} q_{d,w} = 1$.

Assumption 3 : (i) K is zero outside an interior of a bounded set, $L_1 + 1$ times continuously differentiable with bounded derivatives, $\int K(s) ds = 1$, and $\int s_1^{l_1} \cdots s_{L_1}^{l_{L_1}} K(s) ds = 0$ for all nonnegative integers l_1, \dots, l_{L_1} such that $l_1 + \dots + l_{L_1} \leq L_1$ and $\int |s_1^{l_1} \cdots s_{L_1}^{l_{L_1}} K(s)| ds < \infty$ for all nonnegative integers l_1, \dots, l_{L_1} such that $l_1 + \dots + l_{L_1} = L_1 + 1$.

(ii) $\sqrt{n^{-1/2}h^{-L_1} \log n} + n^{1/2}h^{L_1+1} \rightarrow 0$ as $n \rightarrow \infty$.

Assumption 1 is the condition of sample overlap needed for the identification of τ_{wate} . This is violated when $g(X) = 1$ and part of X is only observed among the treated or untreated subsamples. (See Heckman, Ichimura, and Todd (1997) for a discussion in this regard.) Assumption 2(b) controls the tail behavior of either $p_1(x)$ or $p_0(x)$. This condition is satisfied when \mathcal{X} is bounded. Assumption 3(i) is a standard assumption for a higher order kernel. The following theorem establishes the asymptotic distribution of $\hat{\tau}_{wate}$ and $\tilde{\tau}_{wate}$.

Theorem 3 : *Suppose that the condition (8) and Assumptions 1-3 hold. Then*

$$\begin{aligned} \sqrt{n}(\hat{\tau}_{wate} - \tau_{wate}) &\rightarrow_d N(0, V_1), \text{ and} \\ \sqrt{n}(\tilde{\tau}_{wate} - \tau_{wate}) &\rightarrow_d N(0, V_{TS}(\tau_{wate})), \end{aligned}$$

where

$$V_1 \equiv \frac{1}{\{\mathbf{E}[g(X)]\}^2} \sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \frac{p_{d,w}^2}{q_{d,w}} \mathbf{E}_{d,w} \left[g(X)^2 \left\{ \frac{\sigma_d^2(X)}{p_d^2(X)} + (\tau(X) - \tau_{wate})^2 \right\} \right].$$

When the treatment-based sampling is random sampling, the asymptotic variance V of $\hat{\tau}_{wate}$ is reduced to V_{HIR} which we saw that it is less than $V_{TS}(\tau_{wate})$. Therefore, the result of Theorem 3 shows that even when the estimator of HIR is modified to accommodate treatment-based sampling, the estimator is still inefficient. The main reason for the inefficiency seems to lie in the fact that the aggregate shares p_d are not ancillary. Indeed, when the sampling is random sampling, V_1 is reduced to $V_{HIR}(\tau_{wate})$. The efficiency is achieved by an alternative estimator $\tilde{\tau}_{wate}$. The efficient estimator uses estimated fractions $\hat{q}_{d,w}$ and hence can also be used when only the estimated fraction $\hat{q}_{d,w} = n_{d,w}/n$ is available in the data.

4.3 Efficient Estimation of Average Treatment Effect on the Treated

Let us turn to efficient estimation of τ_{atet} . In this case, the identification of τ_{atet} allows us to formulate Assumption 1 differently:

Assumption 1P : For $[a, b] \subset (0, 1)$, $p_1(x) \in [a, b]$ for all $x \in \mathcal{X}$.

This assumption is weaker than Assumption 1 when $g(X) = 1$. We suggest the following

estimator:

$$\tilde{\tau}_{atet} = \frac{1}{p_1} \sum_{w \in \mathcal{W}} \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} Y_i - \frac{\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{p}_1(V_i, w) Y_i / \tilde{p}_0(V_i, w)}{\sum_{w \in \mathcal{W}} \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \tilde{p}_1(V_i, w) / \tilde{p}_0(V_i, w)},$$

where $\tilde{p}_d(x)$ is estimated by (15). We will show that this estimator is efficient.

We saw that in the case of pure treatment-based sampling, the knowledge of p_d is ancillary. One might consider alternatively the estimator of HIR that is adapted to pure treatment-based sampling:

$$\hat{\tau}_{atet,p} = \frac{\frac{p_1}{q_1 n} \sum_{i \in S_1} Y_i - \frac{p_0}{q_0 n} \sum_{i \in S_0} \hat{p}_1(X_i) Y_i / \hat{p}_0(X_i)}{\frac{p_0}{q_0 n} \sum_{i \in S_0} \hat{p}_1(X_i) + \frac{p_1}{q_1 n} \sum_{i \in S_1} \hat{p}_1(X_i)}.$$

The estimator reduces to that of HIR when the sampling is random sampling, i.e., $p_d = q_d$. In a theorem below, we will show that this estimator is efficient. However, this estimator requires knowledge of the aggregate shares p_d . Instead, we suggest the following estimator that does not require knowledge of the aggregate shares in this case.

$$\begin{aligned} \tilde{\tau}_{atet,p} &= \frac{1}{n_1} \sum_{i \in S_1} Y_i - \frac{\sum_{i \in S_0} \tilde{p}_1(X_i) Y_i / \tilde{p}_0(X_i)}{\sum_{i \in S_0} \tilde{p}_1(X_i) / \tilde{p}_0(X_i)} \\ &= \frac{1}{n_1} \sum_{i \in S_1} Y_i - \frac{\sum_{i \in S_0} Y_i \left(\frac{\frac{1}{n_1} \sum_{j \in S_1} K_{ji}}{\frac{1}{n_0} \sum_{j \in S_0} K_{ji}} \right)}{\sum_{i \in S_0} \left(\frac{\frac{1}{n_1} \sum_{j \in S_1} K_{ji}}{\frac{1}{n_0} \sum_{j \in S_0} K_{ji}} \right)}, \end{aligned}$$

where $K_{ji} = K_h(V_{1j} - V_{1i}) 1\{V_{2j} = V_{2i}\}$. The estimator $\tilde{\tau}_{atet,p}$ is in fact an estimator $\tilde{\tau}_{atet}$ that is specialized to pure treatment-based sampling. Hence the estimator is also efficient.

Theorem 4 : *Suppose that the condition (8) and Assumptions 1P, 2-3 hold. Then,*

$$\sqrt{n}(\tilde{\tau}_{atet} - \tau_{atet}) \rightarrow_d N(0, V_{TS}(\tau_{atet})).$$

Suppose further that we are under pure treatment-based sampling. Then

$$\begin{aligned} \sqrt{n}(\tilde{\tau}_{atet,p} - \tau_{atet}) &\rightarrow_d N(0, V_{PTS}(\tau_{atet})) \text{ and} \\ \sqrt{n}(\hat{\tau}_{atet,p} - \tau_{atet}) &\rightarrow_d N(0, V_{PTS}(\tau_{atet})). \end{aligned}$$

5 Conclusion

This paper has established semiparametric efficiency bounds for certain average treatment effects parameters under treatment-based sampling. This paper has also developed an opti-

mal design of treatment-based sampling. The theory of optimal design illuminates the role of treatment-based sampling in improving the quality of efficient inference. Lastly, this paper has suggested efficient estimators. This paper's finding suggests that under treatment-based sampling, tailoring the estimators of HIR to treatment-based sampling works only when the aggregate shares are ancillary.

6 Appendix: Proofs

Proof of Theorem 1 : We need to find $\dot{\psi}_Q^e(z, d, w)$. Let $\mathcal{Q} = \{f_{Z|D,W}(\cdot|\cdot)q : f_{Z|D,W} \in \mathcal{P}_{d,w}, (d, w) \in \mathcal{D} \times \mathcal{W}\}$ and fix $Q \in \mathcal{Q}$. Let $f(z|d, w)$ be the conditional density of $Z = (Y, V)$ given $(D, W) = (d, w)$ and let $z = (y, v)$. We use subscripts P and Q for densities to make it explicit under which probability they are defined when they differ. We do not use the subscripts for the conditional densities given $(D, W) = (d, w)$ or given $(D, W, V) = (d, w, v)$ because they are identical both under P and under Q .

We write the density $f_Q(y, v, d, w)$ of (Y, V, D, W) under Q as $f_{d,P}(y|x)f(v|d, w)q_{d,w}$ where $f_{d,P}(y|x)$ is the conditional density of Y_{di} given $X_i = x$ under P . The second equality follows by the unconfoundedness condition. Hence the score $s(y, v, d, w)$ is written as $s_d(y|x) + s(v|d, w)$, where $\int s_d(y|x)f_{d,P}(y|x)dy = 0$ and $\int s(v|d, w)f(v|d, w)dv = 0$. The closed linear span of such scores constitutes the tangent space \mathcal{T} .

Take a regular parametric submodel $f_Q^t(y, v, d, w) = f^t(y, v|d, w)q_{d,w}$ and let P_t be the parametric submodel with density $f^t(y, v|d, w)p_{d,w}$. We need to find $\dot{\psi}_P$. The weighted average treatment effect under P_t is written as

$$\tau_{wate}(t) = \frac{\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} \int g(v, w) \{ \int y f_{1,t}(y|v, w) dy - \int y f_{0,t}(y|v, w) dy \} p_{d,w} f_t(v|d, w) dv}{\sum_{(d,w) \in \mathcal{D} \times \mathcal{W}} p_{d,w} \int g(v, w) f_t(v|d, w) dv}.$$

The first order derivative of $\tau_{wate}(t)$ with respect to t at $t = 0$ is equal to

$$\begin{aligned} & \frac{1}{\mathbf{E}[g(X)]} \mathbf{E}[g(X) (\mathbf{E}[Y s_1(Y|X)|X] - \mathbf{E}[Y s_0(Y|X)|X])] \\ & - \frac{1}{\mathbf{E}[g(X)]} \mathbf{E}[s(V|D, W)g(X)\{\tau(X) - \tau_{wate}\}]. \end{aligned}$$

Let

$$\begin{aligned} \dot{\psi}_P(y, v, d, w) &= \frac{1}{\mathbf{E}[g(X)]} g(v, w) \left(\frac{d(y - \beta_1(v, w))}{p_1(v, w)} - \frac{(1-d)(y - \beta_0(v, w))}{p_0(v, w)} \right) \\ & - \frac{1}{\mathbf{E}[g(X)]} \zeta_d(v, w). \end{aligned} \quad (16)$$

We can write

$$\frac{\partial \tau_{wate}(t)}{\partial t} = \mathbf{E}[\dot{\psi}_P(Y, V, D, W)s(Y, V, D, W)] = \mathbf{E}_Q \left[\dot{\psi}_Q(Y, V, D, W)s(Y, V, D, W) \right], \quad (17)$$

where $\dot{\psi}_Q(y, v, d, w) = \dot{\psi}_P(y, v, d, w)p_{d,w}/q_{d,w}$. Now, observe that $\dot{\psi}_Q$ belongs to the tangent space \mathcal{T} . (This follows from the unconfounded condition.) Hence the variance bound is given by its $L_2(Q)$ -norm. ■

Proof of Theorem 2 : The tangent space in the proof of Theorem 1 remains the same. The only needed change from Theorem 1 for this case is the computation of the influence function because now $g(x) = p_1(x)$ is not assumed to be known. Let P_t be the submodel as in the proof of Theorem 1. The weighted average treatment effect under P_t is written as

$$\tau_{atet}(t) = \sum_{w \in \mathcal{W}} \int \int y \{f_t(y|v, 1, w) - f_t(y|v, 0, w)\} dy f_t(v|1, w) p_{w|1} dv,$$

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

$$\begin{aligned} & \mathbf{E} [s(V|D, W)\{\tau(X) - \tau_{atet}\} | D = 1] \\ & + \mathbf{E} [\mathbf{E} [Y s_1(Y|X) | X, D = 1] - \mathbf{E} [Y s_0(Y|X) | X, D = 0] | D = 1]. \end{aligned}$$

Therefore, we take

$$\dot{\psi}_P(y, v, d, w) = \frac{1}{p_1} \left\{ d(y - \beta_1(v, w) - \tilde{\zeta}_1(v, w)) - \frac{p_1(v, w)(1-d)(y - \beta_0(v, w))}{p_0(v, w)} \right\}.$$

As shown in the proof of Theorem 1, this yields the semiparametric efficiency bound for τ_{atet} .

Let us turn to the situation with pure treatment-based sampling. The tangent space is the closed linear span of scores of the form $s_d(y|x) + s(v|d)$, where $\int s_d(y|x) f_{d,P}(y|x) dy = 0$ and $\int s(v|d) f(v|d) dx = 0$. Write

$$\tau_{atet}(t) = \int \int y \{f_t(y|x, 1) - f_t(y|x, 0)\} dy f_t(x|1) dx.$$

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

$$\begin{aligned} & \mathbf{E} [s(X|D)\{\tau(X) - \tau_{atet}\} | D = 1] \\ & + \mathbf{E} [\{\mathbf{E} [Y s_1(Y|X) | X, D = 1] - \mathbf{E} [Y s_0(Y|X) | X, D = 0]\} | D = 1]. \end{aligned}$$

Therefore, we take

$$\dot{\psi}_P(y, x, d) = \left\{ \frac{d(y - \beta_1(x) - \{\tau(x) - \tau_{atet}\})}{p_1} - \frac{p_1(x)(1-d)(y - \beta_0(x))}{p_0(x)p_1} \right\}$$

because $\mathbf{E}[\tau(X) - \tau_{atet}|D = 1] = 0$. Let $\dot{\psi}_Q(y, x, d) = \dot{\psi}_P(y, x, d)p_d/q_d$. Now

$$\begin{aligned} \sum_{d \in \mathcal{D}} q_d \mathbf{E} \left[\dot{\psi}_Q^2(Y, X, D) | D = d \right] &= \frac{1}{q_1} \mathbf{E} \left[(Y_1 - \beta_1(X) - \{\tau(X) - \tau_{atet}\})^2 | D = 1 \right] \\ &+ \frac{1}{q_0} \mathbf{E} \left[\frac{p_0^2 p_1(X)^2}{p_0(X)^2 p_1^2} (Y_0 - \beta_0(X))^2 | D = 0 \right]. \end{aligned}$$

By Bayes' rule, $p_0 p_1(X)/(p_1 p_0(X)) = f(X|1)/f(X|0)$, and hence plugging in this, we obtain the wanted result. ■

Lemma A1 : Suppose that S_i is a random variable such that $\mathbf{E}[S_i^2] < \infty$ and $\mathbf{E}[S_i|V_{1i} = \cdot, X_{2i} = x_2]$ is $L_1 + 1$ times continuously differentiable with bounded derivatives.

(i) Suppose that the assumptions of Theorem 3 hold. Then, for $d = 0, 1$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n S_i (p_d(V_i, w) - \hat{p}_d(V_i, w)) \\ &= -\frac{1}{n} \sum_{i=1}^n \frac{\mathbf{E}_Q[S_i|V_i, W_i = w] \mathcal{J}_{d,w,i}}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} \\ &+ \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{E}_Q[S_i|V_i, W_i = w] p_d(V_i, w) \mathcal{J}_{w,i}}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} + o_p(n^{-1/2}), \end{aligned}$$

where $\mathcal{J}_{d,w,i} \equiv L_{d,w,i} - \mathbf{E}_Q[L_{d,w,i}|V_i, W_i = w]$ and $\mathcal{J}_{w,i} = \mathcal{J}_{1,w,i} + \mathcal{J}_{0,w,i}$.

(ii) Suppose that the assumptions of Theorem 4 hold. Then,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n S_i (\hat{p}_1(V_i, w) - \tilde{p}_1(V_i, w)) \\ &= \mathbf{E}_Q [p_0(V_i, w)p_1(V_i, w)S_i] \left(\frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_p(n^{-1/2}). \end{aligned}$$

Proof of Lemma A1 : (i) Observe that 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

$$\begin{aligned} p_1(V_i, w) &= \frac{f(V_i|1, w)p_{1,w}}{f(V_i|1, w)p_{1,w} + f(V_i|0, w)p_{0,w}} \\ &= \frac{(q_1(V_i, w)/q_{1,w})p_{1,w}}{\sum_{d \in \mathcal{D}} (q_d(V_i, w)/q_{d,w})p_{d,w}} = \frac{\mathbf{E}_Q[L_{1,w,i}|V_i, W_i = w]}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]}. \end{aligned} \quad (18)$$

Let $K_{ji} = K_h(V_{1j} - V_{1i})1\{V_{2j} = V_{2i}\}$ for brevity. Observe that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n S_i (p_1(V_i, w) - \hat{p}_1(V_i, w)) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{S_i}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} \left\{ \mathbf{E}_Q[L_{1,w,i}|V_i, W_i = w] - \frac{\sum_{j=1, j \neq i}^n L_{1,w,j} K_{ji}}{\sum_{j=1, j \neq i}^n K_{ji}} \right\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n S_i \left\{ \frac{\sum_{j=1, j \neq i}^n L_{1,w,j} K_{ji}}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w] \sum_{j=1, j \neq i}^n K_{ji}} - \frac{\sum_{j=1, j \neq i}^n L_{1,w,j} K_{ji}}{\sum_{j=1, j \neq i}^n L_{w,j} K_{ji}} \right\} + o_p(n^{-1/2}). \end{aligned} \quad (19)$$

We write the last sum as

$$-\frac{1}{n} \sum_{i=1}^n \frac{S_i p_1(V_i, w)}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} \left\{ \mathbf{E}_Q[L_{w,i}|V_i, W_i = w] - \frac{\sum_{j=1, j \neq i}^n L_{w,j} K_{ji}}{\sum_{j=1, j \neq i}^n K_{ji}} \right\} + o_p(n^{-1/2})$$

using (18). By Lemma B1 below, the last two terms in (19) are asymptotically equivalent to (up to $o_p(n^{-1/2})$)

$$-\frac{1}{n} \sum_{i=1}^n \frac{\mathbf{E}_Q[S_i|V_i, W_i = w] \mathcal{J}_{1,w,i}}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} + \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{E}_Q[S_i|V_i, W_i = w] p_1(V_i, w) \mathcal{J}_{w,i}}{\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]}$$

using the definitions of $\mathcal{J}_{1,w,i}$ and $\mathcal{J}_{w,i}$.

(ii) First, we write $\hat{p}(X_i) - \tilde{p}(X_i)$ as

$$\begin{aligned} &\frac{\sum_{j=1}^n \{L_{1,w,j} - \hat{L}_{1,w,j}\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji}} + \sum_{j=1}^n \hat{L}_{1,w,j} K_{ji} \left\{ \frac{1}{\sum_{j=1}^n L_{w,j} K_{ji}} - \frac{1}{\sum_{j=1}^n \hat{L}_{w,j} K_{ji}} \right\} \\ &= -\frac{\sum_{j=1}^n L_{0,w,j} K_{ji} \sum_{j=1}^n \{\hat{L}_{1,w,j} - L_{1,w,j}\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji} \sum_{j=1}^n L_{w,j} K_{ji}} \\ &\quad + \frac{\sum_{j=1}^n L_{1,w,j} K_{ji} \sum_{j=1}^n \{\hat{L}_{0,w,j} - L_{0,w,j}\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji} \sum_{j=1}^n L_{w,j} K_{ji}} + o_p(n^{-1/2}). \end{aligned}$$

Therefore, we can write $\hat{p}(X_i) - \tilde{p}(X_i)$ as

$$\begin{aligned} & \frac{p_1(V_i, w) \sum_{j=1}^n \left\{ \hat{L}_{0,w,j} - L_{0,w,j} \right\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji}} \\ & - \frac{p_0(V_i, w) \sum_{j=1}^n \left\{ \hat{L}_{1,w,j} - L_{1,w,j} \right\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji}} + o_p(n^{-1/2}) \end{aligned}$$

As for the last term, we can write $\frac{1}{n} \sum_{j=1}^n \left\{ \hat{L}_{1,w,j} - L_{1,w,j} \right\} K_{ji} / \sum_{j=1}^n L_{w,j} K_{ji}$ as

$$\begin{aligned} & \left(\frac{p_{1,w}}{\hat{q}_{1,w}} - \frac{p_{1,w}}{q_{1,w}} \right) \frac{\sum_{j=1}^n 1\{(D_i, W_i) = (1, w)\} K_{ji}}{\sum_{j=1}^n L_{w,j} K_{ji}} \\ = & \left(\frac{p_{1,w}}{\hat{q}_{1,w}} - \frac{p_{1,w}}{q_{1,w}} \right) \frac{q_1(V_i, w)}{q_1(V_i, w)p_{1,w}/q_{1,w} + q_0(V_i, w)p_{0,w}/q_{0,w}} + o_p(n^{-1/2}) \\ = & \left(\frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} \right) \frac{q_1(V_i, w)p_{1,w}/q_{1,w}}{q_1(V_i, w)p_{1,w}/q_{1,w} + q_0(V_i, w)p_{0,w}/q_{0,w}} + o_p(n^{-1/2}) \\ = & \left(\frac{q_{1,w} - \hat{q}_{1,w}}{q_{1,w}} \right) p_1(V_i, w) + o_p(n^{-1/2}) \text{ (by (18).)} \end{aligned}$$

Dealing with $\sum_{j=1}^n \left\{ \hat{L}_{0,w,j} - L_{0,w,j} \right\} K_{ji} / \sum_{j=1}^n L_{w,j} K_{ji}$ similarly, we conclude that

$$\hat{p}(X_i) - \tilde{p}(X_i) = p_0(V_i, w)p_1(V_i, w) \left(\frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_p(n^{-1/2}).$$

Therefore, we write the difference in (ii) as

$$\mathbf{E}_Q [p_0(V_i, w)p_1(V_i, w)S_i] \left(\frac{\hat{q}_{1,w} - q_{1,w}}{q_{1,w}} - \frac{\hat{q}_{0,w} - q_{0,w}}{q_{0,w}} \right) + o_p(n^{-1/2}).$$

Lemma A2 : (i) Suppose that the assumptions of Theorem 3 hold, and let $\varepsilon_{d,w,i} = Y_{di} - \beta_d(V_i, w)$. Then,

$$\begin{aligned} & \frac{p_{1,w}}{q_{1,w}n} \sum_{i \in S_{1,w}} \frac{g(V_i, w)Y_i}{\hat{p}_1(V_i, w)} - \frac{p_{0,w}}{q_{0,w}n} \sum_{i \in S_{0,w}} \frac{g(V_i, w)Y_i}{\hat{p}_0(V_i, w)} \\ = & \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{1,w,i}\varepsilon_{1,w,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{0,w,i}\varepsilon_{0,w,i}}{p_0(V_i, w)} \\ & + \frac{1}{n} \sum_{i=1}^n g(V_i, w)\tau(V_i, w)L_{w,i} + o_p(n^{-1/2}). \end{aligned}$$

(ii) Suppose that the assumptions of Theorem 4 hold, and let $\varepsilon_{d,w,i} = Y_{di} - \beta_d(V_i, w)$. Then,

$$\begin{aligned}
& \frac{p_{1,w}}{n_{1,w}} \sum_{i \in S_{1,w}} \frac{g(V_i, w)Y_i}{\tilde{p}_1(V_i, w)} - \frac{p_{0,w}}{n_{0,w}} \sum_{i \in S_{0,w}} \frac{g(V_i, w)Y_i}{\tilde{p}_0(V_i, w)} \\
= & \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{1,w,i}\varepsilon_{1,w,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{0,w,i}\varepsilon_{0,w,i}}{p_0(V_i, w)} \\
& + \frac{1}{n} \sum_{i=1}^n \{g(V_i, w)\tau(V_i, w) - \mathbf{E}_{1,w}[g(V_i, w)\tau(V_i, w)]\} L_{1,w,i} \\
& + \frac{1}{n} \sum_{i=1}^n \{g(V_i, w)\tau(V_i, w) - \mathbf{E}_{0,w}[g(V_i, w)\tau(V_i, w)]\} L_{0,w,i} \\
& + \mathbf{E}_{1,w}[g(V_i, w)\tau(V_i, w)]p_{1,w} + \mathbf{E}_{0,w}[g(V_i, w)\tau(V_i, w)]p_{0,w} + o_p(n^{-1/2}).
\end{aligned}$$

Proof of Lemma A2 : (i) We first write

$$\begin{aligned}
& \frac{p_{1,w}}{q_{1,w}n} \sum_{i \in S_{1,w}} \frac{g(V_i, w)Y_i}{\hat{p}_1(V_i, w)} - \frac{p_{0,w}}{q_{0,w}n} \sum_{i \in S_{0,w}} \frac{g(V_i, w)Y_i}{\hat{p}_0(V_i, w)} \\
= & \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)Y_i L_{1,w,i}}{\hat{p}_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)Y_i L_{0,w,i}}{\hat{p}_0(V_i, w)} = A_{1n} + A_{2n}, \text{ say.}
\end{aligned}$$

We first write $A_{1n} = \frac{1}{n} \sum_{i=1}^n g(V_i, w)Y_i L_{1,w,i}/p_1(V_i, w) + \tilde{A}_{1n}$, where

$$\tilde{A}_{1n} = \frac{1}{n} \sum_{i=1}^n g(V_i, w)Y_i L_{1,w,i} \left(\frac{1}{\hat{p}_1(V_i, w)} - \frac{1}{p_1(V_i, w)} \right).$$

The term on the right-hand side is equal to

$$\frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)Y_i L_{1,w,i}}{p_1^2(V_i, w)} (p_1(V_i, w) - \hat{p}_1(V_i, w)) + o_p(n^{-1/2}), \quad (20)$$

by the fact that $\sup_{x \in \mathcal{X}} |\hat{p}_1(x) - p_1(x)| = O_p(n^{-1/2}h^{-L_1/2}\sqrt{\log n} + h^{L_1+1}) = o_p(n^{-1/4})$. The uniform convergence is due to Theorem 4 of Hansen (2008) and the last convergence rate is due to Assumption 2(ii). We plug $S_i = g(V_i, w)Y_i L_{1,w,i}/p_1^2(V_i, w)$ in Lemma A1(i) to obtain that the leading sum in (20) is asymptotically equivalent to (up to $o_p(n^{-1/2})$)

$$\begin{aligned}
& -\frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)\mathbf{E}_Q[Y_i L_{1,w,i}|V_i, W_i = w]\mathcal{J}_{1,w,i}}{p_1^2(V_i, w)\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]} \\
& + \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)\mathbf{E}_Q[Y_i L_{1,w,i}|V_i, W_i = w]\mathcal{J}_{w,i}}{p_1(V_i, w)\mathbf{E}_Q[L_{w,i}|V_i, W_i = w]}.
\end{aligned} \quad (21)$$

Using the fact that

$$\frac{\mathbf{E}_Q[Y_i L_{1,w,i} | V_i, W_i = w]}{\mathbf{E}_Q[L_{w,i} | V_i, W_i = w]} = \beta_1(V_i, w) p_1(V_i, w), \quad (22)$$

we write the difference in (21) as

$$-\frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w) \beta_1(V_i, w) p_0(V_i, w)}{p_1(V_i, w)} \mathcal{J}_{1,w,i} + \frac{1}{n} \sum_{i=1}^n g(V_i, w) \beta_1(V_i, w) \mathcal{J}_{0,w,i}.$$

Therefore, we conclude that

$$\begin{aligned} \tilde{A}_{1n} &= -\frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w) \beta_1(V_i, w) p_0(V_i, w)}{p_1(V_i, w)} \mathcal{J}_{1,w,i} \\ &\quad + \frac{1}{n} \sum_{i=1}^n g(V_i, w) \beta_1(V_i, w) \mathcal{J}_{0,w,i} + o_p(n^{-1/2}). \end{aligned}$$

On the other hand, $A_{2n} = \frac{1}{n} \sum_{i=1}^n g(V_i, w) Y_i L_{0,w,i} / p_0(V_i, w) + \tilde{A}_{2n}$, where

$$\tilde{A}_{2n} = \frac{1}{n} \sum_{i=1}^n g(V_i, w) Y_i L_{0,w,i} \left(\frac{1}{\hat{p}_0(V_i, w)} - \frac{1}{p_0(V_i, w)} \right).$$

Similarly as before, we write

$$\begin{aligned} \tilde{A}_{2n} &= \frac{1}{n} \sum_{i=1}^n g(V_i, w) \beta_0(V_i, w) \mathcal{J}_{1,w,i} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w) \beta_0(V_i, w) p_1(V_i, w)}{p_0(V_i, w)} \mathcal{J}_{0,w,i} + o_p(n^{-1/2}). \end{aligned}$$

Combining the two results and using the fact that $\tau(X) = \beta_1(X) - \beta_0(X)$, we conclude

$$\begin{aligned} \tilde{A}_{1n} - \tilde{A}_{2n} &= -\frac{1}{n} \sum_{i=1}^n g(V_i, w) \left(\frac{\beta_1(V_i, w) - \tau(V_i, w) p_1(V_i, w)}{p_1(V_i, w)} \right) \mathcal{J}_{1,w,i} \\ &\quad + \frac{1}{n} \sum_{i=1}^n g(V_i, w) \left(\frac{\tau(V_i, w) p_0(V_i, w) + \beta_0(V_i, w)}{p_0(V_i, w)} \right) \mathcal{J}_{0,w,i} + o_p(n^{-1/2}). \end{aligned}$$

Plugging in this result and rearranging terms, we write

$$\begin{aligned}
& \frac{p_{1,w}}{q_{1,w}n} \sum_{i \in S_{1,w}} \frac{g(V_i, w)Y_i}{\hat{p}_1(V_i, w)} - \frac{p_{0,w}}{q_{0,w}n} \sum_{i \in S_{0,w}} \frac{g(V_i, w)Y_i}{\hat{p}_0(V_i, w)} \\
= & \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{1,w,i}\varepsilon_{1,w,i}}{p_1(V_i, w)} - \frac{1}{n} \sum_{i=1}^n \frac{g(V_i, w)L_{0,w,i}\varepsilon_{0,w,i}}{p_0(V_i, w)} \\
& + \frac{1}{n} \sum_{i=1}^n g(V_i, w)\tau(V_i, w)L_{1,w,i} + \frac{1}{n} \sum_{i=1}^n g(V_i, w)\tau(V_i, w)L_{0,w,i} \\
& + \frac{1}{n} \sum_{i=1}^n g(V_i, w) \left(\frac{\beta_1(V_i, w) - \tau(V_i, w)p_1(V_i, w)}{p_1(V_i, w)} \right) (\mathbf{E}_Q[L_{1,w,i}|V_i, W_i = w]) \\
& - \frac{1}{n} \sum_{i=1}^n g(V_i, w) \left(\frac{\tau(V_i, w)p_0(V_i, w) + \beta_0(V_i, w)}{p_0(V_i, w)} \right) \{\mathbf{E}_Q[L_{0,w,i}|V_i, W_i = w]\} + o_p(n^{-1/2}).
\end{aligned}$$

As for the last two terms, observe that

$$\begin{aligned}
H_n & \equiv \left\{ \frac{\beta_1(V_i, w)}{p_1(V_i, w)} - \tau(V_i, w) \right\} \mathbf{E}_Q[L_{1,w,i}|V_i, W_i = w] \\
& \quad - \left\{ \frac{\beta_0(V_i, w)}{p_0(V_i, w)} + \tau(V_i, w) \right\} \mathbf{E}_Q[L_{0,w,i}|V_i, W_i = w] \\
& = \left\{ \frac{\beta_1(V_i, w)}{p_1(V_i, w)} - \tau(V_i, w) \right\} \frac{q_1(V_i, w)p_{1,w}}{q_{1,w}} \\
& \quad - \left\{ \frac{\beta_0(V_i, w)}{p_0(V_i, w)} + \tau(V_i, w) \right\} \frac{q_0(V_i, w)p_{0,w}}{q_{0,w}}.
\end{aligned}$$

However, by Bayes' rule,

$$\frac{p_{1,w}q_1(V_i, w)}{q_{1,w}} = \frac{p_{1,w}q_1(V_i, w)f_Q(V_i, w)}{q_{1,w}f_Q(V_i, w)} = \frac{p_{1,w}f(V_i|1, w)}{f_Q(V_i, w)} = \frac{p_1(V_i, w)f_P(V_i, w)}{f_Q(V_i, w)}. \quad (23)$$

Using this result and the definition of $\tau(V_i, w)$, we can show that $H_n = 0$. Hence we obtain the wanted result.

