Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand*

Richard K. Crump† V. Joseph Hotz‡ Guido W. Imbens§ Oscar A. Mitnik¶

First Draft: July 2004
This Draft: September 17, 2006

Abstract

Estimation of average treatment effects under unconfoundedness or exogenous treatment assignment is often hampered by lack of overlap in the covariate distributions. This lack of overlap can lead to imprecise estimates and can make commonly used estimators sensitive to the choice of specification. In such cases researchers have often used informal methods for trimming the sample. In this paper we develop a systematic approach to addressing such lack of overlap. We characterize optimal subsamples for which the average treatment effect can be estimated most precisely, as well as optimally weighted average treatment effects. Under some conditions the optimal selection rules depend solely on the propensity score. For a wide range of distributions a good approximation to the optimal rule is provided by the simple selection rule to drop all units with estimated propensity scores outside the range $[0.1, 0.9]$.

JEL Classification: C14, C21, C52

Keywords: *Average Treatment Effects, Causality, Unconfoundedness, Overlap, Treatment Effect Heterogeneity*

*We are grateful for helpful comments by Richard Blundell, Gary Chamberlain, Jinyong Hahn, Gary King, Michael Lechner, Robert Moffitt, Geert Ridder and Don Rubin, and by participants in seminars at the ASSA meetings in Philadelphia, University College London, UCLA, UC–Berkeley, UC–Riverside, MIT–Harvard, Johns Hopkins University, the Malinvaud seminar at CREST, and the IAB Empirical Evaluation of Labour Market Programmes conference.

†Department of Economics, University of California at Berkeley, crump@econ.berkeley.edu, http://socrates.berkeley.edu/∼ crump/.

‡Department of Economics, University of California at Los Angeles, hotz@econ.ucla.edu, http://www.econ.ucla.edu/hotz/.

§Department of Economics, Harvard University, Littauer Center, Cambridge, MA 02138, imbens@harvard.edu, http://www.economics.harvard.edu/faculty/imbens/imbens.html.

¶Dept of Economics, University of Miami, omitnik@miami.edu, http://moya.bus.miami.edu/∼ omitnik/.

1 Introduction

There is a large literature on estimating average treatment effects (ATE) under assumptions of unconfoundedness, ignorability, or exogeneity following the seminal work by Rubin (1974, 1978) and Rosenbaum and Rubin (1983a). Researchers have developed estimators based on regression methods (e.g., Hahn, 1998, Heckman, Ichimura and Todd, 1998), matching (e.g., Rosenbaum, 1989, Abadie and Imbens, 2006), and methods based on the propensity score (e.g., Rosenbaum and Rubin, 1983a, Hirano, Imbens and Ridder, 2003). Related methods for missing data problems are discussed in Robins, Rotnitzky and Zhao (1995) and Robins and Rotnitzky (1995).[1] An important practical concern in implementing these methods is that one needs overlap between covariate distributions in the two subpopulations, i.e., there must be common support in the covariates across the two subpopulations. Even if there exists overlap (common support), there may be parts of this common covariate space with limited numbers of observations for one or the other treatment groups. Such areas of limited overlap can lead to poor finite sample properties for many estimators of average treatment effects. In such cases, many of these estimators can have substantial bias, large variances, as well as considerable sensitivity to the exact specification of the treatment effect regression functions or of the propensity score. LaLonde (1986), Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (1999) discuss the empirical relevance of this overlap issue.[2]

One strand of the literature has focused on assessing the robustness of existing estimators to a variety of potential problems, including limited overlap.[3] A second strand focuses on developing new matching estimators of treatment effects or modifying existing ones to reduce their sensitivity and improve their precision in the face of the overlap problem. For example, Rubin (1977) and Lee (2005b), in situations where there is a single discrete covariate, suggest simply discarding all units with covariate values with either no treated or no control units. Alternatively, Cochran and Rubin (1973) suggest caliper matching where potential matches are dropped if the within-match difference in propensity scores exceeds some threshold level. LaLonde (1986) creates subsamples of the control group by conditioning on covariate values lying in ranges with substantial overlap. Ho, Imai, King and Stuart (2005) propose preprocessing the data by first matching units and carrying out parametric inferences using only the matched data. Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1998), and Smith and Todd (2005), who focus on estimating the average treatment effect for the treated (ATT), discard all observations in both the treated and non-treated groups for values of the estimated propensity scores that have zero or occur infrequently. Dehejia and Wahba (1999), who also focus on estimating the ATT, discard those non-treated group observations

[1] See Rosenbaum (2001), Heckman, LaLonde and Smith (1999), Wooldridge (2002), Blundell and Costa-Diaz (2002), Imbens (2004) and Lee (2005a) for surveys of this literature.

[2] Dehejia and Wahba (1999) write: "... our methods succeed for a transparent reason: They only use the subset of the comparison group that is comparable to the treatment group, and discard the complement." Heckman, Ichimura and Todd (1997) write "A major finding of this paper is that comparing the incomparable—i.e., violating the common support condition for the matching variables—is a major source of evaluation bias as conventionally measured."

[3] See, for example, Rosenbaum and Rubin (1983b), Rosenbaum (2001), Imbens (2003), and Ichino, Mealli, and Nannicini (2005).

for propensity scores that are less than the smallest value of the propensity score for those in the treated group.

Although there are differences across these alternative strategies, they have several things in common. First, all of them discard observations for which there is no overlap between the treated and non-treated group based on either the propensity score or covariate distribution. As a result, each strategy focuses, in essence, on average treatment effect estimands that are defined for subsets of the sample observations and, thus, differ from either the typical ATE or ATT which are defined over the full (population) covariate distributon. Second, each of these strategies is somewhat arbitrary, i.e., the strategies used to discard or reweight observations in forming new estimators are based on criteria with unknown properties.

In this paper, we propose a systematic approach to dealing with samples with limited overlap in the covariates that have optimality properties with respect to the precision of estimating treatment effects, and which are straightforward to implement in practice. As with the previous methods, our approaches also are based on characterizing different estimands relative to the traditional ATE or ATT. We return below to the implications of and some justifications for this latter feature of our approach.

We consider the following two strategies. In the first, we focus on average treatment effects within a selected subpopulation defined in terms of covariate values. Inevitably, conditioning on a subpopulation based on any selection criterion reduces the effective sample size, which, all else the same, increases the variance of the estimated average treatment effect. However, if the subpopulation is chosen appropriately, it may be possible to estimate the average treatment within this subpopulation more precisely than the average effect for the entire population despite the smaller sample size. As we establish below, this tradeoff is, in general, well-defined and, under some conditions, leads to discarding units with propensity scores outside an interval $[\alpha, 1-\alpha]$, where the optimal cutoff value of α is solely determined by the distribution of the propensity score. Our approach is consistent with the practice noted above of researchers dropping units with extreme values of the propensity score, with two important distinctions. First, the role of the propensity score in our procedure is not imposed from the outset; rather, it emerges as a consequence of the criterion of variance minimization. Second, we have a systematic way of choosing the cutoff point, α. We refer to the resulting estimand as the Optimal Subpopulation Average Treatment Effect (OSATE). We note that the determination of the subset of observations that characterize a particular OSATE is based solely on the joint distribution of covariates and the treatment indicator and *not* on the outcome data. As a result, we avoid introducing deliberate bias with respect to the treatment effects being analyzed.

In the second strategy, we formulate weighted average treatment effects, where the weights depend only on covariates. Note that the OSATE can be viewed as a special case of these weighted treatment effects, where the weight function is restricted to be an indicator function. Within a broad class, we characterize the weight function that leads to the most precisely estimated average treatment effect. We note that this class of estimands includes the average treatment effect for the treated, where the weight function is proportional to the propensity score. Under the same conditions as before, the optimal weight function turns out to be a function of the propensity score; in fact, it is proportional to the product of the propensity

score and one minus the propensity score. We refer to this as the Optimally Weighted Average Treatment Effect (OWATE).

Although both strategies we consider are similiar to the more informal ones noted above, it is still the case that both are somewhat uncommon in econometric analyses, precisely because they entail focusing on estimands that depend on sample data.[4] Typically, econometric analyses of treatment effects focus on estimands that are defined *a priori* for populations of interest, as is the case with the population average treatment effect or the average treatment effect for the treated subpopulation. In these cases, estimates are produced that turn out to be more or less precise, depending on the actual sample data. In contrast, we focus on average effects for a statistically defined (weighted) subpopulation.[5] This change of focus is *not* motivated, *per se*, by an intrinsic interest in the subpopulation for which we ultimately estimate the average causal effect. Rather, it acknowledges and addresses the difficulties in making inferences about the population of primary interest.

In our view this approach has several justifications. First, our approach of achieving precision in the estimation of treatment effects has analogues in the statistics literature. In particular, it is similar to the traditional motivation for medians rather than means as more precise measures of central tendency. In particular, by changing the sample from one that was potentially representative of the population of interest, we can gain greater internal validity, although, in doing so, we may sacrifice some of the external validity of the resulting estimates.[6] Furthermore, our proposed approach of placing greater stress on internal versus external validity is similar to that found in the design of randomized experiments which are often carried out on populations unrepresentative of the population of interest in order to improve the precision of the inferences to be drawn. More generally, the relative primacy of internal validity over external validity is advocated in many discussions of causal inference (see, for example, Shadish, Cook, and Campbell, 2002).

Second, our approach may be well-suited to situations where the primary interest is to determine whether a treatment may harm or benefit at least *some* group in a broader population. For example, one may be interested whether there is any evidence that a particular drug could harm or have side effects for some group of patients in a well-defined population. In this context, obtaining greater precision in the estimation of a treatment effect, even if it is not for the entire population, is warranted. We note that the subpopulation for which these estimands are valid are defined in terms of the observed covariate values so that one can determine, for each individual, whether they are in the relevant subpopulation or not.

Third, our approach can provide useful, albeit auxiliary, information when making inferences about the treatment effects for fixed populations. Thus, instead of only reporting the potentially imprecise estimate for the population average treatment effect, one can also report the estimates

[4] We note that the local average treatment effect introduced by Imbens and Angrist (1994) represents another example in which a new estimand is introduced—one in which the average effect of the treatment is defined for the subpopulation of compliers—to deal with a phenomenon quite similar to limited overlap.

[5] This is also true for the method proposed by Heckman, Ichimura and Todd (1998).

[6] A separate issue is that in practice in many cases even the original sample is not representative of the population of interest. For example, we are often interested in policies that would extend small pilot versions of job training programs to different locations and times.

for the subpopulations where we can make more precise inferences.

Fourth, focusing on estimands that discard or reweight observations from the treated and non-treated group subsamples in order to improve precision tends to produce more balance in the distribution of the covariates across these groups. As has been noted elsewhere (Rosenbaum and Rubin, 1984; Heckman, Ichimura and Todd, 1998, among others), increasing the balance in the covariate distributions tends to reduce the sensitivity of treatment effect estimates to changes in the specification. In the extreme case, where the selected sample is completely balanced in covariates in the two treatment arms, one can simply use the average difference in outcomes between treated and control units.

At the same time, our focus on strategies for improving the precision of treatment effect estimators in the face of limited overlap has its limitations. For example, one might seek to devise strategies to deal with limited overlap of the covariate distributions that balance the representativeness of that distributon with precision. While exploring how to achieve such objectives is desirable, we see the results in this paper as an important first step in formulating strategies to deal with the problem of limited overlap that have well-defined properties and that can be implemented on real data.

Finally, it is important to note that the properties we derive below concerning the precision associated with both the OSATE and OWATE estimands are not tied to a specific estimator. Rather, we focus on differences in the efficiency bounds for different subpopulations. As a consequence, a range of efficient estimators—including the ones proposed by Hahn (1998), Hirano, Imbens and Ridder (2003), Imbens, Newey and Ridder (2006), and Robins, and Rotnitzky and Zhao (1995)—can potentially be used to estimate these estimands, especially the OWATE. However, as we make clear below, these standard estimators are not readily applicable to the estimation of the OSATE, due to the complications that arise from having to estimate the optimal subsets of the covariate distribution for this estimand. Accordingly, we develop a new estimator that deals with this case and derive its large sample properties.

We illustrate these methods using data from the non-experimental part of a data set on labor market programs previously used by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2005) and others. In this data set the overlap issue is a well known problem, with the control and treatment group far apart on some of the most important covariates including lagged values for the outcome of interest, yearly earnings. Here our OSATE method suggests dropping 2363 out of 2675 observations (leaving only 312 observations, or just 12% of the original sample) in order to minimize the variance. Calculations suggest that this lowers the variance by a factor $1/160,000$, reflecting the fact that most of the controls are very different from the treated and that it is essentially impossible to estimate the population average treatment effect. More relevant, given the fact that most of the researchers analyzing this data set have focused on the average effect for the treated, is that the variance for the optimal subsample is only 40% of that for the propensity score weighted sample (which estimates the effect on the treated).

The remainder of the paper is organized as follows. In Section 2, we present a simple example in which there is a single and scalar covariate used in the estimation of the average treatment effect. This example allows us to illustrate how the precision of the estimates varies with

changes in the estimand. Section 3 develops the general setup we use throughout the paper. Section 4 reviews the previous approaches to dealing with limited overlap when estimating treatment effects. In Section 5, we develop new estimands and discuss their precision gains. We also show that for a wide class of distributions the optimal set is well approximated by the set of observations with propensity scores in the interval $[0.1, 0.9]$. In Section 6, we discuss the properties of estimators for the OSATE and OWATE estimands. In Section 7, we present the application to the LaLonde data. Section 8 concludes.

2 A Simple Example

To set the stage for the issues to be discussed in this paper, consider the following simplified treatment effect example in which the covariate of interest, X, is a scalar taking on one of two values. In particular, suppose that $X = f$ (female) or $X = m$ (male), so that the covariate space is $\mathbb{X} = \{f, m\}$. For $x = f, m$, let N_x be the sample size for the subsample with $X = x$, and let $N = N_f + N_m$ be the total sample size. Let $W \in \{0, 1\}$ denote the indicator for the treatment. Also, let $p = \mathbb{E}[W]$ be the population share of treated individuals, where $\hat{p} = N_m/N$ is the share of men in the sample. We denote the average treatment effect, conditional on $X = x$, as τ_x. Let N_{xw} be the number of observations with covariate $X_i = x$ and treatment indicator $W_i = w$. It follows that $e_x = N_{x1}/N_x$ is the propensity score for $x = f, m$. Finally, let $\bar{y}_{xw} = \sum_{i=1}^{N} Y_i \cdot 1\{X_i = x, W_i = w\}/N_{xw}$ be the average within each of the four subsamples. We assume that the distribution of the outcomes is homoskedastic, i.e., the variance of $Y(w)$ given $X_i = x$ is σ^2 for all $x = f, m$ and $w = 0, 1$.

At the outset, consider the following two average treatment effects that differ in somewhat subtle ways. In particular, consider the average effect that is averaged over the *sample distribution* of the covariates,

$$\tau_S = \hat{p} \cdot \tau_m + (1 - \hat{p}) \cdot \tau_f,$$

versus the average treatment effect for the full population,

$$\tau_P = p \cdot \tau_m + (1 - p) \cdot \tau_f.$$

where $\tau_x = \mathbb{E}[y_{x1} - y_{x0}]$, $x = f, m$. It is immediately obvious that for either τ_S and τ_P the natural estimator is

$$\hat{\tau}_{\mathbb{X}} = \hat{p} \cdot \hat{\tau}_m + (1 - \hat{p}) \cdot \hat{\tau}_f.$$

However, as we develop below, which estimand is the object of interest makes a difference in terms of the variance for this esimtator and this fact plays a crucial role in the results derived in this paper.

To make things very simple, suppose that subjects are randomly assigned to one of the treatment statuses, $W_i = 0$ or 1, conditional on X. In this case, the natural, unbiased, estimators for the average treatment effects for each of the two subpopulations are

$$\hat{\tau}_f = \bar{y}_{f1} - \bar{y}_{f0}, \qquad \text{and} \qquad \hat{\tau}_m = \bar{y}_{m1} - \bar{y}_{m0},$$

with variances (conditional on the covariates)

$$\mathbb{V}(\hat{\tau}_f) = \mathbb{E}[(\hat{\tau}_f - \tau_f)^2] = \sigma^2 \cdot \left(\frac{1}{N_{f0}} + \frac{1}{N_{f1}}\right) = \frac{\sigma^2}{N \cdot (1-\hat{p})} \cdot \frac{1}{e_f \cdot (1-e_f)},$$

and

$$\mathbb{V}(\hat{\tau}_m) = \mathbb{E}[(\hat{\tau}_m - \tau_m)^2] = \sigma^2 \cdot \left(\frac{1}{N_{m0}} + \frac{1}{N_{m1}}\right) = \frac{\sigma^2}{N \cdot \hat{p}} \cdot \frac{1}{e_m \cdot (1-e_m)}.$$

respectively. The estimator for the sample average treatment effect, τ_S, is

$$\hat{\tau}_{\mathbb{X}} = \hat{p} \cdot \hat{\tau}_m + (1-\hat{p}) \cdot \hat{\tau}_f.$$

Because the two estimates, $\hat{\tau}_f$ and $\hat{\tau}_m$, are independent, it follows that the variance of this estimator is

$$\begin{aligned}\mathbb{V}_S(\hat{\tau}_{\mathbb{X}}) &= \mathbb{E}[(\hat{\tau}_{\mathbb{X}} - \tau_S)^2] = \hat{p}^2 \cdot V(\hat{\tau}_m) + (1-\hat{p})^2 \cdot V(\hat{\tau}_f) \\ &= \frac{\sigma^2}{N} \cdot \left(\frac{\hat{p}}{e_m \cdot (1-e_m)} + \frac{1-\hat{p}}{e_f \cdot (1-e_f)}\right).\end{aligned}$$

It follows that the asymptotic variance of $\sqrt{N}(\hat{\tau}_{\mathbb{X}} - \tau_S)$ converges to

$$AV\left(\sqrt{N}(\hat{\tau}_{\mathbb{X}} - \tau_S)\right) = \sigma^2 \cdot \left(\frac{p}{e_m \cdot (1-e_m)} + \frac{1-p}{e_f \cdot (1-e_f)}\right) = \sigma^2 \cdot \mathbb{E}\left[\frac{1}{e_X \cdot (1-e_X)}\right].$$

Note, however, that the asymptotic variance of $\sqrt{N}(\hat{\tau}_{\mathbb{X}} - \tau_P)$ converges to

$$\begin{aligned}AV\left(\sqrt{N}(\hat{\tau}_{\mathbb{X}} - \tau_P)\right) &= \sigma^2 \cdot \left(\frac{p}{e_m \cdot (1-e_m)} + \frac{1-p}{e_f \cdot (1-e_f)}\right) + p \cdot (\tau_m - \tau_P)^2 + (1-p) \cdot (\tau_f - \tau_P)^2 \\ &= \sigma^2 \cdot \mathbb{E}\left[\frac{1}{e_X \cdot (1-e_X)}\right] + \mathbb{E}[(\tau_X - \tau_P)^2].\end{aligned}$$

where the extra term in this second variance arises because of the difference between the average treatment effect conditional on the sample distribution of X and the one for the full population.

The first formal result of the paper concerns the comparison of $\mathbb{V}_S(\hat{\tau})$, $\mathbb{V}(\hat{\tau}_f)$, and $\mathbb{V}(\hat{\tau}_m)$ according to a variance minimization criterion. In particular, the optimal subset $\mathbb{A}^* \subset \mathbb{X}$ that minimizes

$$\mathbb{V}_{\min} = \min(\mathbb{V}_S(\hat{\tau}_{\mathbb{X}}), \mathbb{V}(\hat{\tau}_f), \mathbb{V}(\hat{\tau}_m)) = \mathbb{V}(\hat{\tau}_{\mathbb{A}^*}).$$

is given by

$$\mathbb{A}^* = \begin{cases} \{f\} & \text{if} \quad \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1-\hat{p}}{2-\hat{p}}, \\ \mathbb{X} & \text{if} \quad \frac{1-\hat{p}}{2-\hat{p}} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1+\hat{p}}{\hat{p}}, \\ \{m\} & \text{if} \quad \frac{1+\hat{p}}{\hat{p}} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)}. \end{cases} \tag{2.1}$$

Note that which estimator has the smallest variances crucially depends on the ratio of the product of the propensity score and one minus the propensity score, $e_m(1-e_m)/(e_f(1-e_f))$.

If the propensity score for women is close to zero or one, we cannot estimate the average treatment effect for women precisely. In that case the ratio $e_m(1-e_m)/(e_f(1-e_f))$ will be high and we may be able to estimate the average treatment effect for men more accurately than the average effect for the sample as a whole, even though we may well lose a substantial number of observations by discarding women. Similarly, if the propensity score for men is close to zero or one, the ratio $e_m(1-e_m)/(e_f(1-e_f))$ is close to zero, and we may be able to estimate the average treatment effect for the women more accurately than for the sample as a whole. If the ratio is close to one, we can estimate the average treatment effect for the population as a whole more accurately than for either of the two subpopulations. Put differently, based on the data, and more specifically the distribution of (X, W), one might prefer to estimate τ_f (or τ_m), rather than the overall average τ, if, *a priori*, it is clear that τ cannot be estimated precisely, and τ_f (or τ_m) can be estimated with accuracy. In this case there is a second obvious advantage of focusing on subpopulation average treatment effects. Within the two subpopulations, we can estimate the within-subpopulation average treatment effect without bias by simply differencing average treatment and control outcomes. As a result, our results are not sensitive to the choice of estimator for the within-subpopulation treatment effects. This need not be the case for the population as a whole, where there is potentially substantial bias from simply differencing average outcomes.

Note that we did *not* define $\mathbb{A}^*$ so as to minimize $(AV(\sqrt{N}(\hat{\tau}_{\mathbb{X}} - \tau_P))/N, \mathbb{V}(\hat{\tau}_f), \mathbb{V}(\hat{\tau}_m))$. While doing so is, in principle, possible, it has two drawbacks, given our desire to determine the estimator which has the smallest variance and that is implementable in practice. First, using $\min(AV(\sqrt{N}(\hat{\tau}\mathbb{X} - \tau_P))/N, \mathbb{V}(\hat{\tau}_f), \mathbb{V}(\hat{\tau}_m))$ as the criteria for estimator selection would require one to evaluate $\mathbb{E}[(\tau_X - \tau_P)^2]$, which would necessarily be difficult to do. Second, this criterion depends on the value of the treatment effect, and, as such, would require analyzing outcome data, Y, before selecting the sample. This would open the door to introducing deliberate biases of the sort avoided by a selection criterion that depends solely on the treatment and covariate data.

A second issue concerns knowledge of $\mathbb{A}^*$. In an actual data set, one typically does not know $\mathbb{A}^*$ and it would have to be estimated, using estimated values for the propensity score and the covariate distribution. Call this estimate $\hat{\mathbb{A}}$. In cases with continuous covariates, the uncertainty stemming from the difference between $\hat{\mathbb{A}}$ and $\mathbb{A}^*$ is not neglible. As a result, in our discussion of statistical inference below, we focus on the distribution of $\sqrt{N}(\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{\hat{\mathbb{A}}})$, rather than at the distribution of $\sqrt{N}(\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{\mathbb{A}^*})$. That is, we focus on the deviation of the estimated average effect relative to the average effect *in the selected subsample*, not relative to the average effect in the subset that would be optimal in the population. To be clear, focusing on $\tau_{\hat{\mathbb{A}}}$ rather than $\tau_{\mathbb{A}^*}$ has consequences. For example, suppose that our estimate is $\hat{\mathbb{A}} = \{m\}$, so that we estimate the average treatment effect using only data for the male subpopulation. It may well be that, in fact, $\mathbb{A}^* = \mathbb{X}$ so that the average treatment effect should be estimated over the population of men *and* women. Nevertheless, we focus on the distribution of $\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{\hat{\mathbb{A}}} = \hat{\tau}_m - \tau_m$, rather than on the asymptotic distribution of $\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{\mathbb{A}^*} = \hat{\tau}_{\hat{\mathbb{A}}} - \tau_{\hat{\mathbb{A}}} + (\tau_{\hat{\mathbb{A}}} - \tau_{\mathbb{A}^*})$. Given that $\hat{\mathbb{A}}$ is known, and $\mathbb{A}^*$ is not, the estimates would seem more interpretable that way.

The second result of the paper takes account of the fact that one need not limit the choice of

average treatment effects to the three discussed so far. In particular, one may wish to consider a weighted average treatment effect of the form

$$\tau_\lambda = \lambda \cdot \tau_m + (1-\lambda) \cdot \tau_f,$$

for fixed λ. It follows that τ_λ can be estimated by

$$\hat{\tau}_\lambda = \lambda \cdot \hat{\tau}_m + (1-\lambda) \cdot \hat{\tau}_f,$$

where the variance for this weighted average treatment effect is given by

$$\begin{aligned}\mathbb{V}(\hat{\tau}_\lambda) &= \lambda^2 \cdot \mathbb{V}(\hat{\tau}_m) + (1-\lambda)^2 \cdot \mathbb{V}(\hat{\tau}_f) \\ &= \lambda^2 \cdot \frac{\sigma^2}{N \cdot \hat{p}} \cdot \frac{1}{e_m \cdot (1-e_m)} + (1-\lambda)^2 \cdot \frac{\sigma^2}{N \cdot (1-\hat{p})} \cdot \frac{1}{e_f \cdot (1-e_f)}.\end{aligned}$$

It follows that the variance of this estimator is minimized by choosing λ to be

$$\lambda^* = \frac{1/\mathbb{V}(\hat{\tau}_m)}{1/\mathbb{V}(\hat{\tau}_m) + 1/\mathbb{V}(\hat{\tau}_f)} = \frac{\hat{p} \cdot e_m \cdot (1-e_m)}{(1-\hat{p}) \cdot e_f \cdot (1-e_f) + \hat{p} \cdot e_m \cdot (1-e_m)}. \tag{2.2}$$

with the minimum value for the variance equal to

$$\mathbb{V}(\tau_{\lambda^*}) = \frac{\sigma^2}{N} \cdot \frac{1}{((1-\hat{p}) \cdot e_f \cdot (1-e_f) + \hat{p} \cdot e_m \cdot (1-e_m))}.$$

The ratio of the variance for the population average to the variance for the optimally weighted average treatment effect is

$$\mathbb{V}(\tau_P)/\mathbb{V}(\tau_{\lambda^*}) \longrightarrow \mathbb{E}\left[\frac{1}{e_X \cdot (1-e_X)}\right] \cdot \mathbb{E}[e_X \cdot (1-e_X)] = \mathbb{E}\left[\frac{1}{\mathbb{V}(W|X)}\right] \cdot \mathbb{E}[\mathbb{V}(W|X)]. \tag{2.3}$$

By Jensen's inequality this is greater than one if $\mathbb{V}(W|X)$ varies with X.

So what are some of the implications of focusing on a criterion of variance minimization for selecting among alternative treatment effect estimators derived from this simplified example? Suppose one is interested in the sample average treatment effect, τ_S. One may find that the efficient estimator for this average effect is likely to be imprecise, even before looking at the outcome data. This would be consistent with two states of the world that correspond to very different sets of information about treatment effects. In one state, the average effect for both of the subpopulations are imprecisely estimable, and, in effect, one cannot say much about the effect of the treatment at all. In the other state of the world it is still possible to learn something about the effect of the treatment because one of the subpopulation average treatment effects can be estimated precisely. In that case, which corresponds to the propensity score for one of the two subpopulations being close to zero or one, it may be useful to report also the estimator for the precisely estimable average treatment effect to convey the information the data contain about the effect of the treatment. It is important to stress that the message of the paper is not that one should report *only* $\hat{\tau}_m$ or $\hat{\tau}_f$ in place of $\hat{\tau}$. Rather, in cases where $\hat{\tau}_m$ or $\hat{\tau}_f$ are precisely estimable and $\hat{\tau}$ is not, we propose one should report *both*.

In the remainder of the paper, we generalize the above analysis to the case with a vector of potentially continuously distributed covariates. We study the existence and characterization of a partition of the covariates space $\mathbb{X}$ into two subsets, $\mathbb{A}^*$ and $\mathbb{X}/\mathbb{A}^*$. For $\mathbb{A}^*$, the average treatment effect is at least as accurately estimable as that for any other subset of the covariate space. This leads to a generalization of (2.1). Under a certain set of assumptions, this problem has a well-defined solution and, under homoskedasticity, these subpopulations have a very simple characterization, namely the set of covariates such that the propensity score is in the closed interval $[\alpha, 1-\alpha]$, or $\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(x) \leq 1-\alpha\}$. The optimal value of the boundary point, α, is determined by the distribution of the propensity score and its calculation is straightforward. Compared to the binary covariate case just considered, it will be difficult to argue in the general setting that this subpopulation is of intrinsic or substantive interest. We will not attempt to do so. Instead, we view it as an interesting average treatment effect because of its statistical properties and, in particular, as a convenient summary measure of the full distribution of conditional treatment effects $\tau(x)$. In addition, we characterize the optimally weighted average treatment effect and its variance, the generalization of (2.2) and (2.3).

3 Setup

The framework we use is standard in this literature.[7] We have a random sample of size N from a large population. For each unit i in the sample, let W_i indicate whether the treatment of interest was received, with $W_i = 1$ if unit i receives the treatment of interest, and $W_i = 0$ if unit i receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), let $Y_i(0)$ denote the outcome for unit i under control and $Y_i(1)$ the outcome under treatment. We observe W_i and Y_i, where

$$Y_i = Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0).$$

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by X_i. Define the two conditional mean functions, $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$, the two conditional variance functions, $\sigma_w^2(x) = \text{Var}(Y(w)|X = x)$, the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x)$, and the propensity score, the probability of selection into the treatment, $e(x) = \Pr(W = 1|X = x) = \mathbb{E}[W|X = x]$.

Initially, we focus on two average treatment effects. The first is the (super-)population average treatment effect

$$\tau_P = \mathbb{E}[Y(1) - Y(0)]. \tag{3.4}$$

We also consider the sample average treatment effect

$$\tau_S = \frac{1}{N} \sum_{i=1}^{N} \tau(X_i), \tag{3.5}$$

[7] For example, see Rosenbaum and Rubin (1983a), Hahn (1998), Heckman, Ichimura and Todd (1998), and Hirano, Imbens and Ridder (2003).

where we condition on the observed set of covariates. The reason for focusing on the second one is twofold. First, it is analogous to the conditioning on covariates commonly used in regression analysis. Second, it can be estimated more precisely if there is variation in the treatment effect by covariates.

To solve the identification problem, we maintain throughout the paper the unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983a), which asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes:

Assumption 3.1 (Unconfoundedness)

$$W \perp (Y(0), Y(1)) \Big| X. \tag{3.6}$$

This assumption is widely used in this literature. See discussions in Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), Lechner (2002a), and others. In addition, we assume there is overlap in the covariate distributions:

Assumption 3.2 (Overlap) *For some $c > 0$, and all $x \in \mathbb{X}$, where $\mathbb{X}$ is the support of X,*

$$c \leq e(x) \leq 1 - c.$$

In addition, one often needs smoothness conditions on the two regression functions $\mu_w(x)$ and the propensity score $e(x)$ for estimation. We make those assumptions explicit in Section 6.

4 Previous Approaches to Dealing with Limited Overlap

In empirical applications, there is often concern about the overlap assumption (e.g., Dehejia and Wahba, 1999; Heckman, Ichimura, and Todd, 1997). As noted in the Introduction, researchers have sometimes trimmed their sample by excluding observations with propensity scores close to zero or one in order to ensure that there is sufficient overlap. Cochran and Rubin (1973) suggest using caliper matching where units whose match quality is too low according to the distance in terms of the propensity score are left unmatched.

Dehejia and Wahba (1999) focus on the average effect for the treated. They suggest dropping all control units with an estimated propensity score lower than the smallest value for the estimated propensity score among the treated units. Formally, they first estimate the propensity score. Let the estimated propensity score for unit i be $\hat{e}(X_i)$. Then let $\underline{e_1}$ be the minimum of the $\hat{e}(X_i)$ among treated units. Dehejia and Wahba drop all control units such that $\hat{e}(X_i) < \underline{e_1}$.

Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2005) also focus on the average effect for the treated. They propose discarding units with covariate values at which the estimated density is below some threshold. The precise method is as follows.[8] First, they estimate the propensity score $\hat{e}(x)$. Next, they estimate

[8] See Heckman, Ichimura and Todd (1997) and Smith and Todd (2005) for details, and Smith and Todd (2005) and Ham, Li and Reagan (2006) for applications of this method.

the density of the estimated propensity score in both treatment arms. Let $\hat{f}_w(e)$ denote the estimated density of the estimated propensity score. The specific estimator they use is a kernel estimator, $\hat{f}_w(e) = \frac{1}{N_w \cdot h} \sum_{i|W_i=w} K\left(\frac{\hat{e}(X_i)-e}{h}\right)$, with bandwidth h.[9] First, Heckman, Ichimura and Todd discard observations with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ exactly equal to zero leaving J observations.[10] Next, they fix a quantile q (Smith and Todd use $q = 0.02$). Using the J observations with positive densities, they rank the $2J$ values of $\hat{f}_0(\hat{e}(X_i))$ and $\hat{f}_1(\hat{e}(X_i))$. They then drop units i with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ less than or equal to c_q, where c_q is the largest real number such that

$$\frac{1}{2J} \sum_{i=1}^{J} \left(1\{\hat{f}_0(\hat{e}(X_i)) < c_q\} + 1\{\hat{f}_1(\hat{e}(X_i)) < c_q\} \right) \leq q.$$

Ho, Imai, King and Stuart (2005) propose combining any specific parametric procedure that the researcher may wish to employ with a nonparametric first stage. In this first stage, all treated units are matched to the closest control unit. Only the treated units and their matches are then used in the second stage. The first stage leads to a data set that is more balanced in terms of covariate distributions between treated and control. It thus reduces sensitivity of the parametric model to specific modelling decisions such as the inclusion of covariates or functional form assumptions.

All these methods tend to make the estimators more robust to specification decisions. However, few formal results are available on the properties of these procedures. They typically also depend on arbitrarily selected values for the "trimming" parameters.

5 Alternative Estimands

This section contains the main results of the paper. First, in Subsection 5.1, we review some results on efficiency bounds and present one new result. These efficiency bounds are used to motivate the estimands that we propose in the remainder of this section. In Subsection 5.2, we discuss the choice of criteria for selecting estimands that have optimal properties with respect to the estimation of average treatment effects. In Subsection 5.3, we derive the optimal subset of covariates over which to define the estimand, and in Subsection 5.4 we derive the optimal weights. Finally we provide some numerical calculations based on the Beta distribution.

5.1 Efficiency Bounds

In this subsection, we discuss some results on efficiency bounds for average treatment effects that will be used to motivate the estimands proposed in this paper. In addition, we present a new result on efficiency bounds.

Various bounds have been derived in the literature.[11] Hahn (1998) derived the semiparametric efficiency bound for $\tau_P = \mathbb{E}[Y(1) - Y(0)]$ under unconfoundedness and overlap (and

[9] In their application Smith and Todd (2005) use Silverman's rule of thumb to choose the bandwidth.

[10] Observations with the estimated density exactly equal to zero may exist when the kernel has finite support. For example, Smith and Todd (2005) use a quadratic kernel with $K(u) = (u^2-1)^2$ for $|u| \leq 1$ and zero elsewhere.

[11] See Hahn (1998), Robins, Rotznitzky and Zhao (1995), Robins, Mark, and Newey (1992), and Hirano, Imbens and Ridder (2003).

some regularity conditions).[12] The efficiency bound for τ_P is

$$\mathbb{V}_P^{\text{eff}} = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X)-\tau_P)^2\right].$$

A generalization of τ_P to the case of the weighted average treatment effect given by

$$\tau_{P,\omega} = \mathbb{E}\left[\tau(X)\cdot\omega(X)\right] \Big/ \mathbb{E}\left[\omega(X)\right],$$

with the weight function $\omega : \mathbb{X} \mapsto \mathbb{R}$ known, is considered in Hirano, Imbens and Ridder (2003), where they establish that the efficiency bound for $\tau_{P,\omega}$ is

$$\mathbb{V}_{P,\omega}^{\text{eff}} = \frac{1}{\mathbb{E}[\omega(X)]^2}\cdot\mathbb{E}\left[\omega(X)^2\cdot\left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X)-\tau_{P,\omega})^2\right)\right]. \tag{5.7}$$

Hirano, Imbens and Ridder (2003) propose the efficient estimator,

$$\hat{\tau}_\omega = \frac{1}{N}\sum_{i=1}^N \omega(X_i)\left(\frac{Y_i\cdot W_i}{\hat{e}(X_i)} - \frac{Y_i\cdot(1-W_i)}{1-\hat{e}(X_i)}\right) \Big/ \sum_{i=1}^N \omega(X_i),$$

for $\tau_{P,\omega}$. The influence function for this estimator is

$$\psi_{P,\omega}(y,w,x) = \frac{\omega(x)}{\mathbb{E}[\omega(X)]}\cdot\left(w\cdot\frac{y-\mu_1(x)}{e(x)} - (1-w)\cdot\frac{y-\mu_0(x)}{1-e(x)} + \mu_1(x) - \mu_0(x) - \tau_{P,\omega}\right),$$

so that

$$\hat{\tau}_\omega = \tau_{P,\omega} + \frac{1}{N}\sum_{i=1}^N \psi_{P,\omega}(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right).$$

Note that $\hat{\tau}_\omega$ also can be interpreted as an estimator of the weighted sample average treatment effect,

$$\tau_{S,\omega} = \sum_{i=1}^N \tau(X_i)\cdot\omega(X_i) \Big/ \sum_{i=1}^N \omega(X_i). \tag{5.8}$$

As an estimator for $\tau_{S,\omega}$, $\hat{\tau}_\omega$ satisfies

$$\sqrt{N}\cdot(\hat{\tau}_\omega - \tau_{S,\omega}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathbb{E}[\omega(X)]^2}\cdot\mathbb{E}\left[\omega(X)^2\cdot\left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right)\right]\right). \tag{5.9}$$

Comparing the efficiency bound for $\tau_{P,\omega}$ in (5.7) with the asymptotic variance in (5.9), it follows that we can estimate $\tau_{S,\omega}$ more accurately than $\tau_{P,\omega}$, so long as there is variation (with X) in the treatment effect $\tau(x)$.

Next we consider the case where the weights depend on the propensity score: $\omega(x) = \lambda(e(x))$, with $\lambda : [0,1] \mapsto \mathbb{R}$ known and the propensity score is unknown. (If the propensity score is known, this is a special case of the previous result.) If the propensity score is unknown, the efficiency bound changes. We establish what it is in the following theorem:

[12] See also Robins, Rotznitzky and Zhao (1995) for a related result in a missing data setting.

Theorem 5.1 (Weighted Average Treatment Effects with Weights Depending on the Propensity Score) *Suppose Assumptions 3.1 and 3.2 hold, and suppose that the weights are a function of the propensity score: $\omega(x) = \lambda(e(x))$ with $\lambda(e)$ known and $e(x)$ unknown. Then the semiparametric efficiency bound for $\tau_{P,\lambda}$ is*

$$\mathbb{V}_{P,\lambda}^{\text{eff}} = \frac{1}{\mathbb{E}[\lambda(e(X))]^2} \cdot \mathbb{E}\left[\lambda(e(X))^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X) - \tau_{P,\lambda})^2\right)\right]$$
$$+ \frac{1}{\mathbb{E}[\lambda(e(X))]^2} \cdot \mathbb{E}\left[e(X)(1-e(X)) \cdot \left[\frac{\partial}{\partial e}\lambda(e(X))\right]^2 (\tau(X) - \tau_{P,\lambda})^2\right]. \quad (5.10)$$

Proof: See Appendix.

The difference between the known weight function case (5.7) and the case with the weight function depending on the unknown propensity score established in Theorem 5.1 is the last term in (5.10). The fact that this term depends on the derivative of the weight function with respect to the propensity score will give rise to problems in formulating an implementable estimator. We address this issue in Section 6 below.

5.2 A Criterion for Choosing the Estimand

We now consider the problem of selecting the estimand that minimize the asymptotic variance in (5.9). Formally, we choose an estimand $\tau_{S,\omega}$ by choosing the weight function $\omega(x)$ that minimizes:

$$\frac{1}{\mathbb{E}[\omega(X)]^2} \cdot \mathbb{E}\left[\omega^2(X) \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right)\right]. \quad (5.11)$$

Under the assumption that the distribution of Y is homoskedastic, the criterion is slightly modified and one minimizes

$$\frac{1}{\mathbb{E}[\omega(X)]^2} \cdot \mathbb{E}\left[\omega^2(X) \cdot \left(\frac{1}{e(X)} + \frac{1}{1-e(X)}\right)\right]. \quad (5.12)$$

However, before implementing this approach, we offer several comments in support of using these variance-minimization criteria.

As we have discussed, one of the consequences of limited overlap in the covariate distributions is the imprecision with which average treatment effects are estimated. Suppose, for now, that the propensity score is known. In that case, no matter how unbalanced the sample is, there is an estimator that is exactly unbiased as long as the propensity score is strictly between zero and one. However, if the sample is severely imbalanced, the efficiency bound—and, in this case, the exact variance of the associated estimator—will be large. Because of this imprecision, it is desirable to try to modify the estimator to improve its precision. One possibility is to utilize a mean-squared-error type criterion for this modification. Unfortunately, implementing such a criterion would be very difficult in practice, as the biases of alternative estimators would be difficult to estimate. This is because these biases will depend on the entire function $\tau(x)$, which is likely to be difficult to estimate for some values of x. More generally, alternative estimators for τ_P will tend to suffer from this same problem.

To get around this problem, we focus on alternative estimands to τ_P. In doing so, the two questions naturally arise: What is the class of estimands and what is the criterion for choosing an estimand within that class? A natural class of estimands would seem to be $\tau_{P,\omega} = \mathbb{E}[\omega(X)\cdot\tau(X)]/\mathbb{E}[\omega(X)]$ and to use a criterion of the asymptotic variance of associated estimators in order to reduce the imprecision associated with limited overap. We do not pursue this class of estimands because evaluating the asymptotic variance of estimators for such estimands requires estimation of $\tau(x)$ over the entire support to deal with the term $\mathbb{E}[\omega^2(X)\cdot(\tau(X)-\tau_{P,\omega})^2]$ in the expression of the asymptotic variance in (5.7), making it difficult to implement in practice. Furthermore, the resulting variance-minimization criterion associated with this estimand would depend on values of the treatment effect, introducing the potential bias discussed in Section 2.

For these two reasons, we limit ourselves to the class of estimands given by $\tau_{S,\omega} = \sum_i \omega(X_i)\cdot\tau(X_i)/\sum_i \omega(X_i)$ for different $\omega(\cdot)$. We note, however, that this is not the only possible approach. There may be alternative classes of estimands or alternative criteria that would lead to effective solutions. The key, however, is to use a systematic approach characterized by a class of estimands and a formal criterion for choosing an estimand within that class that are easy to implement.

5.3 The Optimal Subpopulation Average Treatment Effect

In this section, we characterize the Optimal Subpopulation Averate Treatment Effect (OSATE). We do so by restricting attention to weight functions that are indicator functions: $\omega(x) = 1\{x \in \mathbb{A}\}$, where $\mathbb{A}$ is some closed subset of the covariate space $\mathbb{X}$. For a given set, $\mathbb{A}$, we define corresponding population and sample average treatment effects $\tau_{P,\mathbb{A}}$ and $\tau_{S,\mathbb{A}}$ as

$$\tau_{P,\mathbb{A}} = \mathbb{E}[\tau(X)|X\in\mathbb{A}], \qquad \text{and} \quad \tau_{S,\mathbb{A}} = \frac{1}{N_{\mathbb{A}}}\cdot\sum_{i|X_i\in\mathbb{A}}\tau(X_i),$$

respectively, where $N_{\mathbb{A}} = \sum_{i=1}^N 1\{X_i\in\mathbb{A}\}$ is the number of observations with covariate values in the set $\mathbb{A}$. Let $q(\mathbb{A}) = \Pr(X\in\mathbb{A}) = \mathbb{E}[1\{X\in\mathbb{A}\}]$. With this class of weight functions, the criterion given in (5.11) can be written as

$$\frac{1}{q(\mathbb{A})}\cdot\mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\middle| X\in\mathbb{A}\right]. \tag{5.13}$$

We look for an optimal $\mathbb{A}$, denoted by $\mathbb{A}^*$, that minimizes the asymptotic variance (5.13) among all closed subsets $\mathbb{A}$.

As noted in the Introduction, focusing on estimands that discard observations to reduce the variance of average treatment effect estimators has two opposing effects. First, by excluding units with covariate values outside the set $\mathbb{A}$, one reduces the effective sample size from N to $N\cdot q(\mathbb{A})$. This will increase the asymptotic variance by a factor $1/q(\mathbb{A})$. Second, by discarding units with high values for $\sigma_1^2(X)/e(X)+\sigma_0^2(X)/(1-e(X))$—that is, units with covariate values x such that it is difficult to estimate the average treatment effect $\tau(x)$—one can lower the conditional expectation $\mathbb{E}[\sigma_1^2(X)/e(X)+\sigma_0^2(X)/(1-e(X))|X\in\mathbb{A}]$. Optimally choosing $\mathbb{A}$ involves balancing these two effects.

The following theorem gives the formal result for the optimal $\mathbb{A}^*$ that minimizes the asymptotic variance. Define $k(x) = \sigma_1^2(x)/e(x) + \sigma_0^2(x)/(1-e(x))$.

Theorem 5.2 (Optimal Overlap for the Average Treatment Effect)
Suppose Assumptions 3.1-3.2 hold. Let $\underline{f} \leq f_X(x) \leq \overline{f}$, *and* $\sigma_w^2(x) \leq \overline{\sigma^2}$ *for* $w = 0, 1$ *and all* $x \in \mathbb{X}$. *We consider* $\tau(\mathbb{A})$ *where* $\mathbb{A}$ *is a closed subset of* $\mathbb{X}$. *Then the Optimal Subpopulation Average Treatment Effect (OSATE) is* $\tau_{S,\mathbb{A}^*}$, *where, if* $\sup_{x\in\mathbb{X}} k(x) \leq 2 \cdot \mathbb{E}\left[k(X)\right]$, *then* $\mathbb{A}^* = \mathbb{X}$ *and otherwise,*

$$\mathbb{A}^* = \left\{ x \in \mathbb{X} \,\middle|\, k(x) \leq \frac{1}{\alpha \cdot (1-\alpha)} \right\},$$

where α *is a solution to*

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[k(X) \,\middle|\, k(X) < \frac{1}{\alpha \cdot (1-\alpha)} \right].$$

Proof: See Appendix.

The result in this theorem simplifies in an interesting way under homoskedasticity.

Corollary 5.1 Optimal Overlap for the Average Treatment Effect Under Homoskedasticity *Suppose Assumptions 3.1-3.2 hold. Let* $\underline{f} \leq f_X(x) \leq \overline{f}$. *Suppose that* $\sigma_w^2(x) = \sigma^2$ *for all* $w \in \{0,1\}$ *and* $x \in \mathbb{X}$. *Then the OSATE under homoskedasticity is* $\tau_{S,\mathbb{A}^*}$, *where,*

$$\mathbb{A}_H^* = \{ x \in \mathbb{X} \,|\, \alpha \leq e(x) \leq 1-\alpha \}.$$

If

$$\sup_{x\in\mathbb{X}} \frac{1}{e(x)\cdot(1-e(x))} \leq 2 \cdot \mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))} \right],$$

then $\alpha = 0$. *Otherwise* α *is a solution to*

$$\frac{1}{\alpha\cdot(1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))} \,\middle|\, \frac{1}{e(X)\cdot(1-e(X))} \leq \frac{1}{\alpha\cdot(1-\alpha)} \right].$$

In Section 6, we focus on an estimator for the optimal estimand based on choosing weight functions that minimizes the criterion in (5.12) which corresponds to the case where the Y distribution is homoskedastic rather than using the criterion in (5.11). We use this criterion, even though we do not presume that the true distribution of Y is homoskedastic and, in fact, derive the asymptotic properties of this estimator under heteroskedasticity. We use the heteroskedastic criterion in (5.12) for three distinct reasons. The first—a principled reason—is that estimators of $\mathbb{A}_H^*$ *do not require using outcome data.* The entire analysis of selecting the sample can be carried out without using the outcome data, thus avoiding any deliberate bias that may result from selecting the sample based on outcome data. The second—a practical reason—is that the entire analysis is motivated by the difficulty of estimating $\tau(x)$ for covariates in some subset of the covariate space. For those values, it is even less likely that one can estimate the

conditional variances $\sigma_w^2(x)$ accurately, and, hence, methods that rely on nonparametrically estimating these conditional variances are unlikely to be effective using sample sizes found in practice. Note that the imbalance that precludes accurate estimation of $\tau(x)$ and $\sigma_w^2(x)$ does *not* necessarily preclude accurate estimation of the propensity score $e(x)$. In fact, when it is impossible to estimate $\tau(x)$ because there are no treated or no control units for a particular value of X, it need not be difficult at all to estimate the propensity score accurately. The third reason is that it is rare in applications to find differences of conditional variances that vary by an order of magnitude. In contrast, it is common to find considerable variation in the propensity score so that the dependence of the optimal region on the conditional variances is likely to be less important. For these reasons, we focus on $\mathbb{A}_H^*$, optimal sets based under homoskedasticity, even though all of the estimation used in making inferences will not maintain this assumption.

The final result in this section concerns the case where we are interested only in the average treatment effect for the treated. In this case, it makes sense to limit the estimand to the average over the subpopulation of the treated with sufficient overlap. Formally, we are interested in the set $\mathbb{A}$ that minimizes

$$\frac{1}{\mathbb{E}[e(X)\cdot 1\{X\in\mathbb{A}\}]^2}\cdot\mathbb{E}\left[e^2(X)\cdot 1\{X\in\mathbb{A}\}\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)\right]. \tag{5.14}$$

We only present the result under homoskedasticity.

Theorem 5.3 Optimal Overlap for the Average Effect for the Treated Under Homoskedasticity *Suppose Assumptions 3.1-3.2 hold. Let* $\underline{f}\leq f_X(x)\leq\overline{f}$. *Then the set* $\mathbb{A}_t^*$ *that minimizes*

$$\frac{1}{\mathbb{E}[e(X)\cdot 1\{X\in\mathbb{A}\}]^2}\cdot\mathbb{E}\left[e^2(X)\cdot 1\{X\in\mathbb{A}\}\cdot\left(\frac{1}{e(X)}+\frac{1}{1-e(X)}\right)\right].$$

is equal to $\mathbb{A}_t^*=\{x\in\mathbb{X}\,|e(x)\leq\alpha_t\}$, *where* $\alpha_t=1$ *if*

$$\sup_{x\in\mathbb{X}}\frac{1}{1-e(x)}\leq 2\cdot\mathbb{E}\left[\frac{1}{1-e(X)}\middle|W=1\right],$$

and α_t *is a solution to*

$$\frac{1}{1-\alpha_t}=2\cdot\mathbb{E}\left[\frac{1}{1-e(X)}\middle|W=1,e(X)\leq\alpha_t\right],$$

otherwise.

Proof: See Appendix.

5.4 The Optimally Weighted Average Treatment Effect

In this section, we consider weighted average treatment effects of the form

$$\tau_{S,\omega}=\sum_{i=1}^N\omega(X_i)\cdot\tau(X_i)\Big/\sum_{i=1}^N\omega(X_i),$$

without requiring $\omega(x)$ to be an indicator function. The following theorem gives the most precisely estimable weighted average treatment effect.

Theorem 5.4 (Optimally Weighted Average Treatment Effect)
Suppose Assumptions 3.1-3.2 hold. Let $\underline{f} \leq f_X(x) \leq \overline{f}$, *and* $\sigma_w^2(x) \leq \overline{\sigma^2}$ *for* $w = 0, 1$ *and all* $x \in \mathbb{X}$. *Then the Optimal Weighted Average Treatment Effect (OWATE) is* τ_{S,ω^*}, *where*

$$\omega^*(x) = \left(\frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1 - e(x)} \right)^{-1}.$$

Proof: See Appendix.

Again the result simplifies under homoskedasticity to an estimand in which the weight functions only depend on the propensity score.

Corollary 5.2 (Optimally Weighted Average Treatment Effect under Homoskedasticity)
Suppose Assumptions 3.1-3.2 hold. Let $\underline{f} \leq f_X(x) \leq \overline{f}$. *Suppose that* $\sigma_w^2(x) = \sigma^2$ *for all* $w \in \{0, 1\}$ *and* $x \in \mathbb{X}$. *Then the Optimally Weighted Average Treatment Effect (OWATE) is* τ_{S,ω_H^*}, *where*

$$\omega_H^*(x) = e(x) \cdot (1 - e(x)).$$

5.5 Numerical Simulations for Optimal Estimands when the Propensity Score follows a Beta Distribution

In this section, we assess the implications of the results derived in the previous sections. We do so by presenting simulations for the optimal estimands when the true propensity score follows a Beta distribution. We study the homoskedastic case, where the optimal cutoff value as well as the ratio of the variances depends only on the (true) marginal distribution of the propensity score. The Beta distribution is characterized by two parameters, here denoted by β and γ, both nonnegative. For a Beta distribution with parameters β and γ, denoted by $\mathcal{B}(\beta, \gamma)$, the mean is $\beta/(\gamma+\beta)$, ranging from zero to one. The corresponding variance is $\beta\gamma/((\gamma+\beta)^2(\gamma+\beta+1))$, which lies between zero and $1/4$. The largest value that this variance takes on is for $\gamma = \beta = 0$, leading to a binomial distribution with probability $1/2$ for both zero and one. We focus on distributions for the true propensity score, where $\beta \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$ and $\gamma \in \{\beta, \ldots, 4\}$.[13] For a given pair of values (β, γ), let $\mathbb{V}_P(\beta, \gamma)$ denote the asymptotic variance of the efficient estimator for the sample average treatment effect, τ_S, which is given by

$$\mathbb{V}_P(\beta, \gamma) = \sigma^2 \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1}{1 - e(X)} \middle| e(X) \sim \mathcal{B}(\beta, \gamma) \right].$$

In addition, let $\mathbb{V}_{P,\alpha}(\beta, \gamma)$ denote the asymptotic variance for the sample average treatment effect, where we drop observations with the propensity score outside the interval $[\alpha, 1 - \alpha]$. This variance is given by

$$\mathbb{V}_{P,\alpha}(\beta, \gamma) = \frac{\sigma^2}{\Pr\left(\alpha \leq e(X) \leq 1 - \alpha | \, e(X) \sim \mathcal{B}(\beta, \gamma)\right)}$$

[13] There is no difference from our perspective between a Beta distribution with parameters γ and β and one with parameters β and γ.

$$\cdot \mathbb{E}\left[\frac{1}{e(X)}+\frac{1}{1-e(X)}\middle|\alpha\leq e(X)\leq 1-\alpha, e(X)\sim\mathcal{B}(\beta,\gamma)\right].$$

Finally, let $\alpha(\beta,\gamma)$ denote the optimal cutoff point for the case where the true propensity score has a Beta distribution with parameters γ and β. We calculate the resulting variances, $\mathbb{V}_{P,\alpha}(\beta,\gamma)$, for the optimal cutoff point and two fixed cutoff values, 0.01, and 0.1. For each of the Beta distributions we report the three ratios

$$\frac{\mathbb{V}_P(\beta,\gamma)}{\mathbb{V}_{P,\alpha(\beta,\gamma)}(\beta,\gamma)}, \qquad \frac{\mathbb{V}_{P,0.01}(\beta,\gamma)}{\mathbb{V}_{P,\alpha(\beta,\gamma)}(\beta,\gamma)}, \qquad \text{and} \quad \frac{\mathbb{V}_{P,0.10}(\beta,\gamma)}{\mathbb{V}_{P,\alpha(\beta,\gamma)}(\beta,\gamma)}.$$

Table 1 presents results for this case. There are two main findings. First, the gain from trimming the sample can be substantial, reducing the asymptotic variance of the average treatment effect estimand by a factor of up to ten for some of the values of the propensity score based on the Beta distribution. Second, discarding observations with a propensity score outside the interval $[0.1, 0.9]$ produces variances that are extremely close those produced with optimally chosen cutoff values. In particular, the ratio of the asymptotic variance when using a cutoff value of 0.1 to the variance based on the optimal cutoff value is never larger than 1.04 over the range of distributions we investigate. In contrast, using the smaller fixed cutoff value of 0.01 can lead to considerably larger variances than using the optimal cutoff value.

6 Inference

6.1 Estimands

In this section, we discuss inference for the estimands introduced in the previous sections. Two issues arise with respect to the tractibility of forming estimators for some of these estimands. First, as we have noted at the end of Section 5.1, there is an important difference in the efficiency bounds for *population* average treatment effects (τ_P) and *sample* average treatment effects (τ_S) that complicate the formation of estimators for the former estimand. In particular, the efficiency bound for τ_P requires one to evaluate $\tau(X)$ over its population distribution, which implies that one must know this distribution or be able to estimate it non-parametrically in order to determine this bound.[14] In general, the distribution of $\tau(X)$ is not known and non-parametrically estimating it with any precision is complicated precisely because of the limited overlap problem. Such complications do not arise if we focus on sample average treatment effects, τ_S. Accordingly, we restrict our attention to chracterizing estimators for the latter class of OSATE and OWATE estimands.

A second issue arises in the case of making inferences concerning Optimal Subpopulation Average Treatment Effects (OSATE). In particular, for this class of estimands, the optimal set, $\mathbb{A}^*$, is generally unknown and must be estimated. Moreover, the efficiency bound in Theorem 5.1 implies that, in some cases, the average effect over any subset of the covariate space defined in terms of the propensity score cannot be estimated at root$-N$ rate. For example, suppose that the set of interest is $\mathbb{A} = \{x \in \mathbb{X} | e(x) \leq p\}$. Note that this corresponds to using the

[14] The same problems arise in the estimation of $\tau_{P,\omega}$.

weight function, $\lambda(e(x)) = 1\{e(x) \leq p\}$, in defining the associated estimand. But, the efficiency bound for this estimand given in (5.10) is a function of $\mathbb{E}[(\frac{\partial}{\partial e}\lambda(e(X)))^2]$ which diverges when $\lambda(e(x)$ is an indicator function so that variance $\mathbb{V}_{P,\lambda}$ is unbounded and cannot be attained. In contrast, such problems do not plague estimation if we focus on the subset $\hat{\mathbb{A}}$, even though, as noted in the discussion of our simple motivating example in Section 2, the sets $\mathbb{A}^*$ and $\hat{\mathbb{A}}$ can be quite different. Accordingly, we also restrict our attention to the subset $\hat{\mathbb{A}}$ when chracterizing estimators for OSATE estimands.

6.2 Nonparametric Estimates for Regression Functions

The proposed estimators for the average treatment effects rely on preliminary estimates of the propensity score and the two conditional regression functions. For these conditional means, various estimators have been proposed (Hahn, 1998; Heckman, Ichimura, Todd, 1998; Hirano, Imbens and Ridder, 2003; Imbens, Newey and Ridder, 2006; Chen, Hong, and Tarozzi, 2005). None of them exactly fits the setting we consider here. Specifically, the previously developed estimators do not allow for estimation of the set over which the treatment effect is averaged. It is possible to modify these estimators to allow for this complication, although doing so would not be trivial. However, it is easier to use the generalized partial mean framework developed by Newey (1994) and extended by Imbens and Ridder (2006).

For simplicity, we use the same type of estimator for both conditional means, namely kernel estimators, although it would be possible to use series estimators for the propensity score as in Hirano, Imbens and Ridder (2003). Let $K : [-1, 1]^L \mapsto \mathbb{R}$ be the kernel and $b > 0$ be the bandwidth. Then the standard kernel estimators for the propensity score, the regression and the variance functions are given by

$$\tilde{e}_b(x) = \sum_{i=1}^{N} W_i \cdot K\left(\frac{X_i - x}{b}\right) \Big/ \sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right),$$

$$\tilde{\mu}_{w,b}(x) = \sum_{i=1}^{N} 1\{W_i = w\} \cdot Y_i \cdot K\left(\frac{X_i - x}{b}\right) \Big/ \sum_{i=1}^{N} 1\{W_i = w\} \cdot K\left(\frac{X_i - x}{b}\right),$$

and

$$\tilde{\sigma}^2_{w,b}(x) = \sum_{i=1}^{N} 1\{W_i = w\} \cdot (Y_i - \tilde{\mu}_w(X_i))^2 \cdot K\left(\frac{X_i - x}{b}\right) \Big/ \sum_{i=1}^{N} 1\{W_i = w\} \cdot K\left(\frac{X_i - x}{b}\right),$$

respectively for $w = 0, 1$. To deal with technical boundary issues and to avoid trimming, it is useful to modify this estimator close to the boundary of the covariate space, using the boundary correction suggested by Imbens and Ridder (2006). The key idea behind this boundary modification is to modify the standard estimator for values of x that are close to the boundary, relative to the bandwidth, by using a Taylor series expansion around the nearest point that is sufficiently far away from the boundary. Details for this modification are presented in the Appendix. The resulting estimators will be denoted by $\hat{e}_{m,b}(x)$, and $\hat{\mu}_{w,m,b}(x)$, where m stands for the degree of the Taylor series expansion.

6.3 Assumptions

Here we list three technical assumptions that will be used to control the convergence rate of the nonparametric estimators. These are closely related to the assumptions used in Imbens and Ridder (2006). The first assumption restricts the kernel.

Assumption 6.1 (Kernel)
(i) $K : \mathbb{R}^L \mapsto \mathbb{R}$,
(ii) $K(u) = 0$ *for* $u \notin \mathbb{U}$, *with* $\mathbb{U} = [-1, 1]^L$,
(iii) K *is* r *times continuously differentiable, with the* r*-th derivative bounded on the interior of* $\mathbb{U}$,
(iv) K *is a kernel of order* s, *so that* $\int_{\mathbb{U}} K(u)du = 1$ *and* $\int_{\mathbb{U}} u^\lambda K(u)du = 0$ *for all* λ *such that* $0 < |\lambda| < s$, *for some* $s \geq 1$.

The second assumption requires sufficient smoothness of the distribution of (Y, W, X).

Assumption 6.2 (Distribution)
(i) $(Y_1, W_1, X_1), (Y_2, W_2, X_2), \ldots,$ *are independent and identically distributed,*
(ii) the support of X_i *is* $\mathbb{X} \subset \mathbb{R}^L$, $\mathbb{X} = \bigotimes_{l=1}^L [\underline{x}_l, \overline{x}_l]$, $\underline{x}_l < \overline{x}_l$ *for all* $l = 1, \ldots, L$.
(iii), X_i *is a random vector with probability density function* $f_X(x)$, *which is* q *times continuously differentiable on the interior of* $\mathbb{X}$, *with the* q*-th derivative bounded,*
(iv) $\sup_{x \in \mathbb{X}} \mathbb{E}[|Y|^p | X = x] < \infty$ *for some* p.
(v) $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$ *is* q *times continuously differentiable on the interior of* $\mathbb{X}$ *with the* q*th derivative bounded for* $w = 0, 1$,
(vi) $\sigma_w^2(x) = \mathbb{V}[Y(w)|X = x] = \mathbb{E}[(Y(w) - \mu_w(X))^2 | X = x]$ *is* q *times continuously differentiable on the interior of* $\mathbb{X}$ *with the* q*th derivative bounded for* $w = 0, 1$,
(vii) $e(x) = \mathbb{E}[W|X = x]$ *is* q *times continuously differentiable on the interior of* $\mathbb{X}$ *with the* q*th derivative bounded,*
(viii) $e(X)$ *has a continuous distribution on* $[0, 1]$ *with the probability density function bounded and continuously differentiable.*

The third assumption puts restrictions on the bandwidth and the smoothness of the kernel and the conditional mean functions.

Assumption 6.3 (Bandwidth and Smoothness)
(i) $b_N = N^{-\delta}$,
(ii) $p > 4$,
(iii) $s > \max(L, (L + 2)/(1 - 2/p))$,
(iv) $q \geq 2s - 1$,
(v) $r \geq s - 1$,
(vi) $1/(2s) < \delta < \min(1/(2L), (1 - 2/p)/(L + 2))$.

6.4 The Optimally Selected Average Treatment Effect

Define, for a given set $\mathbb{A} \subset \mathbb{X}$, the estimator

$$\hat{\tau}_{\mathbb{A}} = \frac{1}{N_{\mathbb{A}}} \sum_{i|X_i \in \mathbb{A}} \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right).$$

In this expression we drop the indexing of the kernel estimators $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ on the bandwidth b_N and the degree of the Taylor series expansion in the boundary correction. The latter will be assumed to be equal to s, the degree of the kernel, and subject to conditions given in Assumption 6.3.

We first characterize some preliminary results in order to define the estimator for the optimal set $\mathbb{A}_H^*$. This involves first estimating the propensity score, and then estimating the optimal cutoff value α. First, define

$$\underline{\hat{\gamma}} = \min_{i=1,\ldots,N} (\hat{e}(X_i)(1-\hat{e}(X_i)))^{-1}, \qquad \text{and} \quad \overline{\hat{\gamma}} = \max_{i=1,\ldots,N} (\hat{e}(X_i)(1-\hat{e}(X_i)))^{-1}.$$

By the support and smoothness conditions, $\underline{\hat{\gamma}}$ and $\overline{\hat{\gamma}}$ exist in large enough samples. For $\Gamma = [0, \infty)$, define the function $\hat{r} : \Gamma \mapsto \mathbb{R}$:

$$\hat{r}(\gamma) = \left(\frac{1}{N} \sum_{i=1}^{N} 1\left\{ \frac{1}{\hat{e}(X_i) \cdot (1-\hat{e}(X_i))} \leq \gamma \right\} \right)^2 \Big/ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{e}(X_i) \cdot (1-\hat{e}(X_i))} \cdot 1\left\{ \frac{1}{\hat{e}(X_i) \cdot (1-\hat{e}(X_i))} \leq \gamma \right\},$$

for $\gamma > \underline{\hat{\gamma}}$, and 0 for $0 \leq \gamma \leq \underline{\hat{\gamma}}$. We are interested in the maximand of $\hat{r}(\gamma)$. To deal with nonuniqueness, we define the maximand as

$$\hat{\Gamma} = \left\{ \gamma \in \Gamma \,\middle|\, \hat{r}(\gamma) \leq \sup_{\gamma \in \Gamma} \hat{r}(\gamma) \right\}, \qquad \text{and} \quad \hat{\gamma} = \sup_{\gamma \in \hat{\Gamma}} \gamma.$$

Then let $\hat{\alpha} = 1/2 - \sqrt{1 - 4/\hat{\gamma}}$. Now we are in a position to define the estimator for the set $\mathbb{A}_H^*$:

$$\hat{\mathbb{A}} = \{ x \in \mathbb{X} \,|\, \hat{\alpha} \leq \hat{e}(x) \leq 1 - \hat{\alpha} \}.$$

Theorem 6.1 (OSATE)
Suppose that Assumptions 3.1-3.2 and 6.1-6.3 hold. Then

$$\sqrt{N} \cdot (\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{S,\hat{\mathbb{A}}}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{q(\mathbb{A}^*)} \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \,\middle|\, X \in \mathbb{A}^* \right] \right).$$

Proof: See Appendix.

A key step in the proof is that

$$\sqrt{N} \cdot \left(\hat{\tau}_{\hat{\mathbb{A}}} - \tau_{S,\hat{\mathbb{A}}} \right) - \sqrt{N} \cdot \left(\hat{\tau}_{\mathbb{A}^*} - \tau_{S,\mathbb{A}_H^*} \right) = o_p(1),$$

which allows us to deal with—or, rather, avoid dealing with—the uncertainty in the estimated set $\hat{\mathbb{A}}$.

Next we consider estimation of the OWATE, based on the weight function $\omega_H^*(x) = e(x) \cdot (1 - e(x))$. Define. for all functions $\omega : \mathbb{X} \mapsto \mathbb{R}$. the estimator

$$\hat{\tau}_\omega = \sum_{i=1}^N \omega(X_i) \cdot \hat{\tau}(X_i) \Big/ \sum_{i=1}^N \omega(X_i).$$

The estimator we actually consider is $\hat{\tau}_{\hat{\omega}}$, where $\hat{\omega}(x) = \hat{e}(x) \cdot (1 - \hat{e}(x))$. We present two results for this estimator. First, we consider the normalized difference between $\hat{\tau}_{\hat{\omega}}$ and $\tau_{S,\hat{\omega}}$. Second, we consider the normalized difference between $\hat{\tau}_{\hat{\omega}}$ and τ_{P,ω_H^*}.

Theorem 6.2 (OWATE)
Suppose that Assumptions 3.1-3.2 and 6.1-6.3 hold. Then

$$\sqrt{N} \cdot (\hat{\tau}_{\hat{\omega}} - \tau_{S,\hat{\omega}}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathbb{E}\left[e(X) \cdot (1 - e(X))\right]^2} \cdot \mathbb{E}\left[(e(X) \cdot (1 - e(X)))^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)}\right)\right]\right).$$

Proof: See Supplementary materials (Crump, Hotz, Imbens and Mitnik, 2006b).

Theorem 6.3 (OWATE)
Suppose that Assumptions 3.1-3.2 and 6.1-6.3 hold. Then

$$\sqrt{N} \cdot (\hat{\tau}_{\hat{\omega}} - \tau_{P,\omega_H^*}) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{P,\omega_H^*}\right),$$

with

$$\mathbb{V}_{P,\omega_H^*} = \frac{1}{\mathbb{E}\left[e(X) \cdot (1 - e(X))\right]^2} \cdot \mathbb{E}\left[(e(X) \cdot (1 - e(X)))^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)}\right)\right] \quad (6.15)$$

$$+ \frac{1}{\mathbb{E}\left[e(X) \cdot (1 - e(X))\right]^2} \cdot \mathbb{E}\left[(e(X) \cdot (1 - e(X)))^2 \cdot \left(\tau(X) - \tau_{P,\omega_H^*}\right)^2\right] \quad (6.16)$$

$$+ \frac{1}{\mathbb{E}\left[e(X) \cdot (1 - e(X))\right]^2} \cdot \mathbb{E}\left[e(X) \cdot (1 - e(X)) \cdot (1 - 2e(X))^2 (\cdot \tau(X) - \tau_{P,\omega_H^*})^2\right]. \quad (6.17)$$

Proof: See Supplementary materials (Crump, Hotz, Imbens and Mitnik, 2006b).

The second term in the expression for $\mathbb{V}_{P,\omega_H^*}$, equation (6.16), takes account of the uncertainty in estimating the population versus sample version, and corresponds to the normalized variance of the difference $\tau_{S,\omega_H^*} - \tau_{P,\omega_H^*}$. The last term, equation (6.17), takes account of the uncertainty in estimating the weights, and corresponds to the normalized variance of the difference $\tau_{P,\hat{\omega}} - \tau_{P,\omega_H^*}$.

6.5 Estimating the Asymptotic Variance

In this subsection, we propose consistent estimators for the asymptotic variances. Define, for all sets $\mathbb{A}$, the folloiwng estimators

$$\hat{q}(\mathbb{A}) = \frac{N_\mathbb{A}}{N},$$

$$\hat{\mathbb{V}}_S(\mathbb{A}) = \frac{1}{\hat{q}(\mathbb{A})} \cdot \frac{1}{N_\mathbb{A}} \sum_{i | X_i \in \mathbb{A}} \left(\frac{\hat{\sigma}_1^2(X_i)}{\hat{e}(X_i)} + \frac{\hat{\sigma}_0^2(X_i)}{1 - \hat{e}(X_i)}\right).$$

Theorem 6.4 *Suppose that Assumptions 6.1-6.3 hold. Then*

$$\hat{\mathbb{V}}_S(\hat{\mathbb{A}}) \xrightarrow{p} \frac{1}{q(\mathbb{A}^*)} \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| X \in \mathbb{A}^* \right].$$

Proof: See Appendix.

Next, let

$$\hat{\mathbb{V}}_{S,\hat{\omega}} = \frac{1}{\left(\frac{1}{N}\sum_{i=1}^N \hat{e}(X_i)\cdot(1-\hat{e}(X_i))\right)^2} \cdot \frac{1}{N}\sum_{i=1}^N (\hat{e}(X_i)\cdot(1-\hat{e}(X_i)))^2 \cdot \left(\frac{\hat{\sigma}_1^2(X_i)}{\hat{e}(X_i)} + \frac{\hat{\sigma}_0^2(X_i)}{1-\hat{e}(X_i)} \right),$$

and

$$\hat{\mathbb{V}}_{P,\hat{\omega}} = \hat{\mathbb{V}}_{S,\hat{\omega}} + \frac{1}{\left(\frac{1}{N}\sum_{i=1}^N \hat{e}(X_i)\cdot(1-\hat{e}(X_i))\right)^2} \cdot \frac{1}{N}\sum_{i=1}^N (\hat{e}(X_i)\cdot(1-\hat{e}(X_i)))^2 \cdot (\hat{\tau}(X_i)-\hat{\tau}_\omega)^2$$

$$+ \frac{1}{\left(\frac{1}{N}\sum_{i=1}^N \hat{e}(X_i)\cdot(1-\hat{e}(X_i))\right)^2} \cdot \frac{1}{N}\sum_{i=1}^N \hat{e}(X_i)\cdot(1-\hat{e}(X_i))\cdot(1-2\cdot\hat{e}(X_i)))^2 \cdot (\hat{\tau}(X_i)-\hat{\tau}_\omega)^2 .$$

Theorem 6.5 *Suppose that Assumptions 6.1-6.3 hold. Then*

$$\hat{\mathbb{V}}_{S,\hat{\omega}} \xrightarrow{p} \frac{1}{\mathbb{E}\left[\omega_H^*(X)\right]^2} \cdot \mathbb{E}\left[(\omega_H^*(X))^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \right) \right].$$

Proof: See Supplementary materials (Crump, Hotz, Imbens and Mitnik, 2006b).

Theorem 6.6 *Suppose that Assumptions 6.1-6.3 hold. Then*

$$\hat{\mathbb{V}}_{P,\hat{\omega}} \xrightarrow{p} \mathbb{V}_{P,\omega_H^*}.$$

Proof: See Supplementary materials (Crump, Hotz, Imbens and Mitnik, 2006b).

7 Some Illustrations Based on Real Data

In this section we apply the methods developed in this paper to data from a labor market program. The data set we use was originally constructed by LaLonde (1986) and subsequently used by, among others, Heckman and Hotz (1989), Dehejia and Wahba (1999) and Smith and Todd (2005). The particular sample we use here is the one used by Dehejia and Wahba (1999). The treatment of interest is a job training program. The trainees are drawn from an experimental evaluation of this program. The control group is a sample drawn from the Panel Study of Income Dynamics (PSID). The control and treatment group are very unbalanced. Table 2 presents some summary statistics. The fourth and fifth column present the averages for each of the covariates separately for the control and treatment group. Consider, for example, the average earnings in the year prior to the program, earn '75. For the control group from the PSID, this is 19.06, in thousands of dollars. For the treatment group, it is only 1.53. Given

that the standard deviation is 13.88, this is a very large difference of 1.26 standard deviations, suggesting that simple covariance adjustments are unlikely to lead to credible inferences.

For these data, we compute and compare 9 different estimands. The first is the sample average treatment effect, τ_S (ATE). We then examine average treatment effects derived over three subsamples. In the first, we drop all observations with an estimated propensity score outside of the interval $[0.01, 0.99]$ ($\text{ATE}_{0.01}$). In the second, we drop all observations with an estimated propensity score outside of the interval $[0.10, 0.90]$ ($\text{ATE}_{0.10}$). Finally, we calculate the estimate of the OSATE with optimal cutoff point, α, using the results in Corollary 5.1. The estimated optimal cutoff point is $\hat{\alpha} = 0.0660$. For these calculations, we estimate the propensity score using a logistic model with all nine covariates displayed in Table 2 entered linearly. We also estimate the optimally weighted average treatment effect (OWATE), with weights $\hat{e}(x) \cdot (1 - \hat{e}(x))$. The final four estimates we consider are all versions of the average treatment effect for the treated. We first estimate the conventional average effect for the treated (ATT). We then form ATT estimates similiar to those in Dehejia and Wabha (1999) by dropping observations which have estimated propensity scores greater than 0.99 ($\text{ATT}_{0.01}$) and 0.90 ($\text{ATE}_{0.10}$), respectively. Finally, we form estimates of the optimal subpopulation average treatment effect on the treated (OSATT) by dropping those observations with an estimated propensity score greater than the optimal cutoff point of 0.73. For each of these cases, we display, in Table 3, estimates of the associated estimands and their asymptotic standard errors. Note that the standard errors are calculated separately for each estimator, implying that implicit estimates of the conditional variance σ^2 are different. Hence, the optimal estimators need not have smaller estimated asymptotic variances than the suboptimal ones.

For both the average treatment effect and the average effect for the treated estimands, it makes a substantial difference to the standard errors of the estimators if we drop observations with propensity scores close to their extreme values. For the average treatment effects, the gain in precision is huge. This is not surprising. There are many control observations whose covariate values are so far from those for the treated that it makes little sense to attempt to estimate the treatment effect for those covariate values. Even for the average effect for the treated however, there is a substantial gain to discarding observations with outlying values for the propensity score. This reduces the asymptotic standard error from 2.58 (with no sample selection) to 1.82 (for the fixed cutoff point of 0.10).

The number of observations that should be discarded according to the OSATE is substantial. We report the number of observations dropped for this estimand in Table 4. Out of the original 2675 observations (2490 controls and 185 treated), only 312 are used in estimation (183 controls and 129 treated). We also report in Table 4 the number of observations dropped in the various categories for this criterion and for the suboptimal criteria based on the fixed cutoff points 0.01 ($\text{ATE}_{0.01}$) and 0.10 ($\text{ATE}_{0.10}$), respectively, in the subsequent two panels of this table.

While not the primary focus of our analysis, we also note that the estimates of the various estimands, themselves, vary substantially. This is not surprising, given that the definitions of the underlying estimands are varying. They even differ in sign. At the same time, we make two observations about these estimates. First, the standard errors relative to the estimates tend to be large for all of the alternative estimates, implying that the inferences drawn from them

would not differ across the estimates. Second, the OSATE, OWATE and OSATT estimates are all negative and tend to be closer in magnitude to one another compared to the other estimators. One should not draw strong conclusions from either of this observations, given that the theoretical results established in this paper are focused primarily on the precision of alternative estimands.

8 Conclusion

Estimation of average treatment effects under unconfoundedness or selection on observables is often hampered by lack of overlap in the covariate distributions. This lack of overlap can lead to imprecise estimates and can make commonly used estimators sensitive to the choice of specification. In such cases, researchers have often used informal methods for trimming the sample. In this paper, we develop a systematic approach to addressing such lack of overlap in which we sacrifice some external validity in exchange for improved internal validity. We characterize optimal subsamples where the average treatment effect can be estimated most precisely, as well as optimally weighted average treatment effects. Under some simplifying assumptions, the optimal rules depend solely on the propensity score. We find that the precision for average treatment effects for the optimally selected samples can be much higher than for the overall sample. In addition, we find that a simple *ad hoc* selection rule based on discarding all units with an estimated propensity score outside the interval $[0.1, 0.9]$ can capture most of the precision gains from selecting the sample optimally for a wide range of distributions.

Appendix A: The Kernel Estimator with Boundary Correction

In this appendix we present the details of the boundary correction we use for the kernel estimator. This boundary correction was developed by Imbens and Ridder (2006). We refer to this paper for more details on the estimator. Let $g(x) = \mathbb{E}[Y|X = x]$ be the regression function of interest, and let $f_X(x)$ be the probability density function of X, with the dimension of X equal to L. Then we can write $g(x) = h_1(x)/h_2(x)$, where $h_1(x) = g(x) \cdot f_X(x)$, and $h_2(x) = f_X(x)$. If we define $Y_1 = Y$ and $Y_2 = 1$, then we can write $h_k(x) = \mathbb{E}[Y_k|X = x] \cdot f_X(x)$, with the standard kernel estimator for $h_k(x)$ equal to

$$\tilde{h}_{k,b}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{b^L} Y_{ki} \cdot K\left(\frac{X_i - x}{b}\right).$$

Let $\partial\mathbb{X}$ be the boundary of $\mathbb{X}$, and let $\mathbb{X}_I$ be the "internal" region, more than b_N away from the boundary in all directions, $\mathbb{X}_I = \{x \in \mathbb{X} | \min_{l=1,\ldots,L} \inf_{y\in\partial\mathbb{X}} |y_l - x_l| \geq b_N\}$. Then let $r_b(x)$ be the projection of x onto the set $\mathbb{X}_I$: $r_b(x) = \arg\min_{y\in\mathbb{X}_I} \|x - y\|\}$. Let λ denote an L vector of nonnegative integers, with $|\lambda| = \sum_{l=1}^{L} \lambda_l$, and $\lambda! = \prod_{l=1}^{L} \lambda_l!$. Define for a given, $m-1$ times differentiable function $g : \mathbb{R}^L \to \mathbb{R}$, a point $y \in \mathbb{R}^L$ and an integer m, the $m-1$-th order polynomial function $t : \mathbb{R}^L \to \mathbb{R}$ based on the Taylor series expansion of order $m-1$ of $g(\cdot)$ around the point y:

$$t(x; g(\cdot), y, m) = \sum_{j=0}^{m-1} \sum_{|\lambda|=j} \frac{1}{\lambda!} \frac{\partial^{|\lambda|}}{\partial x^\lambda} g(y) \cdot (x - y)^\lambda. \tag{A.1}$$

Now we define the boundary corrected estimators for $h_k(x)$:

$$\hat{h}_{k,m,b}(x) = \begin{cases} \tilde{h}_{k,b}(x) & \text{if } x \in \mathbb{X}_I \\ t\left(x, \tilde{h}_{k,b}, r_b(x), m\right) & \text{elsewhere.} \end{cases}$$

Finally the boundary corrected estimator for $g(x)$ is

$$\hat{g}_{m,b}(x) = \hat{h}_{1,m,b}(x)/\hat{h}_{2,m,b}(x).$$

Appendix B: Proofs

Proof of Theorem 5.1: The derivation of the efficiency bound follows that of Hahn (1998) and Hirano, Imbens and Ridder (2003). The density of $(Y(0), Y(1), W, X)$ with respect to some σ-finite measure is

$$\begin{aligned} q(y(0), y(1), w, x) &= f(y(0), y(1)|w, x) \cdot f(w|x) \cdot f(x) \\ &= f(y(0), y(1)|x) \cdot f(w|x) \cdot f(x) \\ &= f(y(0), y(1)|x) \cdot e(x)^w \cdot (1 - e(x))^{1-w} \cdot f(x), \end{aligned}$$

where in the second equality we used unconfoundedness. The density of the observed data (y, w, x) is

$$q(y, w, x) = f_1(y|x)^w \cdot e(x)^w \cdot f_0(y|x)^{1-w} \cdot (1 - e(x))^{1-w} \cdot f(x),$$

where $f_w(y|x) = f_{Y(W)|X}(y(w)|x) = \int f(y(1-w), y|x)dy(1-w)$. Consider a regular parametric submodel indexed by θ, with density

$$q(y, w, x|\theta) = f_1(y|x, \theta)^w \cdot e(x|\theta)^w \cdot f_0(y|x, \theta)^{1-w} \cdot (1 - e(x|\theta))^{1-w} \cdot f(x|\theta),$$

which is equal to the true density $q(y, w, x)$ for $\theta = \theta_0$, or $q(y, w, x) = q(y, w, x|\theta_0)$. The score for the parametric model is given by

$$\mathcal{S}(y, w, x|\theta) = \frac{\partial}{\partial\theta} \ln q(y, w, x|\theta) = w \cdot \mathcal{S}_1(y|x, \theta) + (1 - w) \cdot \mathcal{S}_0(y|x, \theta) + \mathcal{S}_x(x|\theta) + \frac{w - e(x|\theta)}{e(x|\theta)(1 - e(x|\theta))} \cdot e'(x|\theta)$$

where

$$\mathcal{S}_1(y|x, \theta) = \frac{\partial}{\partial\theta} \ln f_1(y|x, \theta), \qquad \text{and} \quad \mathcal{S}_0(y|x, \theta) = \frac{\partial}{\partial\theta} \ln f_0(y|x, \theta),$$

$$\mathcal{S}_x(x|\theta) = \frac{\partial}{\partial\theta}\ln f(x|\theta), \qquad \text{and} \quad e'(x|\theta) = \frac{\partial}{\partial\theta}\, e(x|\theta).$$

The tangent space of the model is the set of functions $t(y,w,x)$ of the form

$$\mathcal{T} = \{w\cdot\mathcal{S}_1(y,x) + (1-w)\cdot\mathcal{S}_0(y,x) + \mathcal{S}_x(x) + a(x)\cdot(w-e(x))\}$$

where $a(x)$ is any square-integrable measurable function of x and $\mathcal{S}_1$, $\mathcal{S}_0$, and $\mathcal{S}_x$ satisfy

$$\int \mathcal{S}_1(y,x)f_1(y|x)dy = \mathbb{E}\left[\mathcal{S}_1(Y(1),X)|\, X=x\right] = 0, \quad \forall x,$$
$$\int \mathcal{S}_0(y,x)f_0(y|x)dy = \mathbb{E}\left[\mathcal{S}_0(Y(0),X)|\, X=x\right] = 0, \quad \forall x,$$
$$\int \mathcal{S}_x(x)f(x)dx = \mathbb{E}\left[\mathcal{S}_x(X)\right] = 0.$$

The parameter of interest is

$$\tau_{P,\lambda} = \frac{\iint \lambda(e(x))yf_1(y|x)f(x)dydx - \iint \lambda(e(x))yf_0(y|x)f(x)dydx}{\int \lambda(e(x))f(x)dx}.$$

Thus, for the parametric submodel indexed by θ, the parameter of interest is

$$\tau_{P,\lambda}(\theta) = \frac{\iint \lambda(e(x|\theta))yf_1(y|x,\theta)f(x|\theta)dydx - \iint \lambda(e(x|\theta))yf_0(y|x,\theta)f(x|\theta)dydx}{\int \lambda(e(x|\theta))f(x|\theta)dx}.$$

We need to find a function $\psi(y,w,x)$ such that for all regular parametric submodels,

$$\frac{\partial\tau_{P,\lambda}(\theta_0)}{\partial\theta} = \mathbb{E}\left[\,\psi(Y,W,X)\cdot\mathcal{S}(Y,W,X|\,\theta_0)\right]. \tag{B.1}$$

First, we will calculate $\frac{\partial}{\partial\theta}\tau_{P,\lambda}(\theta_0)$. Let $\mu_\lambda = \int\lambda(e(x))f(x)dx$. Then,

$$\begin{aligned}\frac{\partial}{\partial\theta}\tau_{P,\lambda}(\theta_0) &= \frac{1}{\mu_\lambda}\left[\iint \lambda(e(x|\theta_0))y\left[\mathcal{S}_1(y|x,\theta_0)f_1(y|x,\theta_0) - \mathcal{S}_0(y|x,\theta_0)f_0(y|x,\theta_0)\right]f(x|\theta_0)dydx\right.\\ &\quad + \int\lambda(e(x|\theta_0))\left[\tau(x)-\tau_{P,\lambda}\right]\mathcal{S}_x(x|\theta_0)f(x|\theta_0)dx\\ &\quad \left.+\int\lambda'(e(x|\theta_0))e'(x|\theta_0)\left[\tau(x)-\tau_{P,\lambda}\right]f(x|\theta_0)dx\right]\end{aligned}$$

where $\lambda'(e(x)) = \frac{\partial}{\partial e(x)}\lambda(e(x))$. The following choice for $\psi(y,w,x)$ is shown in the supplementary materials (Crump, Hotz, Imbens and Mitnik, 2006b) to satisfy the condition:

$$\begin{aligned}\psi(y,w,x) &= \frac{w\cdot\lambda(e(x))}{\mu_\lambda\cdot e(x)}(y-\mathbb{E}[Y(1)|X=x]) - \frac{(1-w)\cdot\lambda(e(x))}{\mu_\lambda\cdot(1-e(x))}(y-\mathbb{E}[Y(0)|X=x])\\ &\quad + \frac{\lambda(e(x))}{\mu_\lambda}(\tau(x)-\tau_{P,\lambda}) + \frac{(w-e(x))\cdot\lambda'(e(x))}{\mu_\lambda}(\tau(x)-\tau_{P,\lambda}).\end{aligned}$$

Then by Theorem 2 in Section 3.3 of Bickel, Klaassen, Ritov, and Wellner (1993), the variance bound is the expected square of the projection of $\psi(Y,W,X)$ on the tangent space $\mathcal{T}$. Since $\psi(y,w,x)\in\mathcal{T}$, the variance bound is

$$\begin{aligned}\mathbb{E}[\psi(Y,W,X)^2] &= \mathbb{E}\left[\frac{[\lambda(e(X))]^2}{(\mu_\lambda)^2\cdot e(X)}\cdot\sigma_1^2(X)\right] + \mathbb{E}\left[\frac{[\lambda(e(X))]^2}{(\mu_\lambda)^2\cdot(1-e(X))}\cdot\sigma_0^2(X)\right]\\ &\quad + \mathbb{E}\left[\frac{[\lambda(e(X))+(W-e(X))\cdot\lambda'(e(X))]^2}{(\mu_\lambda)^2}(\tau(X)-\tau_{P,\lambda})^2\right]\\ &= \mathbb{E}\left[\frac{[\lambda(e(X))]^2}{(\mu_\lambda)^2\cdot e(X)}\cdot\sigma_1^2(X)\right] + \mathbb{E}\left[\frac{[\lambda(e(X))]^2}{(\mu_\lambda)^2\cdot(1-e(X))}\cdot\sigma_0^2(X)\right]\\ &\quad + \mathbb{E}\left[\frac{[\lambda(e(X))]^2 + e(X)(1-e(X))\cdot[\lambda'(e(X))]^2}{(\mu_\lambda)^2}(\tau(X)-\tau_{P,\lambda})^2\right]\end{aligned}$$

For the special case of $\lambda(e(x)) = e(x)$ (a case considered by Hahn, 1998) the semiparametric efficiency bound is,

$$\mathbb{E}\left[\frac{e(X)}{(\mu_\lambda)^2}\cdot\sigma_1^2(X)\right]+\mathbb{E}\left[\frac{e(X)^2}{(\mu_\lambda)^2\cdot(1-e(X))}\cdot\sigma_0^2(X)\right]+\mathbb{E}\left[\frac{e(X)}{(\mu_\lambda)^2}(\tau(X)-\tau_{P,\lambda})^2\right].$$

For the special case of $\lambda(e(x)) = e(x)(1-e(x))$ the semiparametric efficiency bound is,

$$\begin{aligned}&\mathbb{E}\left[\frac{e(X)(1-e(X))^2}{(\mu_\lambda)^2}\cdot\sigma_1^2(X)\right]+\mathbb{E}\left[\frac{e(X)^2(1-e(X))}{(\mu_\lambda)^2}\cdot\sigma_0^2(X)\right]\\&\quad+\mathbb{E}\left[\frac{e(X)^2(1-e(X))^2+e(X)(1-e(X))\cdot(1-2\cdot e(X))^2}{(\mu_\lambda)^2}(\tau(X)-\tau_{P,\lambda})^2\right]\end{aligned}$$

which simplifies to

$$\begin{aligned}&\mathbb{E}\left[\frac{e(X)(1-e(X))^2}{(\mu_\lambda)^2}\cdot\sigma_1^2(X)\right]+\mathbb{E}\left[\frac{e(X)^2(1-e(X))}{(\mu_\lambda)^2}\cdot\sigma_0^2(X)\right]\\&\quad+\mathbb{E}\left[\frac{e(X)(1-e(X))(3e(X)^2-3e(X)+1)}{(\mu_\lambda)^2}(\tau(X)-\tau_{P,\lambda})^2\right].\end{aligned}$$

□

Define

$$\tau_{S,\omega}(\mathbb{A}) = \sum_{i|X_i\in\mathbb{A}}\tau(X_i)\cdot\omega(X_i)\Big/\sum_{i|X_i\in\mathbb{A}}\omega(X_i),$$

for nonnegative functions $\omega(\cdot)$. For estimands of this type consider the criterion that encompasses Theorems 5.2 and 5.3:

$$\mathbb{V}_{S,\omega}(\mathbb{A}) = \frac{1}{\mathbb{E}[\omega(X)\cdot 1\{X\in\mathbb{A}\}]^2}\cdot\mathbb{E}\left[\omega(X)^2\cdot 1\{X\in\mathbb{A}\}\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)\right]. \tag{B.2}$$

We are interested in the choice of set $\mathbb{A}$ that minimizes (B.2) among the set of all closed subsets of $\mathbb{X}$. The following theorem provides the characterization.

Theorem B.1 (Weighted OSATE)
Let $\underline{f} \le f(x) \le \overline{f}$, and $\sigma^2(x) \le \overline{\sigma^2}$ for $w = 0, 1$ and all $x \in \mathbb{X}$, and let $\omega : \mathbb{X} \mapsto \mathbb{R}^+$ be continuously differentiable. The set $\mathbb{A}^*$ that minimizes (B.2) is equal to $\mathbb{X}$ if

$$\sup_{x\in\mathbb{X}}\omega(x)\cdot\left(\frac{\sigma_1^2(x)}{e(x)}+\frac{\sigma_0^2(x)}{1-e(x)}\right)\le 2\cdot\mathbb{E}\left[\omega^2(X)\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)\right]\Big/\mathbb{E}\left[\omega(X)\right],$$

and otherwise,

$$\mathcal{A}^* = \left\{x\in\mathbb{X}\,\middle|\,\omega(x)\cdot\left(\frac{\sigma_1^2(x)}{e(x)}+\frac{\sigma_0^2(x)}{1-e(x)}\right)\le\gamma\right\},$$

where γ is a positive solution to

$$\gamma = 2\cdot\frac{\mathbb{E}\left[\omega^2(X)\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)\middle|\omega(X)\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)<\gamma\right]}{\mathbb{E}\left[\omega(X)\middle|\omega(X)\cdot\left(\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\right)<\gamma\right]}.$$

Proof: Define $k(x) = \sigma_1^2(x)/e(x) + \sigma_0^2(x)/(1-e(x))$, $\tilde{f}_X(x) = f_X(x)\cdot\omega(x)/\int_z f_X(z)\cdot\omega(z)dz$, and $\tilde{\omega}(x) = \omega(x)/\int_z f_X(z)\cdot\omega(z)dz$, so that $k(x)$ is bounded, bounded away from zero, and continuously differentiable on $\mathbb{X}$. Let $\tilde{X}$ be a random vector with probability density function $\tilde{f}_X(x)$ on $\mathbb{X}$, and let $\tilde{q}(\mathbb{A}) = \Pr(\tilde{X}\in\mathbb{A})$.[15] Then

$$\begin{aligned}\mathbb{E}\left[\tilde{\omega}(X)\cdot 1\{X\in\mathbb{A}\}\right] &= \int_x\tilde{\omega}(x)\cdot 1\{x\in\mathbb{A}\}\cdot f_X(x)dx\\&= \int_x\frac{\omega(x)}{\int_z\omega(z)\cdot f_X(z)dz}\cdot 1\{x\in\mathbb{A}\}\cdot f_X(x)dx\end{aligned}$$

[15] Note that $\int\tilde{f}_X(x)dx = 1$ by construction, so that $\tilde{f}_X(x)$ is a valid probability density function.

$$= \int_x 1\{x \in \mathbb{A}\} \cdot \tilde{f}_X(x)dx$$
$$= \mathbb{E}\left[1\{\tilde{X} \in \mathbb{A}\}\right] = \Pr\left(\tilde{X} \in \mathbb{A}\right) = \tilde{q}(\mathbb{A}),$$

and similarly,

$$\mathbb{E}\left[\tilde{\omega}(X)^2 \cdot 1\{X \in \mathbb{A}\} \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right)\right] = \mathbb{E}\left[\tilde{\omega}(X)^2 \cdot 1\{X \in \mathbb{A}\} \cdot k(X)\right]$$
$$= \int_x \tilde{\omega}(x)^2 \cdot 1\{x \in \mathbb{A}\} \cdot k(x) \cdot f_X(x)dx$$
$$= \int_x \tilde{\omega}(x) \cdot \frac{\omega(x)}{\int_z \omega(z) \cdot f_X(z)dz} \cdot 1\{x \in \mathbb{A}\} \cdot k(x) \cdot f_X(x)dx$$
$$= \int_x \tilde{\omega}(x) \cdot 1\{x \in \mathbb{A}\} \cdot k(x) \cdot \tilde{f}_X(x)dx$$
$$= \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot 1\{\tilde{X} \in \mathbb{A}\} \cdot k(\tilde{X})\right].$$

Because multiplying $\omega(x)$ by a constant does not change the value of the objective function in (B.2), we have

$$\mathbb{V}_{S,\omega}(\mathbb{A}) = \mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}) = \frac{1}{\mathbb{E}[\tilde{\omega}(X) \cdot 1\{X \in \mathbb{A}\}]^2} \cdot \mathbb{E}\left[\tilde{\omega}(X)^2 \cdot 1\{X \in \mathbb{A}\} \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right)\right]$$
$$= \frac{1}{\tilde{q}^2(\mathbb{A}))} \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot 1\{\tilde{X} \in \mathbb{A}\} \cdot k(\tilde{X})\right].$$
$$= \frac{1}{\tilde{q}(\mathbb{A})} \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| 1\{\tilde{X} \in \mathbb{A}\}\right]. \tag{B.3}$$

Thus the question now concerns the set $\mathbb{A}$ that minimizes (B.3).
We do the remainder of the proof of Theorem B.1 in two stages. First, suppose there is a closed set $\mathbb{A}$ such that $x \in \text{int}(\mathbb{A})$, $z \notin \mathbb{A}$, and $\tilde{\omega}(z) \cdot k(z) < \tilde{\omega}(x) \cdot k(x)$. Then we will construct a closed set $\tilde{\mathbb{A}}$ such that $\mathbb{V}_{S,\tilde{\omega}}(\tilde{\mathbb{A}}) < \mathbb{V}_{S,\tilde{\omega}}(\mathbb{A})$. This implies that the optimal set has the form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \tilde{\omega}(x) \cdot k(x) \leq \gamma\},$$

for some γ. The second step consists of deriving the optimal value for γ.
For the first step define a ball around x with volume ν,

$$\mathcal{B}_\nu(x) = \{z \in \mathbb{X} | \|z - x\| \leq \nu^{1/L} 2^{-1/L} \pi^{-1/2} \Gamma(L/2)^{1/L}\},$$

where $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)dx$ is the gamma function. Let $\mathbb{A}^c$ be the complement of $\mathbb{A}$ in $\mathbb{X}$, and for sets $\mathbb{A}$ and $\mathcal{B}$ let $\mathbb{A}/\mathcal{B} = \mathbb{A} \cap \mathcal{B}^c$. Let ν_0 be small enough so that for $\nu \leq \nu_0$ we have $\mathcal{B}_{\nu/\tilde{f}_X(x)}(x) \subset \mathbb{A}$, $\mathcal{B}_{\nu/\tilde{f}_X(z)}(z) \subset \mathbb{A}^c$ and $\tilde{\omega}(z') \cdot k(z') < \tilde{\omega}(x') \cdot k(x')$ for all $z' \in \mathcal{B}_{\nu/\tilde{f}_X(z)}(z)$ and all $x' \in \mathcal{B}_{\nu/\tilde{f}_X(x)}(x)$. Also, because the volume of the sets $\mathcal{B}_{\nu/\tilde{f}_X(x)}(x)$ and $\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)$ is $\nu/\tilde{f}_X(x)$ and $\nu/\tilde{f}_X(z)$ respectively, it follows that

$$\tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(x)}(x)) - \nu = o(\nu),$$
$$\tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)) - \nu = o(\nu),$$

so that

$$\tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(x)}(x)) - \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)) = o(\nu).$$

Now we construct the set

$$\tilde{\mathbb{A}}_\nu = \left(\mathbb{A}/\mathcal{B}_{\nu/\tilde{f}_X(x)}(x)\right) \cup \mathcal{B}_{\nu/\tilde{f}_X(z)}(z).$$

The objective function for this set is

$$\mathbb{V}_{S,\tilde{\omega}}(\tilde{\mathbb{A}}_\nu) = \frac{1}{\tilde{q}(\tilde{\mathbb{A}}_\nu)} \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| 1\{\tilde{X} \in \tilde{\mathbb{A}}_\nu\}\right]$$

$$= \frac{\tilde{q}(\mathbb{A}) \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathbb{A}\right]}{\left(\tilde{q}(\mathbb{A}) + \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)) - \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(x)}(x))\right)^2}$$

$$+ \frac{\tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)) \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(z)}(z)\right] - \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(x)}(x)) \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(x)}(x)\right]}{\left(\tilde{q}(\mathbb{A}) + \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(z)}(z)) - \tilde{q}(\mathcal{B}_{\nu/\tilde{f}_X(x)}(x))\right)^2}$$

$$= \frac{\tilde{q}(\mathbb{A}) \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathbb{A}\right] + \nu \cdot \left(\mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(z)}(z)\right] - \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(x)}(x)\right]\right)}{q(\mathbb{A})^2} + o(\nu),$$

so that the difference relative to the value of the objective function for the original set $\mathbb{A}$ is

$$\mathbb{V}_{S,\tilde{\omega}}(\tilde{\mathbb{A}}_\nu) - \mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}) = \frac{1}{q(\mathbb{A})^2} \cdot \nu \cdot \left(\mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(z)}(z)\right] - \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(x)}(x)\right]\right) + o(\nu).$$

Since $\mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(z)}(z)\right] - \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{X} \in \mathcal{B}_{\nu/\tilde{f}_X(x)}(x)\right] < 0$ if $\nu \leq \nu_0$, the difference $\mathbb{V}_{S,\tilde{\omega}}(\tilde{\mathbb{A}}_\nu) - \mathbb{V}_{S,\tilde{\omega}}(\mathbb{A})$ is negative for small enough ν, which finishes the first part of the proof.
The question now is to determine the optimal value for γ given that the optimal set has the form

$$\mathbb{A}_\gamma = \{x \in \mathbb{X} | \tilde{\omega}(x) \cdot k(x) \leq \gamma\}.$$

Let $Y = \tilde{\omega}(\tilde{X}) \cdot k(\tilde{X})$, with probability density function $f_Y(y)$. Then

$$\mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}'_\gamma)) = \frac{\mathbb{E}[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) | \tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) < \gamma']}{q(\mathbb{A}'_\gamma)} = \frac{\mathbb{E}[Y|Y < \gamma']}{\Pr(Y < \gamma')} = \frac{\int_0^{\gamma'} y \cdot f_Y(y) dy}{\left(\int_0^{\gamma'} f_Y(y) dy\right)^2}.$$

Denote the minimum and maximum value of the function $k(x)$ over the set $\mathbb{X}$ by $\underline{k}$ and $\overline{k}$. By assumption $\underline{k} > 0$ and $\overline{k} < \infty$. Then $\lim_{\gamma \downarrow \underline{k}} \to \infty$. Because $\mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}_{\overline{k}}) = \mathbb{V}_{S,\tilde{\omega}}(\mathbb{X})$ which is finite by assumption, and because $\mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}_{\overline{k}})$ is continuous as a function of γ, it follows that either $\mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}_{\overline{k}})$ is minimized at $\gamma = \overline{k}$, or there is an interior minimum where the first order conditions are satisfied. Let γ' denote the optimum.
The first derivative with respect to γ is

$$\frac{\partial}{\partial \gamma} \mathbb{V}_{S,\tilde{\omega}}(\mathbb{A}_\gamma) = \frac{\left(\int_0^\gamma f_Y(y) dy\right)^2 \cdot \gamma \cdot f_Y(\gamma) - 2 \int_0^\gamma f_Y(y) dy \cdot \int_0^\gamma y f_Y(y) dy \cdot f_Y(\gamma)}{\left(\int_0^\gamma f_Y(y) dy\right)^4}.$$

This is zero if

$$\gamma' \cdot \int_0^{\gamma'} f_Y(y) dy = 2 \int_0^{\gamma'} y \cdot f_Y(y) dy,$$

implying

$$\gamma' = 2 \cdot \frac{\int_0^{\gamma'} y \cdot f_Y(y) dy}{\int_0^{\gamma'} f_Y(y) dy} = 2 \cdot \mathbb{E}[Y|Y < \gamma']$$

$$= 2 \cdot \mathbb{E}[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) | \tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) < \gamma'].$$

Because $\tilde{\omega}(x) = \omega(x) / \int_z \omega(z) \cdot f_X(z) dz$,

$$\gamma' = 2 \cdot \mathbb{E}\left[\tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) \middle| \tilde{\omega}(\tilde{X}) \cdot k(\tilde{X}) < \gamma'\right],$$

implies

$$\gamma = 2 \cdot \mathbb{E}\left[\omega(\tilde{X}) \cdot k(\tilde{X}) \middle| \omega(\tilde{X}) \cdot k(\tilde{X}) < \gamma\right],$$

for $\gamma = \gamma' \cdot \int \omega(x) \cdot f_X(x) dx$. This in turn implies

$$\gamma = 2 \cdot \mathbb{E}\left[\omega(\tilde{X}) \cdot k(\tilde{X}) \cdot 1\left\{\omega(\tilde{X}) \cdot k(\tilde{X}) < \gamma\right\}\right] \Big/ \Pr\left(\omega(\tilde{X}) \cdot k(\tilde{X}) < \gamma\right)$$

$$
\begin{aligned}
&= 2 \cdot \frac{\int_x \omega(x) \cdot k(x) \cdot 1\{\omega(x) \cdot k(x) < \gamma\} \tilde{f}_X(x) dx}{\int_x 1\{\omega(x) \cdot k(x) < \gamma\} \tilde{f}_X(x) dx} \\
&= 2 \cdot \frac{\int_x \omega(x) \cdot k(x) \cdot 1\{\omega(x) \cdot k(x) < \gamma\} \cdot \frac{\omega(x)}{\int_z \omega(z) \cdot f_X(z) dz} f_X(x) dx}{\int_x 1\{\omega(x) \cdot k(x) < \gamma\} \cdot \frac{\omega(x)}{\int_z \omega(z) \cdot f_X(z) dz} f_X(x) dx} \\
&= 2 \cdot \frac{\int_x \omega^2(x) \cdot k(x) \cdot 1\{\omega(x) \cdot k(x) < \gamma\} f_X(x) dx}{\int_x \omega(x) \cdot 1\{\omega(x) \cdot k(x) < \gamma\} f_X(x) dx} \\
&= 2 \cdot \frac{\mathbb{E}\left[\omega^2(X) \cdot k(X) \cdot 1\{\omega(X) \cdot k(X) < \gamma\}\right]}{\mathbb{E}\left[\omega(X) \cdot 1\{\omega(X) \cdot k(X) < \gamma\}\right]}. \\
&= 2 \cdot \frac{\mathbb{E}\left[\omega^2(X) \cdot k(X) \left|\omega(X) \cdot k(X) < \gamma\right.\right]}{\mathbb{E}\left[\omega(X) \left|\omega(X) \cdot k(X) < \gamma\right|\right]}.
\end{aligned}
$$

Substituting back $k(x) = \sigma_1^2(x)/e(x) + \sigma_0^2(x)/(1-e(x))$ this implies

$$
\gamma = 2 \cdot \frac{\mathbb{E}\left[\omega^2(X) \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right) \middle| \omega(X) \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right) < \gamma\right]}{\mathbb{E}\left[\omega(X) \middle| \omega(X) \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right) < \gamma\right]}.
$$

□

Proof of Theorem 5.2: Substituting $\omega(x) = 1$ into Theorem B.1 implies that the optimal set $\mathbb{A}^*$ is equal to $\mathbb{X}$ if

$$
\sup_{x \in \mathbb{X}} \frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1-e(x)} \leq 2 \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right],
$$

and otherwise,

$$
\mathbb{A}^* = \left\{ x \in \mathbb{X} \middle| \frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1-e(x)} \leq \gamma \right\},
$$

where γ is a positive solution to

$$
\gamma = 2 \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} < \gamma\right].
$$

Then define $\alpha = 1/2 - \sqrt{1/4 - 1/\gamma}$, so that $\gamma = (\alpha(1-\alpha))^{-1}$ and

$$
\mathbb{A}^* = \left\{ x \in \mathbb{X} \middle| \frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1-e(x)} \leq \frac{1}{\alpha \cdot (1-\alpha)} \right\},
$$

where α is a positive solution to

$$
\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} < \frac{1}{\alpha \cdot (1-\alpha)}\right].
$$

□

Proof of Theorem 5.3: Substituting $\omega(x) = e(x)$ and $\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2$ into Theorem B.1 implies that the optimal $\mathbb{A}^*$ is equal to $\mathbb{X}$ if

$$
\sup_{x \in \mathbb{X}} \frac{\sigma^2}{1-e(x)} \leq 2 \cdot \mathbb{E}\left[\frac{\sigma^2 \cdot e(X)}{1-e(X)}\right] \Big/ \mathbb{E}\left[e(X)\right], \tag{B.4}
$$

and otherwise,

$$
\mathbb{A}^* = \left\{ x \in \mathbb{X} \middle| \frac{\sigma^2}{1-e(x)} \leq \gamma \right\},
$$

where γ is a positive solution to

$$
\gamma = 2 \cdot \frac{\mathbb{E}\left[\frac{\sigma^2 \cdot e(X)}{1-e(X)} \middle| \frac{\sigma^2}{1-e(X)} < \gamma\right]}{\mathbb{E}\left[e(X) \middle| \frac{\sigma^2}{1-e(X)} < \gamma\right]}. \tag{B.5}
$$

Condition (B.4) is equivalent to

$$\sup_{x\in\mathbb{X}} \frac{1}{1-e(x)} \leq 2\cdot\mathbb{E}\left[\frac{e(X)}{1-e(X)}\right] \Big/ \mathbb{E}\left[e(X)\right] = 2\cdot\mathbb{E}\left[\frac{1}{1-e(X)}\middle| W=1\right].$$

Let $\alpha_t = 1-\sigma^2/\gamma$, so that $\gamma = \sigma^2/(1-\alpha)$. Then (B.5) implies

$$\frac{1}{1-\alpha_t} = 2\cdot\frac{\mathbb{E}\left[\frac{e(X)}{1-e(X)}\middle|\frac{1}{1-e(X)}<\frac{1}{1-\alpha_t}\right]}{\mathbb{E}\left[e(X)\middle|\frac{1}{1-e(X)}<\frac{1}{1-\alpha_t}\middle|\right]}$$

$$= 2\cdot\frac{\mathbb{E}\left[\frac{e(X)}{1-e(X)}\,|e(X)\leq\alpha_t\right]}{\mathbb{E}\left[e(X)\,|e(X)\leq\alpha_t\right]} = 2\cdot\mathbb{E}\left[\frac{1}{1-e(X)}\middle| W=1, e(X)\leq\alpha_t\right].$$

□

Proof of Theorem 5.4:
We are choosing $\omega:\mathbb{X}\to\mathbb{R}$ to minimize

$$\mathbb{V}_{S,\omega} = \frac{1}{\mathbb{E}[\omega(X)]^2}\cdot\mathbb{E}\left[\frac{\omega(X)^2}{e(X)}\sigma_1^2(X)+\frac{\omega(X)^2}{1-e(X)}\sigma_0^2(X)\right].$$

Again let $k(x) = \frac{\sigma_1^2(x)}{e(x)}+\frac{\sigma_0^2(x)}{1-e(x)}$, so that we minimize

$$\mathbb{V}_{S,\omega} = \frac{1}{\mathbb{E}[\omega(X)]^2}\cdot\mathbb{E}\left[\omega(X)^2k(X)\right] = \frac{\int\omega^2(x)k(x)f(x)dx}{\left(\int\omega(x)f(x)dx\right)^2}.$$

If $\tilde{\omega}(x)$ is a solution, than so is $\tilde{\omega}(x)/c$. Hence we can normalize $\omega(x)$ to satisfy $\int\omega(x)f(x)dx=1$. Then the problem is the minimization of

$$\int\omega^2(x)k(x)f(x)dx \qquad \text{s.t. } \int\omega(x)f(x)dx=1.$$

The solution to this satisfies $0=2\cdot\omega(x)k(x)f(x)-\lambda f(x)$, so that for some constant c, $\omega(x)=c/k(x)$. Hence the optimal weights $\omega^*(x)$ are proportional to $1/k(x)$, and since we do not care about the constant of proportionality we can choose

$$\omega^*(x) = \frac{1}{k(x)} = \frac{e(x)\cdot(1-e(x))}{(1-e(x))\cdot\sigma_1^2(x)+e(x)\cdot\sigma_0^2(x)}.$$

□
Define for sets $\mathbb{A}\subset\mathbb{X}$,

$$N_{\mathbb{A}} = \sum_{i=1}^N 1\{X_i\in\mathbb{A}\},$$

$$\tau_{S,\mathbb{A}} = \begin{cases} \frac{1}{N_{\mathbb{A}}}\sum_{i=1}^N 1\{X_i\in\mathbb{A}\}\cdot(\tau(X_i)) & \text{if } N_{\mathbb{A}}>0,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$\hat{\tau}_{\mathbb{A}} = \begin{cases} \frac{1}{N_{\mathbb{A}}}\sum_{i=1}^N 1\{X_i\in\mathbb{A}\}\cdot(\hat{\mu}_1(X_i)-\hat{\mu}_0(X_i)) & \text{if } N_{\mathbb{A}}>0,\\ 0 & \text{otherwise.}\end{cases}$$

Lemma B.1 *Suppose that* $\sup_{x\in\mathbb{X},w\in\{0,1\}}|\hat{\mu}_w(x)-\mu_w(x)| = o_p(N^{-1/2+\varepsilon})$. *Then for all* $\mathbb{A}\subset\mathbb{X}$,

$$\hat{\tau}_{\mathbb{A}}-\tau_{S,\mathbb{A}} = o_p\left(N^{-1/2+\varepsilon}\right).$$

Proof:

$$\hat{\tau}_{\mathbb{A}}-\tau_{S,\mathbb{A}} = 1\{N_{\mathbb{A}}>0\}\cdot\frac{1}{N_{\mathbb{A}}}\sum_{i=1}^N 1\{X_i\in\mathbb{A}\}\cdot(\hat{\tau}(X_i)-\tau(X_i)) \leq \sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|$$

$$\leq \sup_{x\in\mathbb{X}} |\hat{\mu}_1(x) - \mu_1(x)| + \sup_{x\in\mathbb{X}} |\hat{\mu}_0(x) - \mu_0(x)| = o_p\left(N^{-1/2+\varepsilon}\right).$$

□

Define

$$\underline{\gamma} = \inf_{x\in\mathbb{X}} (e(x)(1-e(x)))^{-1}, \qquad \overline{\gamma} = \sup_{x\in\mathbb{X}} (e(x)(1-e(x)))^{-1},$$

$$\hat{\underline{\gamma}} = \min_{i=1,\ldots,N} (e(X_i)(1-e(X_i)))^{-1}, \qquad \text{and} \quad \hat{\overline{\gamma}} = \max_{i=1,\ldots,N} (e(X_i)(1-e(X_i)))^{-1},$$

For $\Gamma = [0,\infty)$ define the functions $r : \Gamma \mapsto \mathbb{R}$ and $\hat{r} : \Gamma \mapsto \mathbb{R}$:

$$r(\gamma) = \left(\int 1\left\{ \frac{1}{e(x)\cdot(1-e(x))} \leq \gamma \right\} \cdot f_X(x)dx \right)^2 \Big/ \int \frac{1}{e(x)\cdot(1-e(x))} \cdot 1\left\{ \frac{1}{e(x)\cdot(1-e(x))} \leq \gamma \right\} \cdot f_X(x)dx,$$

for $\gamma > \underline{\gamma}$, and 0 for $0 \leq \gamma \leq \underline{\gamma}$, and

$$\hat{r}(\gamma) = \left(\frac{1}{N}\sum_{i=1}^N 1\left\{ \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \leq \gamma \right\} \right)^2 \Big/ \frac{1}{N}\sum_{i=1}^N \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \cdot 1\left\{ \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \leq \gamma \right\},$$

for $\gamma > \hat{\underline{\gamma}}$, and 0 for $0 \leq \gamma \leq \hat{\underline{\gamma}}$. Define

$$\Gamma^* = \left\{ \gamma \in \Gamma | r(\gamma) \leq \sup_{\gamma\in\Gamma} r(\gamma) \right\}, \qquad \text{and} \quad \gamma^* = \sup_{\gamma\in\Gamma^*} \gamma.$$

and

$$\hat{\Gamma} = \left\{ \gamma \in \Gamma | \hat{r}(\gamma) \leq \sup_{\gamma\in\Gamma} \hat{r}(\gamma) \right\}, \qquad \text{and} \quad \hat{\gamma} = \sup_{\gamma\in\hat{\Gamma}} \gamma.$$

Lemma B.2 *Suppose that* $\sup_{x\in\mathbb{X}} |e(x) - \hat{e}(x)| = o_p(N^{-\alpha})$, *and that* $\inf_{x\in\mathbb{X}} e(x)\cdot(1-e(x)) > 0$. *Suppose also that if* $\gamma^* < \overline{\gamma}$, *then* $\Gamma^* = \{\gamma^*\}$, *and* $\frac{\partial^2}{\partial\gamma^2} r(\gamma^*) < 0$. *Then for any* $\delta > 0$,

$$\hat{\gamma} - \gamma^* = o_p\left(N^{-\alpha/2+\delta}\right).$$

Proof: Define

$$p(\gamma) = \int \frac{1}{e(x)\cdot(1-e(x))} \cdot 1\left\{ \frac{1}{e(x)\cdot(1-e(x))} \leq \gamma \right\} \cdot f_X(x)dx,$$

$$\hat{p}(\gamma) = \frac{1}{N}\sum_{i=1}^N \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \cdot 1\left\{ \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \leq \gamma \right\},$$

$$q(\gamma) = \int 1\left\{ \frac{1}{e(x)\cdot(1-e(x))} \leq \gamma \right\} \cdot f_X(x)dx,$$

$$\hat{q}(\gamma) = \frac{1}{N}\sum_{i=1}^N 1\left\{ \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \leq \gamma \right\},$$

$$r(\gamma) = q^2(\gamma)/p(\gamma),$$

and

$$\hat{r}(\gamma) = \hat{q}^2(\gamma)/\hat{p}(\gamma),$$

The proof consists of three parts. First we show that

$$\sup_{\gamma\in\Gamma} |\hat{p}(\gamma) - p(\gamma)| = O_p\left(N^{-\alpha}\right), \qquad \text{and} \quad \sup_{\gamma\in\Gamma} |\hat{q}(\gamma) - q(\gamma)| = O_p\left(N^{-\alpha}\right). \tag{B.6}$$

In the second step we show that this implies that

$$\sup_{\gamma\in\Gamma} |\hat{r}(\gamma) - r(\gamma)| = O_p\left(N^{-\alpha}\right). \tag{B.7}$$

In the third step we show that this in turn implies for any $\delta > 0$ that

$$\hat{\gamma} - \gamma^* = \arg\max_{\gamma \in \Gamma} \hat{r}(\gamma) - \arg\max_{\gamma \in \Gamma} r(\gamma) = o_p\left(N^{-\alpha/2+\delta}\right). \tag{B.8}$$

First consider the convergence of $\hat{q}(\gamma)$. Define $k(x) = \frac{1}{e(x)\cdot(1-e(x))}$, and $\hat{k}(x) = \frac{1}{\hat{e}(x)\cdot(1-\hat{e}(x))}$, so that $q(\gamma) = \int 1\{k(x) \le \gamma\} \cdot f_X(x)dx$, and $\hat{q}(\gamma) = \frac{1}{N}\sum_{i=1}^N 1\left\{\hat{k}(X_i) \le \gamma\right\}$. By the Triangle Inequality,

$$\sup_{\gamma\in\Gamma} |\hat{q}(\gamma) - q(\gamma)| \le \sup_{\gamma\in\Gamma} \left|\hat{q}(\gamma) - \frac{1}{N}\sum_{i=1}^N 1\{k(X_i) \le \gamma\}\right| \tag{B.9}$$

$$+ \sup_{\gamma\in\Gamma} \left|\frac{1}{N}\sum_{i=1}^N 1\{k(X_i) \le \gamma\} - q(\gamma)\right|. \tag{B.10}$$

As the difference between the distribution function and the empirical distribution function of $k(X)$,

$$\sup_{\gamma\in\Gamma} \left|\frac{1}{N}\sum_{i=1}^N 1\{k(X_i) \le \gamma\} - q(\gamma)\right| = O_p\left(N^{-1/2}\right),$$

e.g., Billingsley (1985). Next, consider the righthand side of (B.9):

$$\begin{aligned}
&\sup_{\gamma\in\Gamma} \left|\frac{1}{N}\sum_{i=1}^N 1\left\{\hat{k}(X_i) \le \gamma\right\} - \frac{1}{N}\sum_{i=1}^N 1\{k(X_i) \le \gamma\}\right| \\
&\quad \le \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N \left|1\left\{\hat{k}(X_i) \le \gamma\right\} - 1\{k(X_i) \le \gamma\}\right| \\
&\quad \le \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N \left(1\left\{\hat{k}(X_i) \le \gamma \le k(X_i)\right\} + 1\left\{k(X_i) \le \gamma \le \hat{k}(X_i)\right\}\right).
\end{aligned} \tag{B.11}$$

Define the indicator

$$B_N = 1\left\{\sup_{x\in\mathbb{X}} |\hat{k}(x) - k(x)| \ge N^{-\alpha}\right\},$$

so that $B_N = o_p(1)$. Then (B.11) can be written as

$$B_N \cdot \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N \left(1\left\{\hat{k}(X_i) \le \gamma \le k(X_i)\right\} + 1\left\{k(X_i) \le \gamma \le \hat{k}(X_i)\right\}\right). \tag{B.12}$$

$$+(1 - B_N) \cdot \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N \left(1\left\{\hat{k}(X_i) \le \gamma \le k(X_i)\right\} + 1\left\{k(X_i) \le \gamma \le \hat{k}(X_i)\right\}\right). \tag{B.13}$$

Because B_N is binary and $o_p(1)$, it follows that (B.12) is $o_p(N^{-\alpha})$. If $B_N = 0$, then $\hat{k}(x) \le \gamma \le k(x)$ or $k(x) \le \gamma \le \hat{k}(x)$ both imply $|k(x) - \gamma| < N^{-\alpha}$, so that (B.13) can be bounded by

$$(1 - B_N) \cdot 2 \cdot \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N 1\left\{|k(X_i) - \gamma| < N^{-\alpha}\right\} \le 2 \cdot \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N 1\left\{|k(X_i) - \gamma| < N^{-\alpha}\right\}. \tag{B.14}$$

This is the sum of independent and identically distributed binary random variables with mean bounded by $C_0 \cdot N^{-\alpha}$, implying by Markov's inequality that (B.14) is $O_p(N^{-\alpha})$:

$$\begin{aligned}
&\Pr\left(N^\alpha \cdot 2 \cdot \sup_{\gamma\in\Gamma} \frac{1}{N}\sum_{i=1}^N 1\left\{|k(X_i) - \gamma| < N^{-\alpha}\right\} > C\right) \le \frac{2N^\alpha C_0 \cdot N^{-\alpha}}{C} \\
&\quad = 2 \cdot C_0/C,
\end{aligned}$$

which can be made arbitrarily small by choosing C large. This finishes the proof that (B.9) is $O_p(N^{-\alpha})$. and thus that $\sup_{\gamma\in\Gamma} |\hat{q}(\gamma) - q(\gamma)| = O_p(N^{-\alpha})$. The proof for the claim that $\sup_{\gamma\in\Gamma} |\hat{p}(\gamma) - p(\gamma)| = O_p(N^{-\alpha})$ is similar and is omitted.

Next consider (B.7). This follows directly from the convergence of $\hat{p}(\gamma)$ and $\hat{q}(\gamma)$ to $p(\gamma)$ and $q(\gamma)$ respectively. Finally, consider (B.8). Let $a = -\frac{\partial^2}{\partial\gamma^2} r(\gamma^*) > 0$. Let $\Gamma_0 = \{\gamma \in \Gamma | \frac{\partial^2}{\partial\gamma^2} r(\gamma) < -a/2\}$, so that $\gamma^* \in \text{int}(\Gamma_0)$, and let $\Gamma_N = \{\gamma \in \Gamma | |\gamma - \gamma^*| < N^{-\alpha/2}\}$. For $N > N_0$, $\Gamma_N \subset \Gamma_0$. Let $r_0 = \sup_{\gamma \in \Gamma/\Gamma_0} r(\gamma)$. Then $r_0 < r(\gamma^*) = \sup_{\gamma\in\Gamma} r(\gamma)$. Define the two events

$$A_N = 1\{\inf_{\gamma\in\Gamma} |\hat{r}(\gamma) - r(\gamma)| > |r_0 - r(\gamma^*)|/2\},$$

and

$$B_N = 1\{\inf_{\gamma\in\Gamma} |\hat{r}(\gamma) - r(\gamma)| > (a/8)N^{-\alpha}\}.$$

For $N > N_1$, B_N implies A_N. Since $\Pr(B_N = 1) \to 0$, it follows that $B_N = o_p(N^{-\alpha})$.
Let $N > \max(N_0, N_1)$, and consider $\gamma \in \Gamma_0/\Gamma_N$. We will show that for such γ, $|r(\gamma^*) - r(\gamma)| = r(\gamma^*) - r(\gamma) > (a/4) \cdot N^{-\alpha}$. Suppose $\gamma > \gamma^*$. First note that for $c \in \Gamma_0$, $c > \gamma^*$, it follows that

$$\frac{\partial}{\partial\gamma} r(c) = \frac{\partial^2}{\partial\gamma^2} r(\tilde{c}) \cdot (c - \gamma^*) < -(a/2) \cdot (c - \gamma^*).$$

Hence for $\gamma > \gamma^*$,

$$\begin{aligned} r(\gamma) &= r(\gamma^*) + \int_{\gamma^*}^{\gamma} \frac{\partial}{\partial c} r(c) dc \\ &< r(\gamma^*) - \int_{\gamma^*}^{\gamma} (a/2) \cdot (c - \gamma^*) dc \\ &= r(\gamma^*) - \frac{a}{4}(\gamma - \gamma^*)^2. \end{aligned}$$

Because $\gamma \notin \Gamma_N$, $|\gamma - \gamma^*| \geq N^{-\alpha/2}$ so that

$$r(\gamma) - r(\gamma^*) < -\frac{a}{4} \cdot N^{-\alpha},$$

and thus

$$|r(\gamma) - r(\gamma^*)| = r(\gamma^*) - r(\gamma) > |a/4| \cdot N^{-\alpha}.$$

Therefore, if $B_N = 0$, it must be that $\gamma \notin \Gamma_N$ implies

$$\hat{r}(\gamma) \leq r(\gamma) + (a/8)N^{-\alpha} < r(\gamma^*) - (a/8)N^{-\alpha} \leq \hat{r}(\gamma^*),$$

and thus $\hat{\gamma} \in \Gamma_N$. Finally, write

$$\hat{\gamma} - \gamma^* = B_N \cdot (\hat{\gamma} - \gamma^*) + (1 - B_N) \cdot (\hat{\gamma} - \gamma^*). \tag{B.15}$$

The first term on the righthand side is $o_p(N^{-\alpha/2})$ because B_N is binary and $o_p(1)$, and the second term is $o_p(N^{-\alpha/2+\delta})$ because if $B_N = 0$, then $|\hat{\gamma} - \gamma^*| < N^{-\alpha/2}$. Thus (B.15) is $o_p(N^{-\alpha/2+\delta})$. □
Define

$$A_N = 1 - 1\left\{\sup_{x\in\mathbb{X}} \left| \frac{1}{\hat{e}(x) \cdot (1 - \hat{e}(x))} - \frac{1}{e(x) \cdot (1 - e(x))} \right| \leq N^{-1/2+\epsilon}, |\hat{\gamma} - \gamma| \leq N^{-1/4+\varepsilon/2+\delta} \right\},$$

and

$$\tilde{\tau}(\mathbb{A}) = (1 - A_N) \cdot \hat{\tau}(\mathbb{A}).$$

Lemma B.3 *Suppose that for some* $\varepsilon, \delta > 0$ $\sup_{x\in\mathbb{X}, w\in\{0,1\}} |\hat{\mu}_w(x) - \mu_w(x)| = o_p(N^{-1/2+\epsilon})$, $\sup_{x\in\mathbb{X}} |\hat{e}(x) - e(x)| = o_p(N^{-1/2+\epsilon})$, *and that* $\inf_{x\in\mathbb{X}} e(x) \cdot (1 - e(x)) > 0$. *Then, for all sets* $\mathbb{A} \subset \mathbb{X}$,

$$\tilde{\tau}(\mathbb{A}) - \hat{\tau}(\mathbb{A}) = o_p\left(N^{-1/2}\right).$$

Proof: First we show that $A_N = o_p(1)$. By the assumptions and Lemma B.2 it follows that $\hat{\gamma} - \gamma = O_p\left(N^{-1/2+\varepsilon/2+\delta}\right)$. Thus

$$\begin{aligned}&\Pr(A_N = 1)\\ &\quad \leq \Pr\left(\sup_{x\in\mathbb{X}}\left|\frac{1}{\hat{e}(x)\cdot(1-\hat{e}(x))} - \frac{1}{e(x)\cdot(1-e(x))}\right| > N^{-1/2+\varepsilon}\right) + \Pr\left(|\hat{\gamma}-\gamma| > N^{-1/4+\varepsilon/2+\delta}\right) = o(1).\end{aligned}$$

Second,

$$\Pr\left(N^{1/2}\cdot|\tilde{\tau}(\mathbb{A}) - \hat{\tau}(\mathbb{A})| > C\right) = \Pr\left(N^{1/2}\cdot A_N\cdot|\hat{\tau}(\mathbb{A})| > C\right) \leq \Pr(A_N = 1) = o(1).$$

so that $\tilde{\tau}(\mathbb{A}) - \hat{\tau}(\mathbb{A}) = o_p(N^{-1/2})$. □

Lemma B.4 *Suppose that* $\sup_{x\in\mathbb{X}}|\hat{e}(x)-e(x)| = o_p(N^{-1/2+\varepsilon})$*, and that* $\inf_{x\in\mathbb{X}} e(x)\cdot(1-e(x)) > 0$*. Then for any* $\delta > 0$*, (i),* $\frac{N_{\hat{\mathbb{A}}}-N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$*, (ii),* $\frac{N_{\hat{\mathbb{A}}/\mathbb{A}^*}}{N_{\mathbb{A}^*}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$*, (iii),* $\frac{N_{\mathbb{A}^*/\hat{\mathbb{A}}}}{N_{\mathbb{A}^*}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$*, (iv),* $\frac{N_{\hat{\mathbb{A}}}-N_{\mathbb{A}^*}}{N_{\hat{\mathbb{A}}}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$*, (v),* $\frac{N_{\hat{\mathbb{A}}/\mathbb{A}^*}}{N_{\hat{\mathbb{A}}}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$*, (vi),* $\frac{N_{\mathbb{A}^*/\hat{\mathbb{A}}}}{N_{\hat{\mathbb{A}}}} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$.

Proof: We prove (i). The other claims follow the same argument. Define $B_{Ni} = 1\{|k(X_i)-\gamma| < 2N^{-1/4+\varepsilon/2+\delta}\}$. Then for all N the random variables B_{Ni} and B_{Nj} are independent and identically distributed with mean bounded by $C\cdot N^{-1/4+\varepsilon/2+\delta}$. Hence $\sum_{i=1}^N B_{Ni}/N$ is $O_p(N^{-1/4+\varepsilon/2+\delta})$. Now

$$\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}} = A_N\cdot\left(\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}}\right) + (1-A_N)\cdot\left(\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}}\right). \tag{B.16}$$

Because

$$\Pr\left(N^{1/2}\cdot A_N\cdot\left(\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}}\right) > C\right) \leq \Pr(A_N > 0) = o(1),$$

it follows that the first term of the right hand side of (B.16) is $o_p(N^{-1/2})$. Next, consider the second term of the right hand side of (B.16):

$$\begin{aligned}(1-A_N)\cdot\left|\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}}\right| &= \frac{1-A_N}{N_{\mathbb{A}}^*}\cdot\left|\sum_{i=1}^N 1\{\hat{k}(X_i)\leq\hat{\gamma}\} - 1\{k(X_i)\leq\gamma\}\right|\\ &\leq \frac{1-A_N}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N\left|1\{\hat{k}(X_i)\leq\hat{\gamma}\} - 1\{k(X_i)\leq\gamma\}\right|\\ &\leq \frac{1-A_N}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N 1\{\hat{k}(X_i)\leq\hat{\gamma}, \gamma < k(X_i)\} + 1\{k(X_i)\leq\gamma, \hat{\gamma} < \hat{k}(X_i)\}\\ &= \frac{1}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N(1-A_N)\cdot B_{Ni}\cdot 1\{\hat{k}(X_i)\leq\hat{\gamma}, \gamma < k(X_i)\} \qquad \text{(B.17)}\\ &\quad + \frac{1}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N(1-A_N)\cdot(1-B_{Ni})\cdot 1\{\hat{k}(X_i)\leq\hat{\gamma}, \gamma < k(X_i)\} \qquad \text{(B.18)}\\ &\quad + \frac{1}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N(1-A_N)\cdot B_{Ni}\cdot 1\{k(X_i)\leq\gamma, \hat{\gamma} < \hat{k}(X_i)\} \qquad \text{(B.19)}\\ &\quad + \frac{1}{N_{\mathbb{A}}^*}\cdot\sum_{i=1}^N(1-A_N)\cdot(1-B_{Ni})\cdot 1\{k(X_i)\leq\gamma, \hat{\gamma} < \hat{k}(X_i)\}. \qquad \text{(B.20)}\end{aligned}$$

$A_N = 0$ implies $|\hat{k}(x) - k(x)| < N^{-1/2+\varepsilon}$ and $|\gamma - \hat{\gamma}| < N^{-1/4+\varepsilon/2+\delta}$, so that $\hat{k}(X_i)\leq\hat{\gamma}$ and $\gamma < k(X_i)$ combined with $A_N = 0$ implies that $\gamma < k(x) < \gamma + 2N^{-1/4+\varepsilon/2+\delta}$ and thus $B_{Ni} = 1$. Hence (B.18) is equal to zero, and similarly (B.20) is equal to zero. Thus

$$(1-A_N)\cdot\left|\frac{N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*}}{N_{\mathbb{A}^*}}\right|$$

$$\leq \frac{1}{N^*_{\mathbb{A}}} \cdot \sum_{i=1}^{N} (1 - A_N) \cdot B_{Ni} \cdot 1\{\hat{k}(X_i) \leq \hat{\gamma}, \gamma < k(X_i)\}$$

$$+ \frac{1}{N^*_{\mathbb{A}}} \cdot \sum_{i=1}^{N} (1 - A_N) \cdot B_{Ni} \cdot 1\{k(X_i) \leq \gamma, \hat{\gamma} < \hat{k}(X_i)\}$$

$$\leq \frac{1}{N^*_{\mathbb{A}}} \cdot \sum_{i=1}^{N} B_{Ni} = \frac{N}{N^*_{\mathbb{A}}} \cdot \sum_{i=1}^{N} B_{Ni}/N = O_p(1) \cdot O_p\left(N^{-1/4+\varepsilon/2+\delta}\right) = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right).$$

Combined with the fact that the first term of the right hand side of (B.16) is $o_p(N^{-1/2})$, this implies that $(N_{\hat{\mathbb{A}}} - N_{\mathbb{A}^*})/N_{\mathbb{A}^*} = O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)$.
The other parts of the Lemma follow by similar arguments. For that reason their proofs are omitted. □

Lemma B.5 *Suppose that* $\sup_{x\in\mathbb{X}} |e(x) - \hat{e}(x)| = o_p(N^{-1/2+\varepsilon})$ *for some* $\varepsilon < 1/6$*, and that* $\inf_{x\in\mathbb{X}} e(x)\cdot(1-e(x)) > 0$*. Suppose also that if* $\gamma^* < \overline{\gamma}$*, then* $\Gamma^* = \{\gamma^*\}$*, and* $\frac{\partial^2}{\partial\gamma^2} r(\gamma^*) < 0$*. Then*

$$\left(\hat{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\hat{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*)) = o_p\left(N^{-1/2}\right).$$

Proof: Note that by Lemma B.2 for any $\delta > 0$ we have $\hat{\gamma} - \gamma^* = o_p(N^{-1/4+\varepsilon/2+\delta})$. Because $\varepsilon < 1/6$, it follows that $-3/4 + \varepsilon(3/2) < -1/2$, and so we can choose δ so that $-3/4 + \varepsilon(3/2) + \delta < -1/2$. Next,

$$\left(\hat{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\hat{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*))$$

$$= \frac{N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\hat{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\hat{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*)) + \frac{N^*_{\mathbb{A}} - N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\hat{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right)$$

$$= \left(\hat{\tau}(\hat{\mathbb{A}}) - \tilde{\tau}(\hat{\mathbb{A}})\right) \tag{B.21}$$

$$+ \frac{N_{\hat{\mathbb{A}}} - N^*_{\mathbb{A}}}{N^*_{\mathbb{A}}} \cdot \left(\hat{\tau}(\hat{\mathbb{A}}) - \tilde{\tau}(\hat{\mathbb{A}})\right) \tag{B.22}$$

$$- (\hat{\tau}(\mathbb{A}^*) - \tilde{\tau}(\mathbb{A}^*)) \tag{B.23}$$

$$+ \frac{N^*_{\mathbb{A}} - N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\hat{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) \tag{B.24}$$

$$+ \frac{N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\tilde{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\tilde{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*)). \tag{B.25}$$

By Lemma B.3 (B.21) and (B.23) are $o_p(N^{-1/2})$. By Lemma B.3 the second factor in (B.22) is $o_p(N^{-1/2})$, and by Lemma B.4(i) the first factor is $O_p(N^{-1/4+\varepsilon/2+\delta})$, hence the product (B.22) is $o_p(N^{-1/2})$. Next, consider (B.24). The first factor is $O_p(N^{-1/4+\varepsilon/2+\delta})$ by Lemma B.4(i). The second factor is $o_p(N^{-1/2+\varepsilon})$ by Lemma B.1, so the product is $o_p(N^{-3/4+\varepsilon(3/2)+\delta}) = o_p(N^{-1/2})$ because δ can be chosen to satisfy $\delta + \varepsilon(3/2) < 1/4$.
Finally, consider (B.25):

$$\frac{N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\tilde{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\tilde{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*))$$

$$= A_N \cdot \left(\frac{N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\tilde{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\tilde{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*))\right) \tag{B.26}$$

$$+ (1 - A_N) \cdot \left(\frac{N_{\hat{\mathbb{A}}}}{N^*_{\mathbb{A}}} \cdot \left(\tilde{\tau}(\hat{\mathbb{A}}) - \tau(\hat{\mathbb{A}})\right) - (\tilde{\tau}(\mathbb{A}^*) - \tau(\mathbb{A}^*))\right). \tag{B.27}$$

Because A_N is binary and $o_p(1)$, it follows that $A_N = o_p(N^{-1/2})$, and thus (B.26) is $o_p(N^{-1/2})$. Next, (B.27) is equal to

$$(1 - A_N) \cdot \left(\frac{1}{N^*_{\mathbb{A}}} \sum_{i=1}^{N} 1\{X_i \in \hat{\mathbb{A}}\} \cdot (\hat{\tau}(X_i) - \tau(X_i)) - \frac{1}{N^*_{\mathbb{A}}} \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}^*\} \cdot (\hat{\tau}(X_i) - \tau(X_i))\right)$$

$$= \frac{1 - A_N}{N^*_{\mathbb{A}}} \sum_{i=1}^{N} 1\{X_i \in \hat{\mathbb{A}}/\mathbb{A}^*\} \cdot (\hat{\tau}(X_i) - \tau(X_i)) \tag{B.28}$$

$$-\frac{1-A_N}{N^*_{\mathbb{A}}}\sum_{i=1}^{N}1\{X_i\in\mathbb{A}^*/\hat{\mathbb{A}}\}\cdot(\hat{\tau}(X_i)-\tau(X_i))\,. \tag{B.29}$$

Consider (B.28):

$$\left|\frac{1-A_N}{N^*_{\mathbb{A}}}\sum_{i=1}^{N}1\{X_i\in\hat{\mathbb{A}}/\mathbb{A}^*\}\cdot(\hat{\tau}(X_i)-\tau(X_i))\right|\leq\frac{1}{N^*_{\mathbb{A}}}\sum_{i=1}^{N}1\{X_i\in\hat{\mathbb{A}}/\mathbb{A}^*\}\cdot|\hat{\tau}(X_i)-\tau(X_i)|$$

$$\leq\frac{1}{N^*_{\mathbb{A}}}\sum_{i=1}^{N}1\{X_i\in\hat{\mathbb{A}}/\mathbb{A}^*\}\cdot\sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|$$

$$=\frac{N_{\hat{\mathbb{A}}/\mathbb{A}^*}}{N^*_{\mathbb{A}}}\cdot\sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|=O_p\left(N^{-1/4+\varepsilon/2+\delta}\right)\cdot o_p\left(N^{-1/2+\varepsilon}\right)=O_p\left(N^{-3/4+\varepsilon(3/2)+\delta}\right)=o_p\left(N^{-1/2}\right).$$

(B.29) is $o_p(N^{-1/2})$ by the same argument. □

Lemma B.6 *Suppose that for some* $0<\varepsilon<1/4$, $\sup_{x\in\mathbb{X},w\in\{0,1\}}|\hat{\mu}_w(x)-\mu_w(x)|=o_p(N^{-1/2+\varepsilon})$ *and* $\sup_{x\in\mathbb{X}}|\hat{e}(x)-e(x)|=o_p(N^{-1/2+\varepsilon})$, *and that* $\inf_{x\in\mathbb{X}}e(x)\cdot(1-e(x))>0$. *Then for* $\lambda(x)=e(x)\cdot(1-e(x))$ *and* $\hat{\lambda}(x)=\hat{e}(x)\cdot(1-\hat{e}(x))$,

$$(\hat{\tau}_{\hat{\lambda}}-\tau_{\hat{\lambda}})-(\hat{\tau}_{\lambda}-\tau_{\lambda})=o_p\left(N^{-1/2}\right).$$

Proof: By the rate of the convergence of $\hat{e}(x)$ to $e(x)$, it follows that $\sup_{x\in\mathbb{X}}|\hat{\lambda}(x)-\lambda(x)|=o_p\left(N^{-1/2+\varepsilon}\right)$, and in combination with the positive lower bound on $\lambda(x)$ it follows that

$$\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\right)^{-1}-\left(\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\right)^{-1}=o_p\left(N^{-1/2+\varepsilon}\right).$$

By the rate of the convergence of $\hat{\mu}_w(x)$ to $\mu_w(x)$, it follows that $\sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|=o_p\left(N^{-1/2+\varepsilon}\right)$. Then

$$|(\hat{\tau}_{\hat{\lambda}}-\tau_{\hat{\lambda}})-(\hat{\tau}_{\lambda}-\tau_{\lambda})|=\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)}-\frac{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)}\right|$$

$$=\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)}-\frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)}\right|$$

$$+\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)}-\frac{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))}{\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)}\right|$$

$$\leq\left|\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\cdot(\hat{\tau}(X_i)-\tau(X_i))\right|\cdot\left|\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\right)^{-1}-\left(\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\right)^{-1}\right|$$

$$+\left(\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\right)^{-1}\cdot\left|\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}(X_i)-\lambda(X_i)\right)\cdot(\hat{\tau}(X_i)-\tau(X_i))\right|$$

$$\leq\sup_{x\in\mathbb{X}}\lambda(x)\cdot\sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|\cdot\left|\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}(X_i)\right)^{-1}-\left(\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\right)^{-1}\right|$$

$$+\inf_{x\in\mathbb{X}}\lambda(x)^{-1}\cdot\sup_{x\in\mathbb{X}}\left|\hat{\lambda}(x)-\lambda(x)\right|\cdot\sup_{x\in\mathbb{X}}|\hat{\tau}(x)-\tau(x)|$$

$$=o_p\left(N^{-1/2+\varepsilon}\right)\cdot o_p\left(N^{-1/2+\varepsilon}\right)+o_p\left(N^{-1/2+\varepsilon}\right)\cdot o_p\left(N^{-1/2+\varepsilon}\right)=o_p\left(N^{-1/2}\right).$$

□

Define

$$\theta=\mathbb{E}\left[\lambda(X)\cdot(\mu_1(X)-\mu_0(X))\right],\qquad\theta_S=\frac{1}{N}\sum_{i=1}^{N}\lambda(X_i)\cdot(\mu_1(X_i)-\mu_0(X_i))\,,$$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \lambda(X_i)\cdot(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))\,.$$

and

$$\phi(y,w,x) =$$
$$\lambda(x)\cdot\left(\left(\frac{y\cdot w}{e(x)} - \mu_1(x)\right) - \left(\frac{y\cdot(1-w)}{1-e(x)} - \mu_0(x)\right) - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)}\right)\cdot(w-e(x))\right).$$

Lemma B.7 (Asymptotic Linearity)
Suppose Assumptions 3.1-3.2 and 6.1-6.3 hold. Then

$$\sqrt{N}\cdot\left(\hat{\theta} - \theta_S\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi(Y_i, W_i, X_i) + o_p(1).$$

Proof: We apply Theorem 4.1 in Imbens and Ridder (2006). Define the vector $\tilde{Y}$ as

$$\tilde{Y} = \begin{pmatrix} Y\cdot W \\ Y\cdot(1-W) \\ W \end{pmatrix},$$

and define the functions $\omega : \mathbb{X}\to\mathbb{R}$, $g:\mathbb{X}\to\mathbb{R}^3$, and $m:\mathbb{R}^3\to\mathbb{R}$:

$$\omega(x) = \lambda(x),$$
$$g(x) = \mathbb{E}[\tilde{Y}|X=x] = \begin{pmatrix} g_1(x) \\ g_2(x) \\ g_3(x) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y\cdot W|X=x] \\ \mathbb{E}[Y\cdot(1-W)|X=x] \\ \mathbb{E}[W|X=x] \end{pmatrix} = \begin{pmatrix} \mu_1(x)\cdot e(x) \\ \mu_0(x)\cdot(1-e(x)) \\ e(x) \end{pmatrix},$$
$$m(z) = z_1/z_3 - z_2/(1-z_3).$$

Then

$$\theta = \mathbb{E}[\omega(X)\cdot m(g(X))], \qquad \theta_S = \frac{1}{N}\sum_{i=1}^{N}\omega(X_i)\cdot m(g(X_i)),$$

and

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\omega(X_i)\cdot m(\hat{g}(X_i)).$$

Assumptions 3.1-3.2 and 6.1-6.3 imply Assumptions 3.2, 3.3, 4.1, and 4.2 in Imbens and Ridder (2006). Then by Theorem 4.1 in Imbens and Ridder (2006) we have

$$\sqrt{N}\cdot\left(\hat{\theta} - \theta\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega(X_i)\cdot\frac{\partial m}{\partial g'}(g(X))\left(\tilde{Y} - g(X)\right) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega(X_i)\left(m(g(X_i)) - \theta\right) + o_p(1).$$

Thus

$$\sqrt{N}\cdot\left(\hat{\theta} - \theta_S\right) = \sqrt{N}\cdot\left(\hat{\theta} - \theta\right) + \sqrt{N}\cdot(\theta - \theta_S) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega(X_i)\cdot\frac{\partial m}{\partial g'}(g(X))\left(\tilde{Y} - g(X)\right) + o_p(1).$$

Because

$$\frac{\partial m}{\partial g}(g(X)) = \begin{pmatrix} 1/g_3(X) \\ -1/(1-g_3(X)) \\ -g_1(X)/g_3^2(X) - g_2(X)/(1-g_3(X))^2 \end{pmatrix} = \begin{pmatrix} 1/e(X) \\ -1/(1-e(X)) \\ -\mu_1(X)/e(X)^2 - \mu_0(X)/(1-e(X))^2 \end{pmatrix},$$

and

$$\tilde{Y} - g(X) = \begin{pmatrix} Y\cdot W - g_1(X) \\ Y\cdot(1-W) - g_2(X) \\ W - g_3(X) \end{pmatrix} = \begin{pmatrix} Y\cdot W - \mu_1(X)\cdot e(X) \\ Y\cdot(1-W) - \mu_0(X)\cdot(1-e(X)) \\ W - e(X) \end{pmatrix},$$

we have

$$\sqrt{N}\cdot\left(\hat{\theta}-\theta_S\right)=$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda(X_i)\cdot\left(\left(\frac{Y_i\cdot W_i}{e(X_i)}-\mu_1(X_i)\right)-\left(\frac{Y_i\cdot(1-W_i)}{1-e(X_i)}-\mu_0(X_i)\right)-\left(\frac{\mu_1(X_i)}{e(X_i)}+\frac{\mu_0(X_i)}{1-e(X_i)}\right)\cdot(W_i-e(X_i))\right)$$
$$+o_p(1).$$

□

Lemma B.8 (Asymptotic Normality)
Suppose Assumptions 3.1-3.2 and 6.1-6.3 hold. Then

$$\sqrt{N}\cdot(\hat{\tau}(\mathbb{A}^*)-\tau(\mathbb{A}^*))\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{q(\mathbb{A}^*)}\cdot\mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)}+\frac{\sigma_0^2(X)}{1-e(X)}\middle|X\in\mathbb{A}^*\right]\right).$$

Proof: By Lemma B.7, independent sampling, and because the second moment of $\phi(Y,W,X)$ exists, it follows that

$$\sqrt{N}\cdot(\hat{\theta}-\theta_S)\to\mathcal{N}\left(0,\mathbb{E}\left[\phi(Y,W,X)^2\right]\right).$$

In addition, $N_{\mathbb{A}}^*/N\xrightarrow{p}q(\mathbb{A}^*)$, so that

$$\sqrt{N}\cdot(\hat{\tau}(\mathbb{A}^*)-\tau(\mathbb{A}^*))=\sqrt{N}\cdot\frac{N}{N_{\mathbb{A}}^*}\cdot\left(\hat{\theta}-\theta_S\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{q^2(\mathbb{A}^*)}\cdot\mathbb{E}\left[\phi(Y,W,X)^2\right]\right).$$

Because

$$\mathbb{E}\left[\left(\left(\frac{Y\cdot W}{e(X)}-\mu_1(X)\right)-\left(\frac{Y\cdot(1-W)}{1-e(X)}-\mu_0(X)\right)-\left(\frac{\mu_1(X)}{e(X)}+\frac{\mu_0(X)}{1-e(X)}\right)\cdot(W-e(X))\right)^2\middle|X\in\mathbb{A}^*\right]$$

$$=\mathbb{E}\left[\left(\frac{Y\cdot W}{e(X)}-\mu_1(X)\right)^2\middle|X\in\mathbb{A}^*\right]$$
$$-2\cdot\mathbb{E}\left[\left(\frac{Y\cdot W}{e(X)}-\mu_1(X)\right)\cdot\left(\frac{Y\cdot(1-W)}{1-e(X)}-\mu_0(X)\right)\middle|X\in\mathbb{A}^*\right]$$
$$-2\cdot\mathbb{E}\left[\left(\frac{Y\cdot W}{e(X)}-\mu_1(X)\right)\cdot\left(\frac{\mu_1(X)}{e(X)}+\frac{\mu_0(X)}{1-e(X)}\right)\cdot(W-e(X))\middle|X\in\mathbb{A}^*\right]$$
$$+\mathbb{E}\left[\left(\frac{Y\cdot(1-W)}{1-e(X)}-\mu_0(X)\right)^2\middle|X\in\mathbb{A}^*\right]$$
$$+2\cdot\mathbb{E}\left[\left(\frac{Y\cdot(1-W)}{1-e(X)}-\mu_0(X)\right)\cdot\left(\frac{\mu_1(X)}{e(X)}+\frac{\mu_0(X)}{1-e(X)}\right)\cdot(W-e(X))\middle|X\in\mathbb{A}^*\right]$$
$$+\mathbb{E}\left[\left(\left(\frac{\mu_1(X)}{e(X)}+\frac{\mu_0(X)}{1-e(X)}\right)\cdot(W-e(X))\right)^2\middle|X\in\mathbb{A}^*\right]$$

$$=\mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)}+\mu_1^2(X)\cdot\frac{1-e(X)}{e(X)}\middle|X\in\mathbb{A}^*\right]$$
$$+\mathbb{E}\left[2\cdot\mu_0(X)\cdot\mu_1(X)\middle|X\in\mathbb{A}^*\right]$$
$$-\mathbb{E}\left[2\cdot\mu_0(X)\cdot\mu_1(X)+2\cdot\mu_1^2(X)\cdot\frac{1-e(X)}{e(X)}\middle|X\in\mathbb{A}^*\right]$$
$$+\mathbb{E}\left[\frac{\sigma_0^2(X)}{1-e(X)}+\mu_0^2(X)\cdot\frac{e(X)}{1-e(X)}\middle|X\in\mathbb{A}^*\right]$$
$$-\mathbb{E}\left[2\cdot\mu_0^2(X)\cdot\frac{e(X)}{1-e(X)}+2\cdot\mu_0(X)\cdot\mu_1(X)\middle|X\in\mathbb{A}^*\right]$$
$$+\mathbb{E}\left[2\cdot\mu_0(X)\cdot\mu_1(X)+\mu_1^2(X)\cdot\frac{1-e(X)}{e(X)}+\mu_0^2(X)\cdot\frac{e(X)}{1-e(X)}\middle|X\in\mathbb{A}^*\right]$$

$$= \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| X \in \mathbb{A}^* \right],$$

it follows that

$$\mathbb{E}\left[\phi(Y,W,X)^2\right] = q(\mathbb{A}^*)\cdot$$
$$\mathbb{E}\left[\left(\left(\frac{Y \cdot W}{e(X)} - \mu_1(X) \right) - \left(\frac{Y \cdot (1-W)}{1-e(X)} - \mu_0(X) \right) - \left(\frac{\mu_1(X)}{e(X)} + \frac{\mu_0(X)}{1-e(X)} \right) \cdot (W - e(X)) \right)^2 \middle| X \in \mathbb{A}^* \right]$$
$$= q(\mathbb{A}^*) \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| X \in \mathbb{A}^* \right],$$

and the result in the Lemma follows. □

Proof of 6.1: This follows directly from Lemmas B.5 and B.8 □

The proofs of Theorems 6.2-6.6 are omitted here in the interest of space. They are available on the web (Crump, Hotz, Imbens and Mitnik, 2006b).

References

Abadie, A., and G. Imbens, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1): 235-267.

Angrist, J., (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2): 249-288.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A., (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Billingsley, P. (1985), *Probability and Measure,* Wiley Series in Probability and Mathematical Statistics.

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Chen, X., Hong, H., and Tarozzi, A. (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Error," Unpublished manuscript, Duke Univesity.

Cochran, W., and D. Rubin (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya*, Series A,35: 417-446.

Crump, R., V. J. Hotz, G. Imbens and O. Mitnik, (2006a), "Nonparametric Tests for Treatment Effect Heterogeneity', NBER Technical Working Paper No. 324.

Crump, R., V. J. Hotz, G. Imbens and O. Mitnik, (2006b), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand", Supplemental Proofs, http://www.economics.harvard.edu/faculty/imbens/papers/chim_goalpost_supp.pdf.

Dehejia, R., and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94: 1053-1062.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66(2): 315-331.

Ham, J. C., X. Li and P. B. Reagan, (2006), "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men," Unpublished manuscript, USC.

Heckman, J., and V. J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs," (with discussion), *Journal of the American Statistical Association.*, 84(804): 862-874.

Heckman, J., H. Ichimura, and P. Todd, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64(4): 605-654.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65: 261-294.

Heckman, J., H. Ichimura, J. Smith, and P. Todd, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5): 1017-1098.

Heckman, J., R. LaLonde, and J. Smith, (1999), "The economics and econometrics of active labor market programs," in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3A, North-Holland, Amsterdam, 1865-2097.

Hirano, K., and G. Imbens (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2: 259-278.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189.

Ho, D., K. Imai, G. King, and E. Stuart, (2005), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," mimeo, Department of Government, Harvard University.

Ichino, A., F. Mealli, and T. Nannicini, (2005), "Sensitivity of Matching Estimators to Unconfoundedness. An Application to the Effect of Temporary Work on Future Employment," EUI

Imbens, G. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, Papers and Proceedings.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 61(2): 467-476.

Imbens, G., W. Newey and G. Ridder, (2006), "Mean-squared-error Calculations for Average Treatment Effects," unpublished manuscript, Department of Economics, UC Berkeley.

Imbens, G., and G. Ridder, (2006), "Estimation and Inference for Generalized Partial Means," unpublished manuscript, Department of Economics, UC Berkeley.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76: 604-620.

Lechner, M, (2002a), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review of Economics and Statistics*, 84(2): 205-220.

Lechner, M, (2002b), "Some Practical Issues in the Evaluation of Heterogenuous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165: 659–82.

Lee, M.-J., (2005a), *Micro-Econometrics for Policy, Program, and Treatment Effects* Oxford University Press, Oxford.

Lee, M.-J., (2005b), "Treatment Effect and Sensitivity Analysis for Self-selected Treatment and Selectively Observed Response," mimeo, Singapore Management University.

Newey, W., (1994). "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, Vol 10, 233-253.

Robins, J.M., and A. Rotnitzky, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90: 122-129.

Robins, J.M., Rotnitzky, A., Zhao, L-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90: 106-121.

Robins, J.M., S. Mark, and W. Newey, (1992), "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48(2): 479-495.

Robinson, P., (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 67: 645-662.

Rosenbaum, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84: 1024-1032.

Rosenbaum, P., (2001), *Observational Studies*, second edition, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70: 41-55.

Rosenbaum, P., and D. Rubin, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45: 212-218.

Rosenbaum, P., and D. Rubin, (19884), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *JASA*.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66: 688-701.

Rubin, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1): 1-26.

Rubin, D., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6: 34-58.

Shadish, W., T. Cook, and D. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston, MA.

Smith, J., and P. Todd, (2005), "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, 125: 305-353.

Stock, J., (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406): 567-575.

Wooldridge, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press

Table 1: Variance Ratios for Beta Distributions

$\gamma \longrightarrow$		0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$\beta = 0.5$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$	13.38	11.68	13.28	13.71	13.83	13.54	13.24	12.83
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$	1.70	1.64	1.70	1.71	1.70	1.66	1.63	1.58
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$	1.00	1.00	1.00	1.00	1.01	1.01	1.03	1.04
$\beta = 1.$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$		2.68	2.41	2.65	2.97	3.13	3.28	3.36
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$		1.39	1.36	1.39	1.44	1.46	1.47	1.47
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$		1.00	1.00	1.00	1.00	1.00	1.01	1.01
$\beta = 1.5$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$			1.34	1.28	1.34	1.41	1.46	1.51
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$			1.19	1.17	1.19	1.23	1.25	1.26
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$			1.00	1.00	1.00	1.00	1.00	1.00
$\beta = 2.0$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$				1.11	1.09	1.11	1.15	1.16
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$				1.09	1.08	1.09	1.13	1.12
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$				1.00	1.00	1.00	1.00	1.00
$\beta = 2.5$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$					1.04	1.04	1.04	1.06
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$					1.04	1.04	1.04	1.06
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$					1.00	1.00	1.00	1.00
$\beta = 3.0$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$						1.02	1.04	1.02
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$						1.02	1.04	1.02
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$						1.00	1.02	1.00
$\beta = 3.5$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$							1.02	1.02
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$							1.02	1.02
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$							1.00	1.00
$\beta = 4.0$	$\mathbb{V}_S(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$								1.02
	$\mathbb{V}_{S,0.01}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$								1.02
	$\mathbb{V}_{S,0.10}(\gamma,\beta)/\mathbb{V}_{S,\alpha(\gamma,\beta)}(\gamma,\beta)$								1.00

Table 2: Covariate Balance for LaLonde Data

	Mean	Stand. Dev.	Mean Contr.	Mean Treat.	Normalized Dif. in Treat. and Contr. Ave's All	[t-Stat]	$\alpha < e(x) < 1-\alpha$	Optimal Weights	Prop. Score, Weighted
age	34.23	10.50	34.85	25.82	-0.86	[-16.0]	-0.18	-0.08	-0.12
educ	11.99	3.05	12.12	10.35	-0.58	[-11.1]	-0.04	-0.25	-0.35
black	0.29	0.45	0.25	0.84	1.30	[21.0]	0.20	-0.79	-0.70
hispanic	0.03	0.18	0.03	0.06	0.15	[1.5]	0.07	0.27	0.37
married	0.82	0.38	0.87	0.19	-1.76	[-22.8]	-0.81	-0.01	-0.08
unempl '74	0.13	0.34	0.09	0.71	1.85	[18.3]	0.78	-0.23	-0.26
uenmpl '75	0.13	0.34	0.10	0.60	1.46	[13.7]	0.51	-0.18	-0.18
earn '74	18.23	13.72	19.43	2.10	-1.26	[-38.6]	-0.20	0.78	1.19
earn '75	17.85	13.88	19.06	1.53	-1.26	[-48.6]	-0.14	0.47	0.90
Prop. Score	0.07	0.20	0.02	0.68	3.22	[29.9]	1.90	1.86	2.15
Log Odds Ratio	-7.87	4.91	-8.53	1.08	1.96	[53.6]	0.42	0.48	0.56

Table 3: Estimates and Asymptotic Standard Errors for LaLonde Data

	ATE	$\text{ATE}_{0.01}$	$\text{ATE}_{0.10}$	OSATE	OWATE	ATT	$\text{ATT}_{0.01}$	$\text{ATT}_{0.10}$	OSATT
Est.	-14.75	4.97	-0.74	-1.17	-0.19	2.67	2.67	-0.30	-1.43
(s.e.)	637.90	2.09	1.26	1.69	1.29	2.58	2.58	1.82	2.08

Table 4: Subsample Sizes for LaLonde Data

OSATE ($\alpha = 0.066$)	$e(x) < \alpha$	$\alpha \leq e(x) \leq 1 - \alpha$	$1 - \alpha < e(x)$	All
Controls	2302	183	5	2490
Treated	9	129	47	185
All	2311	312	52	2675
$\text{ATE}_{0.10}$	$e(x) < \alpha$	$\alpha \leq e(x) \leq 1 - \alpha$	$1 - \alpha < e(x)$	All
Controls	2354	128	8	2490
Treated	12	98	75	185
All	2366	226	83	2675
$\text{ATE}_{0.01}$	$e(x) < \alpha$	$\alpha \leq e(x) \leq 1 - \alpha$	$1 - \alpha < e(x)$	All
Controls	1999	491	0	2490
Treated	3	182	0	185
All	2002	673	0	2675