ABSTRACT. In weighted moment condition models, we point a subtle link between identification and
estimability that limits the practical usefulness of estimators based on these models. In particular, if it
is necessary for (point) identification that the weights take arbitrarily large values, then the parameter of
interest, though point identified, cannot be estimated at the regular (parametric) rate, and is said to be
irregularly identified. This rate depends on relative tail conditions and can be as slow in some examples
as $n^{-1/4}$. This nonstandard rate of convergence can lead to numerical instability, and/or large standard
errors. We examine two weighted model examples: 1) The binary response model under mean restriction
introduced by Lewbel (1997) and further generalized to cover endogeneity and selection where the estimator
in this class of models is weighted by the density of a special regressor; and, 2) The treatment effect model
under exogenous selection where the resulting estimator of the average treatment effect (or ATE) is one that
is weighted by a variant of the propensity score. Without strong support conditions, these models, similar
to well known “identified at infinity” models, lead to estimators that converge at slower than parametric
rate, since essentially, to ensure point identification, one requires some variables to take values on sets with
arbitrarily small probabilities, or thin sets. For the two models above, we derive rates of convergence under
different tail conditions and also propose rate adaptive inference procedures that are similar to Andrews and

JEL Classification: C14, C25, C13.

Key Words: Irregular and Robust Identification, inverse weighting, rates of convergence.

*We thank a Co-editor and 3 referees for comments that improved the content and exposition in the paper. We
also thank J. Powell and seminar participants at many universities, as well as conference participants at the 2007
NASM at Duke University and 2008 NAWM in New Orleans, for helpful comments.

†Department of Economics, Duke University, 213 Social Sciences Building, Durham, NC 27708. Phone: (919)
660-1873.

‡Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208. Phone: (847)
491-8218. Support from the National Science Foundation is gratefully acknowledged.
1 Introduction

There is a class of models in econometrics, mainly arising in limited dependent variable models, that attain identification by requiring that covariate variables take support in regions with arbitrary small probability mass. These identification strategies sometimes lead to estimators that are weighted by a density or a conditional probability, and these weights take arbitrarily large values on these regions of small mass. For example, an estimator that is weighted by the inverse of the density of a continuous random variable, effectively only uses observations for which that density is small. Taking values on these “thin sets,” is essential (necessary) for point identification in these models. We explore the effect of this weighting on the rates of convergence of resulting estimators, and show that under general conditions, it is often not possible to attain the regular parametric rate (square root of the sample size) with these models. We label these identification strategies as “irregular” in the sense that estimators based on them do not converge at the parametric (square root of the sample size) rate. Consequently we argue that these strategies belong to the class of “identified at infinity” approaches (See Chamberlain (1986) and Heckman (1990a)). Note that it is part of econometrics folklore to equate the parameter being identified at infinity with slow rates of convergence, as was done in Chamberlain (1986) for some particular models. See also Andrews and Schafgans (1998), who show how rates can vary with relative tail behavior in a sample selection model. Furthermore, these rates of convergence depend on (unknown) explicit functions of the relative tail behavior properties of observed and unobserved random variables, making inference more complicated than in standard problems. This paper establishes rate adaptive inference procedures for this class of models and studies properties of the corresponding studentized estimators.

The results in this paper are connected to two models that we examine in detail. First, we consider the binary choice model under an exclusion restriction and mean restrictions. This model was introduced to the literature by Lewbel (1997), Lewbel (1998) and Lewbel (2000). There, Lewbel demonstrates that this binary model is point identified with only a mean restriction on the unobservables, by requiring the presence of a special regressor that is conditionally independent of the error. Under these conditions, Lewbel provides a density weighted estimator for the finite dimensional parameter (including the constant term). This estimator contains a random denominator that can take arbitrarily small values in regions that are necessary for point identification. We show that the parameters in a simple version of that model are irregularly identified unless special relative tail behavior conditions are imposed. Two such conditions are: a support condition such that the propensity score hits limiting values with positive measure, in which case root $n$ estimation can be attained; or infinite moment restrictions on the regressors. In general, the optimal rate of convergence in this model depends on the tail behavior of the special regressor relative to the error distribution.

Second, we consider the important treatment effects model under exogenous selection. This is an important model that is widely used in economics and statistics to estimate the average
treatment effect (ATE), or the treatment on the treated (ATT) parameters in program evaluations. For an exhaustive review of this literature, see Imbens (2004). Hahn (1998), in important work, derived the efficiency bound for this model and provided an estimator that reaches that bound (See also Hirano, Imbens, and Ridder (2003)). In the case where the covariates take relative large support, the propensity score is arbitrarily close to zero or one\(^4\)(as opposed to bounded away from zero or one). We show that the ATE in this case can be irregularly identified (the semiparametric efficiency bound can be infinite) resulting in non-regular rates of convergence. The optimal rates will depend on the relative tail thickness of the error in the treatment equation and the covariates. For empirical users, this results in instability of the estimator in cases where there is “limited overlap” and that care should be taken in making inference on these parameters. Busso, DiNardo, and McCrary (2008), in an important recent paper, highlight this instability in extensive Monte Carlo simulations where the ATE and ATT parameter appear to exhibit bias at moderate sample sizes. See also Frolich (2004).

The next section introduces our class of models. Section 3 considers the binary choice model under mean restrictions, and section 4 considers the average treatment effect. In each of these sections we also derive optimal rates of convergence for estimators of parameters of interest that are sample analogs to the moments conditions used to identify these parameters, and propose rate adaptive inference procedures. Section 5 concludes by summarizing and suggesting areas for future research. An online Appendix collects proofs to all the results in the paper.

2 Irregular Identification and Inverse Weighting

The notion of regularity of statistical models is linked to efficient estimation of parameters- see, e.g. Stein (1956). The concept of irregular identification has been related to consistent estimation of the parameters of interest at the parametric rate. An important early reference to irregular identification in the literature appears in Chamberlain (1986) dealing with efficiency bounds for sample selection models\(^5\). There, Chamberlain showed that even though a slope parameter in the selection equation is identified, it is not possible to estimate it at the parametric root \(n\) rate (under the assumptions he maintains in the paper). Chamberlain added that point identification in this case relies “on the behavior of the regression function at infinity.” In essence, since the disturbance terms are allowed to take support on the real line, one requires that the regression function takes “large” values so that at the limit the propensity score is equal to one (or is arbitrarily close to one). The identified set in the model shrinks to a point when the propensity score hits one (or no selection). Heckman (1990b) highlighted the importance of identified at infinity parameters in

---

\(^4\)The identification strategy for these models is based on writing the estimand as a weighted sum where the weights (in the denominator) can get arbitrary small. This intuition applies equally to a matching strategy since a matching estimator can be written as a weighted estimator where weights can get small.

\(^5\)See also Newey (1990) for other impossibility theorems and notions of regular and irregular estimators.
various selection models and linked this type of identification essentially to the propensity score. Andrews and Schafgans (1998) used a smoothed variation of an estimator proposed by Heckman (1990a) and confirmed slow rates of convergence of the estimator for the constant and also showed that this rate depends on the thickness of the tails of the regression function distribution relative to that of the error term. This rate can be as slow as cube root and can be arbitrarily close to root \( n \).

The class of models we consider in this paper are ones where the parameter of interest can be written as a closed form solution to a weighted moment condition. The weight in these models can take arbitrary small values, and causes point identification to be fragile. This means that the rate of convergence is generally slower than the regular rate. This exact rate is determined by relative tail conditions on observed and unobserved variables. In some models, not only is point identification delicate, but the model will not be able to have any information about the true parameter, i.e., the identified set is the whole parameter space. This leads to non-robust irregular identification where one either learns the parameter precisely, or, the model provides no information about the parameter of interest in the sense that the identified region equates with the parameter space. This is a drawback of inference in this class of models.

We consider models where the parameter of interest can be written as an expectation of some function of observed variables:

\[
\theta_0 = E[g(y_i, x_i)]
\] (2.1)

where \( \theta_0 \) is a finite dimensional parameter of interest, and \((y_i, x_i)\) is finite dimensional vector of observed random variables. What we are particularly interested in this paper is the class of models above in which

\[
E[\|g(y_i, x_i)\|^2] = \infty
\] (2.2)

where \( \| \cdot \| \) denotes the Euclidean norm. The models arise in various settings, especially in cases where \( g \) is a ratio and where the denominator takes arbitrarily small values in a region of its support. Typical examples are weighted estimators and estimator with “random denominators.” An important recent example that fits into this framework is Graham, and Powell (2009).

We distinguish between two classes of irregularly identified models, robust and non-robust. We say that a parameter vector is non-robustly irregularly point identified if it is irregularly point identified yet the function \( g(\cdot) \) in (2.1) is unbounded on the support of its arguments. This relates to the classical definitions of robustness (see, e.g. Hampel (1986)) where we focus attention to irregularly identified models. We show that estimators based on (2.1) generally converge at a rate slower than the standard parametric rate, and whose optimal rate depends on relative tail conditions on the random variable \( x_i \) and the unobserved components also governing the behavior of \( y_i \). This general setting will encompass many models in econometrics, but for illustrative purposes,
our discussion will focus on the binary response model under mean restrictions, and the average treatment effect estimator under conditional independence.

2.1 Preview of Approach

The key problem with an empirical analog estimator of $\hat{\theta}$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i)$$  \hspace{1cm} (2.3)

is that although consistent (assume that $\theta_0 = 0$), has the unattractive property that $E[\|g(y_i, x_i)\|^2] = \infty$ which does not allow us to use a standard CLT to say that $\hat{\theta}$ is root-$n$ consistent and asymptotically normal. One approach is to introduce a trimming sequence $\tau_{ni} = \tau(y_i, x_i, n) > 0, \tau_{ni} \xrightarrow{a.s.} 1$ such that

$$E[\tau_{ni}^2\|g(y_i, x_i)\|^2] \equiv \gamma_n < \infty$$  \hspace{1cm} (2.4)

But note that since $\tau_{ni}$ is converging to 1, $\gamma_n$ will generally diverge to infinity as the sample size increases. Thus we can explore the asymptotic properties of the trimmed estimator:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \tau_{ni} g(y_i, x_i)$$  \hspace{1cm} (2.5)

This trimming, which disregards observations in the sample where the value of the “denominator” is small, will introduce bias which has to be taken into account. So, ultimately for the binary response model and the ATE model we consider, we address the following questions.

1. What is the optimal rate of convergence for $\hat{\theta}_n$?

2. What is the limiting distribution for $\hat{\theta}_n$?

3. How does one conduct inference for $\theta_0$ given the rate of convergence of its estimator will generally be unknown?

Addressing 1) is analogous to deriving optimal rates for nonparametric estimation. That is, the variance and bias are inversely related in the sense that here $\tau_{ni}$ controls the variance, but at the expense of inducing bias. Therefore, as in nonparametric estimation, we can solve for an “optimal” sequence $\tau_{ni}^*$ that balances the bias and the variance. To address 2), we can apply the Lindeberg theorem to the centered trimmed estimator that should hold under a version of a Lindeberg condition. Finally, to address the third question, our proposed approach to conduct inference based on $\hat{\theta}_n$ will be to studentize the estimator. As we will show, under our conditions the $t$-statistic will be rate adaptive in the sense that it will converge in probability to a normal distribution regardless of the rate of convergence of $\hat{\theta}$.

We next apply this framework to specific models, beginning with the binary choice model. Next, we examine the average treatment effect model under conditional independence.
3 Binary Choice with Mean Restrictions

In the standard binary response model

\[ y = 1[x'\beta + \epsilon \geq 0] \]

Manski (1988) showed that a conditional mean restriction does not identify the parameter \( \beta \). In fact, he showed that the mean independence assumption is not strong enough in binary response models to even bound \( \beta \). Hence, we modify this model by adding more assumptions to ensure point identification. To simplify the discussion here, consider the special case where \( x \) consists only of a constant and a single regressor \( v \) with a coefficient that is normalized to equal one, as introduced by Lewbel (1997):

\[ y_i = 1[\alpha + v_i - \epsilon_i \geq 0] \quad (3.1) \]

where \( v_i \) is a scalar random variable that is independent of \( \epsilon_i \) and both \( \epsilon_i \) and \( v_i \) have support on the real line and \( E[\epsilon_i] = 0 \). We observe both \( y \) and \( v \) and the object of interest is the parameter \( \alpha \). The location restriction on \( \epsilon \) point identifies \( \alpha \). We start with the following:

\[ P(y_i = 1|v_i) = F_\epsilon(\alpha + v_i) \quad (3.2) \]

where \( F_\epsilon(.) \) is the cdf of \( \epsilon \) and we assume that this cdf is a strictly increasing function. Lewbel (1997) derived the following relation

\[ \alpha = E[(y_i - I[v_i > 0])f(v_i)^{-1}] \equiv E[(y_i - I[v_i > 0])w(v_i)] \quad (3.3) \]

with the weight function \( w(v_i) = f(v_i)^{-1} \) where \( f(.) \) here denotes the density of \( v_i \). This type of identification is sensitive to conditions imposed on the support of \( v \) and \( \epsilon \). For example, point identification of \( \alpha \) is lost when one excludes sets of \( v \) of arbitrary small probability when \( v \) is in the tails. The model in this case will not contain any information about \( \alpha \) (in the sense of trivial bounds). This is a case of non-robust or thin-set identification. To see this, note that if we restrict the support of \( v \) to lie on the set \([-K,K]\) for any \( K > 0 \), \( \alpha \) will not be point identified:

\[ \alpha = \left[ -\int_{-\infty}^{-K}vf_\epsilon(\alpha + v)dv - \int_{-K}^{K}vp'(v)dv - \int_{K}^{\infty}vf_\epsilon(\alpha + v)dv \right] \quad \text{(1)} \]

\[ \quad \quad \quad \quad \text{where } p'(\cdot) \text{ denotes the derivative of the probability function } P(y_i = 1|v_i). \]

We see that only (2) is identified, while (1) and (3) are not since, in (3.2), we can only learn \( F_\epsilon \) from \( \alpha - K \) to \( \alpha + K \). So, no matter how large \( K \) is, the model is set identified. In addition, we see that it is possible to choose the unidentified portion of \( f_\epsilon(.) \) (parts (1) and (3) above) in such a way that the model provides no information about \( \alpha \). This is similar to results in Magnac and Maurin (2005) (we view this property...
as similar to the nonrobustness of the sample mean). To guarantee point identification when $\epsilon$ takes support on the real line $([-\infty, +\infty])$, $v$ is also required to take arbitrary large values. This will mean that the density of $v_i$ vanishes on a set of arbitrarily small measure, in this case $|v| > M$, for an arbitrarily large constant $M$. Here, the weight function in (3.3) becomes unbounded. This suggests identification of $\alpha$ based on (3.3) is irregular in the sense that an analog estimator will not generally converge at the parametric rate. We confirm this later by showing how optimal rates of the analog estimator can vary widely with tail behavior conditions on observed and unobserved variables. This nonuniform behavior of the analog estimator makes conducting inference difficult, and motivates our proposed rate adaptive inference procedure which we explain in the following section.

### 3.1 Rate Adaptive Inference in the Binary Choice Model

We propose the use of rate-adaptive inference procedures to estimate the binary response model above. This is similar to what Andrews and Schafgans (1998) did for the selection model. Let $\hat{\alpha}_n$ denote the trimmed variant of the estimator proposed in Lewbel (1997) for the intercept term:

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{\hat{f}(v_i)} I[|v_i| \leq \gamma_{2n}]$$

(3.4)

where $\hat{f}(v_i)$ is a kernel estimator for the density of $v$ (details of the kernel estimation procedure can be found in the Appendix). The estimator includes the additional trimming term $I[|v_i| \leq \gamma_{2n}]$ where $\gamma_{2n}$ is a deterministic sequence of numbers that satisfies $\lim_{n \to \infty} \gamma_{2n} = \infty$ and $\lim_{n \to \infty} \gamma_{2n}/n = 0$. Effectively, this extra term helps govern tail behavior. Trimming this way suffices (under the assumptions in Lewbel (1997)) to deal with the denominator problem associated with the density function $f(v_i)$ getting small, which, under stated conditions in Lewbel (1997)), only happens in the tails of $v_i$. Let $\hat{S}_n$ denote a trimmed estimator of its asymptotic variance if conditions were such that the asymptotic variance were finite

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}(v_i)(1 - \hat{p}(v_i))}{\hat{f}(v_i)^2} I[|v_i| \leq \gamma_{2n}]$$

(3.5)

where $\hat{p}()$ denotes a kernel estimator of the choice probability function (or the propensity score). Our main result in this section is that the “studentized” estimator converges to a standard normal

---

6This is a property shared by the estimator in Heckman (1990b) and Andrews and Schafgans (1998) for an estimator of the intercept term in a sample selection model. As is the case in that setting, the slower rate attained does not necessarily imply inefficiency of the analog estimator. In fact the appendix derives a result analogous to an impossibility theorem in Chamberlain (1986)

7In cases where $\epsilon$ has bounded support, it is possible that $\alpha$ is identified regularly, if the support of $v$ is large relative to that of $\epsilon$. One can check that easily since for those large values of $v$, the choice probability hits 0 or 1.

8We also derive the efficiency bounds for this model in the Appendix.
distribution. We consider this result as rate-adaptive in the sense that the limiting distribution holds regardless of the rate of convergence of the un-studentized estimator. We state this formally as a theorem, which first requires the following definitions. Define

\[ v(\gamma_2n) = E[p(v_i)(1 - p(v_i))]/f(v_i) I[v_i \leq \gamma_2n] \]  

(3.6)

\[ b(\gamma_2n) = \alpha_0 - E[y_i - I[v_i > 0]/f(v_i) I[v_i \leq \gamma_2n] \]  

(3.7)

and

\[ X_{ni} = y_i - p(v_i) I[v_i \leq \gamma_2n] \]  

(3.8)

The first piece \( v(\gamma_2n) \) is the variance of the infeasible estimator where in \( (\gamma_3) \), the true \( f(.) \) is used. This term is of the same order as the variance of the feasible estimator (See the Appendix). Similarly, the term in \( (3.7) \) is the bias term for the infeasible estimator. The form of \( X_{ni} \) is the linear representation of \( \hat{\alpha}_n \), which involves first linearizing the estimator around \( \bar{f} \) and then taking the usual U-statistic projections. See section A.4 of the appendix for details on this.

**Theorem 3.1** Suppose (i) \( \gamma_2n \rightarrow \infty \), (ii) \( \forall \epsilon > 0 \),

\[
\lim_{n \rightarrow \infty} \frac{1}{v(\gamma_2n)} E[X_{ni}^2 I[X_{ni}] > \epsilon \sqrt{nv(\gamma_2n)}] = 0
\]

(3.9)

and (iii) \( \sqrt{nv(\gamma_2n)} b(\gamma_2n) \rightarrow 0 \), then

\[
\sqrt{n} S_n^{-1/2} (\hat{\alpha} - \alpha_0) \Rightarrow N(0, 1)
\]

(3.10)

The proof is left to the appendix and here we only comment on the conditions. Condition (ii) is a Lindberg that allows us to apply a central limit theorem. Condition (iii) on the other hand is needed to ensure that the bias will vanish with sample size. It is possible to provide more interpretable sufficient conditions for (ii) and (iii). For example, for (ii), it is sufficient that \( (\gamma_2n f(\gamma_2n))^{1+\delta} v(\gamma_2n)^{\delta/2} f(\gamma_2n) \) converges to zero for some \( \delta > 0 \) (See section A.6.1 in the appendix for more details). This happens if 1) \( \epsilon \) has thinner tails than \( v \), and 2) \( \gamma_2n f(\gamma_2n) \) converges to zero as \( \gamma_2n \) increases, which will happen with exponentially decaying densities such as logistic, normal, laplace, and will also happen with \( t \) distributions. Note if \( \epsilon_i \) has thinner tails than \( v_i \) the term in the square root converges to 0, so we will focus on \( \gamma_2n f(\gamma_2n) n^{1/2} \). Suppose the density of \( \epsilon_i \) has exponential tails: \( f(\gamma_2n) = \exp(-g(\gamma_2n)) \) where \( g(\gamma_2n) \) increases to infinity at a polynomial rate, i.e. \( g(\gamma_2n) = \gamma_n^a \) where \( a > 0 \). Then, in cases where \( \gamma_2n \) converges to infinity very slowly, i.e. a logarithmic rate (which recall was often close to the "optimal rate" in many of our examples), the
term $\gamma_{2n}f_\epsilon(\gamma_{2n})n^{1/2}$ will converge to 0. So we see that both exponential tails of $\epsilon_i$ and relative thin tails of $\epsilon_i$ to $v_i$, will jointly suffice for (iii) to hold. The exponential tail is sufficient in the sense that (iii) will be satisfied with some members of the $t$ family of distributions. Of course (iii) will not be satisfied when the tails of $v_i$ are too thin relative to $\epsilon_i$, for example when $v_i$ has exponential tails and $\epsilon_i$ is in the $t$ family. But we note in such cases, as with nonparametric density estimation, an alternative to undersmoothing would be to estimate $b(\gamma_{2n})$ and show the centered estimator is asymptotically normal with 0 mean. Other conditions on the kernel and the other smoothing parameters are given in section A.3 in the appendix.

The overall result in the above theorem is analogous to the results in Andrews and Schafgans (1998) who considered asymptotic properties of the identification at infinity estimator proposed in Heckman (1990a). Note that the rate of convergence of this estimator depends on the behavior of $\hat{S}_n$. For example, in cases where $\hat{S}_n$ converges in probability to a non-random matrix, then the estimator above will converge at the regular root $n$ rate. However, in some interesting cases that we go over below, the variance of the estimator goes to infinity which results in slower than regular rates. Finally, for a model with regressors $x$, the estimator for the vector of parameters (which include both slope and intercept) can be obtained in closed form (See Lewbel (1997)):

$$
\hat{\beta} = \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right\}^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i \frac{y_i - 1[v_i \geq 0]}{f(v_i|x_i)^2} I[|v_i| \leq \gamma_{2n}]
$$

and so, the studentized estimator is similar to the one above.

3.2 Relative Tail Behavior and Rates of Convergence in Special Cases

In this section, we derive the rate of convergence for the estimator of $\alpha$ in some examples. This will shed light on the rates of convergence in simple and generic cases. This rate of convergence depends on the relative tail behavior of the density of $v$ vs the “tails” of the propensity score, or the rate at which the propensity score approaches 0 and 1. This rate for the estimator of the intercept term in (3.1) is slower than parametric rate. The result holds for any model where the tail of the special regressor $v$ is as thin or thinner than the tail of the error term. We present examples also where the rate of convergence reaches the regular parametric rate. For example, as we show below, when $v_i$ is Cauchy, then the estimator of $\alpha$ converges at root $n$ rate\footnote{The notion of attaining a faster rate of convergence when moments are infinite is not new. For example, it is well known that when regressors have a cauchy distribution in the basic linear model, OLS is super-consistent.} Here, we focus on the rate of convergence for the infeasible estimator. This rate of convergence is of the same order as the one of the feasible estimator (See the Appendix for more on this).
Consider the infeasible estimator of $\alpha$ that was proposed by Lewbel (1997):

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n]$$ (3.11)

Next let $\bar{\alpha}_n = E[\hat{\alpha}]$. In what follows we will establish a rate of convergence and limiting distribution for $\hat{\alpha} - \bar{\alpha}_n$. To do so we first define the sequence of constants $v(\gamma_2n) = \text{Var}(\frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n])$, and let $h_n = v(\gamma_2n)^{-1}$.

To derive the distribution of the above, we will apply the Lindeberg condition to the following triangular array:

$$\sqrt{n}h_n(\hat{\alpha} - \bar{\alpha}_n) = \sum_{i=1}^{n} \sqrt{\frac{h_n}{n}} \left( \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n] - \bar{\alpha}_n \right)$$ (3.12)

Before claiming asymptotic normality of the above term we verify that the Lindeberg condition for the array is indeed satisfied. Specifically, we need to establish the limit of

$$h_n E\left[ \left( \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n] - \bar{\alpha}_n \right)^2 I\left[ \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \bar{\alpha}_n \right)^2 > \frac{n}{h_n} \epsilon^2 \right] \right]$$ (3.13)

for $\epsilon > 0$. The above limit is indeed zero since $h_n E\left[ \left( \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n] - \bar{\alpha}_n \right)^2 \right] = 1$ and $\frac{n}{h_n} \to \infty$. Therefore, we can conclude that

$$\sqrt{n}h_n(\hat{\alpha} - \bar{\alpha}_n) \Rightarrow N(0,1)$$ (3.14)

So we have established that the rate of convergence of the centered estimator is governed by the limiting behavior of $h_n$. As we will show below, $h_n \to 0$ in some generic examples, resulting in a slower rate of convergence, as is to be expected given our efficiency bound calculations in [A.1].

The above calculations only derives rates for the centered estimator, and not the estimator itself. To complete the distribution theory we derive the rate of convergence of the bias term:

$$\sqrt{n}h_n(\bar{\alpha}_n - \alpha_0) \quad \text{and} \quad \alpha_0 = E[\frac{y_i - I[v_i > 0]}{f(v_i)}]$$

We have:

$$b_n \equiv \bar{\alpha}_n - \alpha_0 = - \int_{\gamma_2n}^{\infty} (p(v) - 1)dv - \int_{-\infty}^{-\gamma_2n} p(v)dv$$ (3.15)

where $p(v) = P(y_i = 1|v_i = v)$ is the “propensity score”. Clearly we have $\lim_{n \to \infty} b_n = 0$ if we maintain that $\lim_{n \to \infty} \gamma_2n = \infty$.

There we can see that the limiting distribution of the estimator can be characterized by two components, the variance, which we see depends on the limiting ratio of the propensity score divided
by the regressor density, and the bias, which depends on the rate at which the propensity score converges to 1. Our results are somewhat abstract as stated since we have not yet stated conditions on \( \gamma_{2n} \) except that it increases to infinity. Nor have we stated what the sequences \( h_n \) and \( b_n \) look like as functions of \( \gamma_{2n} \). We show now how they relate to tail behavior assumptions on the regressor distribution and latent error term.

First we calculate the form of the variance term \( v(\gamma_{2n}) \) as a function of \( \gamma_{2n} \). Note we can express \( v(\gamma_{2n}) \) as the following integral:

\[
v(\gamma_{2n}) = \int_{-\gamma_{2n}}^{\gamma_{2n}} \frac{p(v)(1 - p(v))}{f(v)} dv + \text{Var}((E[(\frac{y_i - I[v_i > 0]}{f(v_i)}I[|v_i| \leq \gamma_{2n}])[v_i] - \bar{\alpha}_n])) \tag{3.16}
\]

We will focus on the first term because we will now show the second term is negligible when compared to the first. The second term in the variance is of the form:

\[
\text{Var}(E[(\frac{y_i - I[v_i > 0]}{f(v_i)}I[|v_i| \leq \gamma_{2n}])[v_i] - \bar{\alpha}_n]) \tag{3.17}
\]

The above variance is of the form:

\[
E[(\bar{\alpha}_n(v_i) - \bar{\alpha}_n)^2] \tag{3.18}
\]

where \( \bar{\alpha}_n(v_i) \) denotes the conditional expectation in (3.17). Note that since \( \bar{\alpha}_n \) converges to \( \alpha_0 \) and \( E[\bar{\alpha}_n(v_i)] = \bar{\alpha}_n \) the only term in the above expression that may diverge is

\[
E[\bar{\alpha}_n(v_i)^2] = \int_{-\gamma_{2n}}^{\gamma_{2n}} \frac{(p(v) - I[v_i > 0])^2}{f(v)} dv = \int_{0}^{\gamma_{2n}} \frac{(1 - p(v))^2}{f(v)} dv + \int_{-\gamma_{2n}}^{0} \frac{(p(v))^2}{f(v)} dv \tag{3.19}
\]

which is of equal or smaller order than \( \int_{-\gamma_{2n}}^{\gamma_{2n}} \frac{p(v)(1 - p(v))}{f(v)} dv \), the first piece of the variance term in (3.16). So, as far as deriving the optimal rate of convergence for the estimator, we can ignore the second term in (3.16) and focus on the first term.

As for the bias term \( b_n \), we see that

\[
b_n = E[\bar{\alpha}_n] - \alpha = \int_{-\gamma_{2n}}^{\gamma_{2n}} \{F_{\epsilon}(v + \alpha) - 1[v \geq 0]\} dv - \alpha \tag{3.20}
\]

where \( F_{\epsilon}(\cdot) \) denotes the cdf of \( \epsilon_i \). This bias term behaves asymptotically like:

\[
b_n \approx \gamma_{2n}(1 - p(\gamma_{2n})) + \gamma_{2n}p(-\gamma_{2n}) \tag{3.21}
\]

Note that the bias term does not directly depend on the density of \( v_i \). With these general results, we now calculate the rate of convergence for some special cases corresponding to various tail behavior conditions on the regressors and the error terms.

---

\(^{10}\)This can be shown most easily by applying Hopital’s rule to the ratio of \( \int_{-\gamma_{2n}}^{\gamma_{2n}} \frac{(1 - p(v))^2}{f(v)} dv / \int_{-\gamma_{2n}}^{\gamma_{2n}} \frac{p(v)(1 - p(v))}{f(v)} dv \) when the denominator diverges. When the denominator converges to a constant, it can also be shown that the numerator converges to a constant.
- **[Logit Errors/Logit Regressors]** Here we assume the latent error term and the regressors both have a standard logistic distribution. We consider this to be the main example, as the results that arise here (a slower than root-$n$ rate) will generally also arise anytime we have distributions whose tails decline exponentially, such as the normal, logistic and Laplace distributions. From our calculations in the previous section we can see that first

\[ v(\gamma_{2n}) = 2\gamma_{2n} \quad \text{and} \quad b_n = \ln \frac{1 + e^{-\gamma_{2n} - \alpha}}{1 + e^{-\gamma_{2n} + \alpha}} \]  

(3.22)

and by our approximation

\[ b_n \approx -2 \exp(-\gamma_{2n}) \]  

(3.23)

So to minimize mean squared error we set \( \frac{\gamma_{2n}}{n} = \gamma_{2n}^2 \exp(-2\gamma_{2n}) \) to get \( \gamma_{2n} = O(\log n^{1/2}) \), resulting in the rate of convergence of the estimator:

\[ \sqrt{\frac{n}{\log n^{1/2}}} (\hat{\alpha} - \alpha_0) = O_p(1) \]  

(3.24)

Furthermore, from the results in the previous section, the estimator will be asymptotically normally distributed, and have a limiting bias.

- **[Normal Errors/Normal Regressors]** Here we assume that the latent error term and the regressors both have a standard normal distribution. To calculate \( v(\gamma_{2n}) \) in this case, we can use the asymptotic properties of the Mills Ratio (see Gordon (1941)):

\[ r(t) \sim \frac{1}{t} \quad \text{as} \quad t \to \infty \]

to approximate \( v(\gamma_{2n}) \approx \log(\gamma_{2n}) \)\(^{11}\) For this case we have \( b_n \approx (1 - \Phi(\gamma_{2n})) / \gamma_{2n} \). So to minimize MSE, we set \( (1 - \Phi(\gamma_{2n})) / \gamma_{2n} = \sqrt{\log \gamma_{2n}/n} \) and we get \( \gamma_{2n} \approx O(\sqrt{\log n}) \) so the estimator is \( O_p(\sqrt{\log(\log n)}/n) \).

- **[Logit Errors/Normal Regressors]** Here we assume the latent error term has a logistic distribution but the regressors have a normal distribution. We expect the rate of convergence

\(^{11}\)This approximation is based on the case where \( \alpha \neq 0 \); when \( \alpha = 0 \) the estimator is unbiased, and converges at a rate arbitrarily close to root \( n \). Actually, the approximation in (3.21) is not sharp in this case, as the limit of the actual bias over the approximate bias converges to 0. Hence for this example we worked with the approximation \( b_n \approx p(\gamma_{2n}) - 1 - p(-\gamma_{2n}) \)

\(^{12}\)The way these approximations are derived here and in all the examples is: 1) we guess at an approximating function (say in this case \(-2 \exp(-\gamma_{2n})\) for the bias in this example), and then 2) we take the ratio of \( b_n \) to its proposed approximation (here \( b_n \) in (3.22) divided by its approximating function in (3.23)) and show via L’Hospital’s rule that the limit is a nonzero constant that is finite.

\(^{13}\)Again, this can be shown by applying L’Hospital’s rule to the ratio \( v(\gamma_{2n}) / \log(\gamma_{2n}) \) and showing that the limit is a nonzero constant.
to be slow since the regressor has much thinner tails than the error. The variance is of the form:

\[ v(\gamma_n^2) = \int_{-\infty}^{\gamma_n^2} \exp(v^2/2)dv \]  

(3.25)

which can be approximated as: \( v(\gamma_n^2) = O(\exp(\gamma_n^2/2)\gamma_n^{-1}) \). The bias is of the form:

\[ b_n \approx \exp(-\gamma_n^2) \]  

(3.26)

So the MSE minimizing sequence is of the form: \( \gamma_n = O(\sqrt{\log n}) \) Resulting in the rate of convergence

\[ O_p\left( \frac{n^{-1/4}}{\sqrt{\log n}} \right) \]

- **[Logit Errors/Cauchy Regressors]** Here we assume the latent error term has a standard logistic distribution but the regressor has a standard cauchy distribution. From our calculations in the previous section we can see that

\[ v(\gamma_n^2) \approx \gamma_n^2 \exp(-\gamma_n^2) (1 + \gamma_n^2) \]  

(3.27)

and

\[ b_n = \gamma_n^2 \exp(-\gamma_n^2) (1 + \gamma_n^2) \]  

(3.28)

We note in this situation \( v(\gamma_n^2) \) remains bounded as \( \gamma_n \rightarrow \infty \), so the variance of the estimator is \( O(\frac{1}{n}) \). Therefore we can let \( \gamma_n \) increase to infinity as quickly as desired to ensure \( b_n = o_p(n^{-1/2}) \). Therefore in this case we can conclude that the estimator is root-n consistent and asymptotically normal with no asymptotic bias.

- **[Probit Errors/Cauchy Regressors]** Here we assume the latent error term has a standard normal distribution but the regressor has a standard cauchy distribution. From our calculations in the previous section we can see that

\[ v(\gamma_n^2) \approx \gamma_n^2 \exp(-\gamma_n^2/2)\gamma_n^{-1} (1 + \gamma_n^2) \]  

(3.29)

where the above approximation is based on the asymptotic series expansion of the erf. Furthermore

\[ b_n = \gamma_n^2 \exp(-\gamma_n^2/2)\gamma_n^{-1} \]  

(3.30)

We can see immediately that no matter how quickly \( \gamma_n \rightarrow \infty \), \( v(\gamma_n^2) = O(1) \), so we can set \( \gamma_n = O(n^2) \) to that

\[ \sqrt{n}(\hat{\alpha} - \alpha_0) = O_p(1) \]  

(3.31)

and furthermore the estimator will be asymptotically normal and asymptotically unbiased.
- **Differing Support Conditions: Regular Rate** Let the support of the latent error term be strictly smaller than the support of the regression function. For $\gamma_{2n}$ sufficiently large $p(\gamma_{2n})$ takes the value 1, so in this case we need not use the above approximation and $\lim_{\gamma_{2n} \to \infty} v(\gamma_{2n})$ is finite. This implies we can let $\gamma_{2n}$ increase as quickly as possible to remove the bias, so the estimator is root-$n$ consistent and asymptotically normal with no limiting bias. This is a case where for example $\alpha + v$ in (3.1) has strictly larger support than $\epsilon$’s.

As the results in this section illustrate, the rates of convergence for the analog inverse weight estimators vary with tail behavior conditions on both observed and unobserved random variables, and the optimal rate rarely coincides with the parametric rate. Hence, semiparametric efficiency bounds might not be as useful for this class of models. Overall, this confirms the results in Lewbel (1997) about the relationship between the thickness of the tails and the rates of convergence. So to summarize our results, we say that the optimal rate is directly tied to the relative tail behavior of the special regressor and the error term. Rates will generally be slower than the parametric rate, which can only be attained with infinite moments on the regressor, or strong support conditions. The latter ensures the propensity scores attains its extreme values with positive probability.

### 4 Treatment Effects Model under Exogenous Selection

This section studies another example of a parameter that is written as a weighted moment condition and where regular rates of convergence requires that support conditions be met, essentially guaranteeing that the denominator be bounded away from zero. We show that the Average Treatment Effect estimator under exogenous selection cannot generally be estimated at the regular rate unless the propensity score is bounded away from 0 and 1, and so, estimation of this parameter can lead to unstable estimates if these support (or overlap) conditions are not met.

A central problem in evaluation studies is that potential outcomes that program participants would have received in the absence of the program is not observed. Letting $d_i$ denote a binary variable taking the value 1 if treatment was given to agent $i$, and 0 otherwise, and letting $y_{0i}, y_{1i}$ denote potential outcome variables, we refer to $y_{1i} - y_{0i}$ as the treatment effect for the $i$’th individual. A parameter of interest for identification and estimation is the average treatment effect, defined as:

$$\beta = E[y_{1i} - y_{0i}]$$

(4.1)

One identification strategy for $\beta$ was proposed in Rosenbaum and Rubin (1983), under the following assumption:

**Assumption 1 (ATE under Conditional Independence)** Let the following hold:

(i) There exists an observed variable $x_i$ s.t.

$$d_i \perp (y_{0i}, y_{1i}) | x_i$$
(ii) \(0 < P(d_i = 1|x_i) < 1\ \forall x_i\)

See also Hirano, Imbens, and Ridder (2003). The above assumption can be used to identify \(\beta\) as

\[
\beta = E[E[y_i|d_i = 1, x_i] - E[y_i|d_i = 0, x_i]]
\]

or

\[
\beta = E_X[E[y_i|d_i = 1, p(x_i)] - E[y_i|d_i = 0, p(x_i)]]
\]

where \(p(x_i) = P(d_i = 1|x_i)\). The above parameter can be written as:

\[
\beta = E[y_i(d_i - p(x_i)) p(x_i)(1 - p(x_i))]
\]

This parameter is a weighted moment condition where the denominator gets small if the propensity score approaches 0 or 1. Also, identification is lost when we remove any region in the support of \(x_i\) (so, fixed trimming will not identify \(\beta\) above). Without any further restrictions, there can be regions in the support of \(x\) where \(p(x)\) becomes arbitrarily large, suggesting an analog estimator may not converge at the parametric rate. Whether or not it can is a delicate question, and depends on the support conditions and on the propensity score.

Hahn (1998), in very important work, derived efficiency bounds for the model based on Assumption 1. These bounds can be infinite and in fact, there will be many situations where this can happen. For example, when the treatment equation corresponds to a propensity score function of \(p(x) = F(x)\) where \(F()\) denotes the cdf of the error term in the treatment equation, and there is one continuous regressor \(x_i\) where the distribution of the \(x_i\) is the same as the treatment equation error term so (ii) is satisfied, the variance bound in Hahn (1998) is infinity. Thus we see that, such as was the case in the binary choice model, rates of convergence will depend on relative tail behaviors of regressors and error terms. Consequently, inference can be more complicated, and one might want to supplement standard efficiency bound calculations. For inference, we again propose a rate adaptive procedure based on a studentized estimator.

### 4.1 Rate Adaptive Inference in the Treatment Effects Model

We derive the asymptotic distribution of the feasible ATE estimator when the propensity score \(p(x)\) is replaced by a nonparametric estimate \(\hat{p}(x)\) which is kernel based. The feasible estimator is

\[
\hat{\beta}_n = \frac{1}{n} \sum_{i=1}^n \left( \frac{d_i y_i}{\hat{p}(x_i)} - \frac{(1 - d_i) y_i}{1 - \hat{p}(x_i)} \right) I[|x_i| \leq \gamma 2^n]
\]

Moreover, let
This represents the first term in the linear representation (see the appendix for more on this). As in the previous section, let \( \hat{S}_n \) denote the trimmed estimator of its asymptotic variance, which is the variance of the above linear term and can be written as (assuming a homoskedastic model with unit variances)

\[
\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( \hat{E}[y_i|d_i = 1, x_i] - \hat{E}[y_i|d_i = 0, x_i] - \hat{\alpha}_n \right)^2 + \frac{1}{\hat{p}(x_i)(1 - \hat{p}(x_i))} \right] 1[|x_i| \leq \gamma_{2n}]
\]

Again, we provide below the main result which is the asymptotic distribution of the studentized estimator. First, we make the following definitions:

\[
\bar{\beta}_n = E[\hat{\beta}_n]; \quad v(\gamma_{2n}) = E\left[ \frac{1}{\hat{p}(x_i)(1 - \hat{p}(x_i))} I[|X_{ni}| \leq \gamma_{2n}] \right]; \quad b_n = \bar{\beta}_n - \beta_0
\]

where here

\[
\beta_0 = E\left( \frac{d_i y_i}{p(x_i)} - \frac{(1 - d_i) y_i}{1 - p(x_i)} \right) \tag{4.6}
\]

We next state the main theorem of this section.

**Theorem 4.1** Suppose (i) \( \gamma_{2n} \to \infty \), (ii) \( \forall \epsilon > 0 \),

\[
\lim_{n \to \infty} \frac{1}{v(\gamma_{2n})} E[X_{ni}^2 I[|X_{ni}| > \epsilon \sqrt{nv(\gamma_{2n})}]] = 0 \tag{4.7}
\]

and (iii) \( \sqrt{nv(\gamma_{2n})} b(\gamma_{2n}) \to 0 \), then

\[
\sqrt{n} \hat{S}^{-1/2} (\hat{\beta}_n - \beta_0) \Rightarrow N(0, 1) \tag{4.8}
\]

### 4.2 Average Treatment Effect Estimation Rates

Here we conduct the same rate of convergence exercises for estimating the average treatment effect as the ones we computed for the binary response model in the previous section. As a reminder, we do not compute what the variance or the bias are exactly, rather, we are interested in the rate at which these need to behave as sample size increases. We will explore the asymptotic properties of the following estimator:

\[
\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{d_i y_i}{p(x_i)} - \frac{(1 - d_i) y_i}{1 - p(x_i)} \right) I[|x_i| \leq \gamma_{2n}] \tag{4.9}
\]

We note that this estimator is infeasible, this time because we are assuming the propensity score is known. Like in the binary choice model this will not effect our main conclusions (of course its
asymptotic variance will be different than the feasible estimator, but will have the same order). Also, like in the binary choice setting, we assume that it suffices to trim on regressor values, and that the denominator only vanishes in the tails of the regressor. Furthermore, to clarify our arguments we will assume here that the counterfactual outcomes are homoskedastic with a variance of 1. Define $h_n = v(\gamma_2 n)^{-1}$.

Carrying the same arguments as before we explore the asymptotic properties of $v(\gamma_2 n)$ and $b_n$ as $\gamma_2 n \to \infty$. Generally speaking, anytime $v(\gamma_2 n) \to \infty$, the fastest rate for the estimator will be slower than root-$n$.

We can express $v(\gamma_2 n)$ as the following integral:

$$v(\gamma_2 n) = \int_{-\gamma_2 n}^{\gamma_2 n} \frac{f(x)}{p(x)(1-p(x))} dx$$  \hspace{1cm} (4.10)

For the problem at hand we can write $b_n$ approximate as the following integral:

$$b_n \approx \gamma_2 n \int_{-\gamma_2 n}^{\infty} m(x)f(x)dx + \int_{-\gamma_2 n}^{-\infty} m(x)f(x)dx$$  \hspace{1cm} (4.11)

where $m(x)$ is the conditional ATE (CATE). We now consider some particular examples corresponding to tail behavior on the error term in the treatment equation and the regressor.

**Logit Errors/Regressors/Bounded CATE** Here we assume the latent error term and the regressors both have a standard logistic distribution, and to simplify calculations we will assume the conditional average treatment effect is bounded on the regressor support. We consider this to be a main example, as the results that arise here (a slower than root-$n$ rate) will generally also arise anytime we have distributions whose tails decline exponentially, such as the normal, logistic and Laplace distributions. From our calculations in the previous section we can see that

$$v(\gamma_2 n) = \gamma_2 n$$  \hspace{1cm} (4.12)

and

$$b_n \approx \gamma_2 n \frac{\exp(-\gamma_2 n)}{1+\exp(-\gamma_2 n)^2}$$  \hspace{1cm} (4.13)

Clearly $v(\gamma_2 n) \to \infty$ resulting in a slower rate of convergence. To get the optimal rate we solve for $\gamma_2 n$ that set $v(\gamma_2 n)/n = b_n^2$. We see this matches up when $\gamma_2 n = \frac{1}{2} \log n$, resulting in the rate of convergence of the estimator:

$$\sqrt{\frac{n}{\log n^{1/2}}} (\hat{\alpha} - \alpha_0) = O_p(1)$$  \hspace{1cm} (4.14)

Furthermore, from the results in the previous section, the estimator will be asymptotically normally distributed, and have a limiting bias.
We notice this the exact same result attained for the binary choice model considered previously. As mentioned, similar slower than parametric rates will be attained when the error distribution and regressor have similar tail behavior.

**Normal Errors/Logistic Regressors/Bounded CATE** Here we assume the latent error term has a standard normal distribution and the regressors both have a standard logistic distribution, and to simplify calculations we will assume the conditional average treatment effect is bounded on the regressor support.

In this situation we have:

\[
v(\gamma_2n) \approx \int_{-\gamma_2n}^{\gamma_2n} \frac{\exp(-x)}{\Phi(x)(1 - \Phi(x))} dx \tag{4.15}
\]

which by multiplying and dividing the above fraction inside the integral by \(\phi(x)\), the standard normal PDF, and using the asymptotic approximation to the inverse Mills ratio, we get

\[
v(\gamma_2n) \approx \int_{-\gamma_2n}^{\gamma_2n} \exp(x^2/2)dx = O(\exp(\frac{1}{2} \gamma_2^2n)) \tag{4.16}
\]

In this setting the bias is approximately of the form

\[
b_n = \int_{\gamma_2n}^{\infty} f(x)dx + \int_{-\gamma_2n}^{-\infty} f(x)dx \approx \exp(-\gamma_2n) \tag{4.17}
\]

so the MSE minimizing value of \(\gamma_2n\) here is

\[
\gamma_2n = O(\sqrt{\log n}) \tag{4.18}
\]

This leads to a rate of convergence that is \(O_p(n^{-1/4})\).

**Differing Support Conditions/Bounded CATE** Here we assume the support of the latent error term in the treatment equation is strictly larger than the support of the regressor and the support of the regression function in the treatment equation. For \(\gamma_2n\) sufficiently large \(p(\gamma_2n)\) remains bounded away from 0 and 1 so \(\lim_{\gamma_2n \to \infty} v(\gamma_2n)\) is finite. We can let \(\gamma_2n\) increase as quickly as possible to remove the bias, resulting in a root-\(n\) consistent estimator that is asymptotically normal, completely analogous to the result attained before under differing supports in the binary choice model.

**Cauchy Errors/Logistic Regressors/Bounded CATE** Here we impose heavy tails on treatment effect error term, assuming it has a Cauchy distribution, but we impose exponential tails on the regressor distribution, assuming here that it has a logistic distribution, though similar results would hold if we assumed normality.

\[
v(\gamma_2n) \approx \gamma_2n \frac{\exp(-\gamma_2n)}{\frac{1}{4} - \arctan(\gamma_2n)^2} \tag{4.19}
\]
where after using L'Hospital’s rule we can conclude that:

\[ v(\gamma_{2n}) \approx \gamma_{2n} \exp(-\gamma_{2n})(1 + \gamma_{2n}^2) \]  

(4.20)

which remains bounded as \( \gamma_{2n} \to \infty \). Therefore we can let \( \gamma_{2n} \) increase arbitrarily quickly in \( n \), say \( \gamma_{2n} = O(n^2) \) in which case the estimator will be root-\( n \) consistent and asymptotically normal with no asymptotic bias.

5 Conclusion

This paper points out a subtle property of estimators based on weighted moment conditions. In some of these models, there is an intimate link between identification and estimability, in that if it is necessary for (point) identification that the weights take arbitrary large values, then the parameter of interest, though point identified, cannot be estimated at the regular rate. This rate depends on relative tail conditions and in some examples can be as slow as \( O(n^{-1/4}) \). Practically speaking, this nonstandard rate of convergence can lead to numerical instability, and/or large standard errors. See for example the recent work of Busso, DiNardo, and McCrary (2008) for some numerical evidence on this based on the ATE parameter. We provide in this paper a studentized approach to inference in which there is no need to know explicitly the rate of convergence of the estimator. While we illustrate our points in the context of two specific models (binary choice and treatment effects), the ideas here can be extended to other models. For example, for randomly censored regression models, inverse weighting (by the censoring probability) procedures are often employed, see, e.g. Koul, Susarla, and Ryzin (1981). It would be worth exploring how much rates of convergence vary with relative tail conditions, and develop rate adaptive procedures, for those models as well.

There are a few remaining issues that need to be addressed in future work. One is the choice of the trimming parameter, \( \gamma_{2n} \), in small samples. We provide bounds on the rate of divergence to infinity of this parameter in the paper. One approach to ‘pick’ a rate is similar to what is typically done for kernel estimators for densities, mainly pick the parameter that minimizes a the first term expansion of the mean squared error. For example, in the ATE case, \( \gamma_{2n} \) can be chosen to minimize the sum of an empirical version of (4.10) and the squared of the empirical version of (4.11).

In addition, in this paper, we saw examples of non-robust point identification in which the model has no information on the parameter if a set with arbitrary small probability is ‘taken out’ of the support. This notion of identification deserves further study also.

References


A Appendix

In section A.1 we provide impossibility results for the binary choice model above. In sections A2-A6, we provide proofs for the theorems in Section 3, for the binary choice model. A7 outlines the proof for the linear representation for the ATE model.

B Appendix: Infinite Bounds

As alluded to earlier in the paper the slower than parametric rates of the inverse weight estimators do not necessarily imply a form of inefficiency of the procedure. The parametric rate is unattainable by any procedure. We show that efficiency bounds are infinite for a variant of model (3.1) above.

Now, introducing regressors $x_i$, we alter the assumption so that $\epsilon|x, v = d \epsilon|x$ and that $E[\epsilon|x] = 0$ (here, $= d$ means “has the same distribution as”). We also impose other conditions such that $\epsilon|x$ and $v|x$ have support on the real line for all $x$. The model we consider now is

$$
y = 1[\alpha + x\beta + v - \epsilon \geq 0]
$$

(B.1)

We restate Theorem 3.1 first and its proof follows.

**Theorem B.1** In the binary choice model (A.7) with exclusion restriction and unbounded support on the error terms, and where $\epsilon|x, v = d \epsilon|x$ and $E[\epsilon|x] = 0$, if the second moments of $x, v$ are finite, then the semiparametric efficiency bound for $\alpha$ and $\beta$ is not finite.

**Remark B.1** The proof of the above results are based on the assumption of second moments of the regressors being finite. This type of assumption was also made in Chamberlain (1986) in establishing his impossibility result for the maximum score model.

**Proof:** We follow the approach taken in Chamberlain (1986). Specifically, look for a subpath around which the variance of the parametric submodel is unbounded. We first define the function:

$$
g_0(t, x) = P(\epsilon_i \leq t|x_i = x)
$$

Note it does not depend on the $v$ because of our assumption of conditional independence. In what follows $t$ will generally correspond to the index $x'\beta + v$. The likelihood function we will be working with is based on the following density function:

$$
f(y, z, \beta, g) = g(x\beta + v, x)y(1 - g(x\beta + v, x))^{1-y}
$$

We first define the family of conditional distributions, $\Gamma$:

**Definition B.1** $\Gamma$ consists of all functions $g : R^k \rightarrow R$ such that for all $(t, x) \in R \times R^{k-1}$ we have
1. $g$ is continuous.

2. $g'(t,x)$, the partial derivative of $g(t,x)$ with respect to its first argument, is continuous and positive.

3. $\lim_{s \to -\infty} g(s,x) = 0, \ \lim_{s \to +\infty} g(s,x) = 1$.

4. $\int s g'(s,x)ds = 0$

We next define the set of sub-paths, $\Lambda$, we will work with:

**Definition B.2** $\Lambda$ consists of the paths:

$$\lambda(\delta) = g_0 C \left( \frac{h}{\delta - \delta_0} \right)$$

where $g_0$ is the “true” distribution function, assumed to be an element of $\Gamma$, $C()$ denotes the cdf of random variable with a cauchy distribution, location parameter 0 and scale parameter 1, and $h : R^{k+1} \to R$ is a positive, continuously differentiable function.

We note that

$$\frac{d}{d\delta} \lambda(\delta)|_{\delta = \delta_0} = \frac{-g_0}{\pi h}$$

So for any $h$ that satisfies $\int s g_0 h ds = 0 \lambda(\delta) \in \Gamma$ for $\delta$ in a neighborhood of $\delta_0$, as all four conditions will be satisfied.

We also note that the scores of the root likelihood function are:

$$\psi_j(y,x,v) = \frac{1}{2} \left\{ y g_0^{-1/2}(x \beta_0 + v, x) - (1 - y) (1 - g_0(x \beta_0 + v, x))^{-1/2} \right\} g'_0(x \beta_0 + v, x)x(j)$$

where $x(j)$ the $j^{th}$ component of $x$ and

$$\psi(\lambda,y,x,v) = \frac{1}{2} \left\{ y g_0^{-1/2}(x \beta_0 + v, x) - (1 - y) (1 - g_0(x \beta_0 + v, x))^{-1/2} \right\} - \frac{g_0(x \beta_0 + v, x)}{\pi h(x \beta_0 + v, x)}$$

We will show that:

**Theorem B.2** Let $I_{\lambda,j}$ denote the partial information for the $j^{th}$ component of $\beta_0$ as defined in Chamberlain (1986).

If $P(x_i \beta_0 + v_i = 0) = 0$ then if the second moment of the vector $(x_i,v_i)$ is finite,

$$\inf_{\lambda \in \Lambda} I_{\lambda,j} = 0$$
Note heuristically we will get the desired result if we define \( \frac{1}{\pi h(t,x,v)} = a(x,v)c(t) \) where \( a(x,v) \) is arbitrarily close to \( f \in |X(\beta_0 + v|x)/g_0(\beta_0 + v) : x(j) \), and \( c(t) \) is the function that takes the value 1 on its support.

To fill in these details, we define \( Q_\lambda \) and \( \Pi \) as:

\[
\Pi(A) = \int_A g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v))^{-1}f_V(v) dF_X(x)
\]

\[
Q_\lambda = \int [b(x\beta_0 + v, x) x(j) - h(x\beta_0 + v, x)]^2 d\Pi(v, x)
\]

where \( b(\cdot, \cdot) = g'_0(\cdot, \cdot)/g_0(\cdot, \cdot) \). We can then add and subtract \( a(x,v) \) to the above integrand, inside the square.

Note we can make the term

\[
\int [b(x\beta_0 + v, x) x(j) - a(x,v)]^2 d\Pi(v, x)
\]

arbitrarily small by the denseness of the space of continuously differentiable functions.

This result, that the above integral can be made arbitrarily small, will follow from Lemma A.2 in Chamberlain (1986) if we can show that \( b(x\beta_0 + v) \in L^2(\Pi) \), where

\[
d\Pi(v, x) = g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))^{-1}f_X(x)f_V(v) dx dv
\]

So in other words all we need to show is the finiteness of the following integral:

\[
\int \frac{g'_0(x\beta_0 + v, x)^2}{g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))} x(j)^2 f_X(x)f_V(v) dx dv
\]  \( (B.3) \)

Finiteness will follow for all distributions satisfying the assumptions in the definition of \( \Gamma \), which will imply the uniform (in \( v,x \) boundedness of the term

\[
\frac{g'_0(x\beta_0 + v, x)^2}{g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))}
\]

This is true because under our assumptions, for any finite \( t \), \( \frac{g'_0(t,x)}{g_0(t,x)(1-g_0(t,x))} \) is finite, so the only possibility of the fraction becoming unbounded is as \( t \to \pm \infty \). However, by considering \( t \to \infty \) and applying Hospitale’s rule, this would be equivalent to the unboundedness of \( \lim_{t \to +\infty} g''_0(t,x) \), (with \( g''_0(t,x) \) the second partial derivative of \( g_0(t,x) \)). But that would contradict \( g'_0(t,x) \) being a density function limiting to 0 as \( t \to \pm \infty \).
The boundedness of the above ratio will imply finiteness of the integral in (A.1) by the finite second moment assumption of $x$ and the fact that $f_V(v)$ is a density function and integrates to 1.

Finally, the term:

$$\int [a(x,v) - a(x,v)c(t)]^2 d\Pi(v,x)$$

can be made arbitrarily small by setting $c(t)$ to 1 in most of its support. Furthermore, with this same definition of $c(t)$ we can satisfy $\int c'(t)dt = 0$.

### B.1 Kernel Regularity Conditions and Uniform Rates

Our proofs to follow will rely on previously established theorems for uniform rates of convergence for kernel density and regression estimators. Such results are based on the following conditions on the kernel function in our kernel density and regression estimator, which is also assumed in, e.g., Lewbel (1997), Collomb and Hardle (1986), and Stoker (1991). See also Newey and McFadden (1994).

1. The kernel function used has support $(-1,1)$ and is of order $p$.
2. The density and regression functions estimated satisfy the $p$th order smoothness conditions described in, e.g., Lewbel (1997) and Stoker (1991).
3. The bandwidth sequence $h_n$ satisfies

   (a) $\sqrt{n}h_n^p \rightarrow 0$

   (b) $nh_n \rightarrow \infty$.

Given these conditions we may appeal to standard theorems related to kernel estimation of density and regression functions. Here, we use results from Collomb and Hardle (1986) which were applied in the context of average derivatives in Stoker (1991) and Hardle and Stoker (1989).

**Theorem B.3** Under the above conditions:

$$\sup_v |\hat{f}(v) - f(v)| I[|v| < \gamma_{2n}] = O_p((n^{1-\delta}h_n)^{-1/2})$$

$$\sup_v |\hat{f}(v) - f(v)/\tilde{f}(v)| I[f(v) > b_n] = O_p(b_n^2n^{1-\delta}h_n^{-1/2})$$

for $\delta > 0$, bandwidth sequence $h_n$, and $b_n \rightarrow 0$. 

24
An important consequence of the above theorem, one that we will be using frequently in our proofs that follow, is that under smoothness conditions on \( f(\cdot) \) which are standard in the literature, we let \( h_n \rightarrow 0 \) sufficiently slowly such that
\[
\sup_v |\hat{f}(v) - f(v)|I[|v| \leq \gamma_2n] = O_p((n^{-1/4-\varepsilon})
\]
for an arbitrarily small \( \varepsilon > 0 \).

### B.2 Linear Representation for Inverse Weighting Estimator

In this section we derive a linear representation for the inverse density weighted estimator. The arguments used in the proof will be based on Powell, Stock, and Stoker (1989), hereafter referred to as PSS. Our linear representation will be the key step in deriving the proof of Theorem 3.1.

We have
\[
\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_2n] = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_2n] + R_n
\]
\[
= \frac{2}{n(n-1)} \sum_{i,j} \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \frac{y_j - I[v_j > 0]}{f(v_j)} \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_2n] + R_n
\]
where \( R_n \) is a remainder term of the form:
\[
R_n = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_2n]
\]

We start by showing that \( R_n \) is \( o_p(n^{-1/2}) \). To do so we split the sum into separate regions. First multiply the term inside the summation by \( I[f(v_i) > (\log n)^{-1}]I[\hat{f}(v_i) > (\log n)^{-1}] \). Now we can apply Theorem A.3 to conclude the term
\[
= O_p(\frac{1}{n^{1-\delta} h_n (\log n)^2}) \sum_{i=1}^{n} \left( \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_2n, f(v_i) > (\log n)^{-1}, \hat{f}(v_i) > (\log n)^{-1}]
\]
This can be made \( o_p(n^{-1/2}) \) by our stated uniform rates for the kernel density estimator and the fact that
\[
\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - I[v_i > 0]}{f(v_i)} \right| I[|v_i| \leq \gamma_2n] = O_p(1)
\]
The above equation follows by the LLN. Note we have \( E[|y_i - I[v_i > 0]|f(v_i)] < \infty \) for the LLN to apply, and follows from the assumption that \( E[\epsilon_i^2] < \infty \).
Next, we consider the average with the indicator $I[0 < f(v_i) \leq (\log n)^{-1}]I[0 < \hat{f}(v_i) \leq (\log n)^{-1}]$ and consider the asymptotic properties of:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right)^{-1} I[|v_i| \leq \gamma_{2n}] I[f(v_i) \in (0, (\log n)^{-1}]$$

The absolute value of this term is bounded by the absolute value of

$$\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_{2n}] I[f(v_i) \in (0, (\log n)^{-1}]$$

which we show is $o_p(n^{-1/2})$.

Let

$$V_n = \frac{y_i - I[v_i > 0]}{f(v_i)} \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_{2n}]$$

and let $\epsilon > 0$ be given; we wish to evaluate

$$P(\sqrt{n}^{-1} \sum_{i=1}^{n} V_n I[f(v_i) \in (0, (\log n)^{-1}]) > \epsilon)$$

We have the event

$$\{\sqrt{n}^{-1} \sum_{i=1}^{n} V_n I[f(v_i) \in (0, (\log n)^{-1}]) > \epsilon \} \implies \{\text{for at least one } i \in \{1, 2, \ldots, n\} \ f(v_i) \in (0, (\log n)^{-1})\}$$

The RHS can be bounded above by

$$1 - P(f(v_i) > (\log n)^{-1})^n \rightarrow_{n \rightarrow \infty} 0$$

Similar arguments can be used for the remaining regions. Therefore, collecting all our results it follows that

$$R_n = o_p(n^{-1/2}) \quad \text{(B.4)}$$

So we will turn attention to the expression for $\hat{\alpha}_n$ without $R_n$, which is a second order $U$-statistic. The (symmetric) kernel of this U-statistic is:

$$m_n(z_i, z_j) = \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \frac{y_i - I[v_i > 0]}{f(v_i)} \left( f(v_i) - \hat{f}(v_i) \right) \right) \left( \frac{\hat{f}(v_i) - f(v_i)}{f(v_i)} \right) I[|v_i| \leq \gamma_{2n}]$$

$$+ \left( \frac{y_j - I[v_j > 0]}{f(v_j)} - \frac{y_j - I[v_j > 0]}{f(v_j)} \left( f(v_j) - \hat{f}(v_j) \right) \right) \left( \frac{\hat{f}(v_j) - f(v_j)}{f(v_j)} \right) I[|v_j| \leq \gamma_{2n}]$$
Now, (as in PSS) define the projection

\[ r_n(z_i) = E(m_n(z_i, z_j)|z_i) \]

This is equal

\[
r_n(z_i) = \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \frac{y_i - I[v_i > 0]}{f(v_i)} \left( \int K_{h_n} (v_j - v_i) f(v_j) \, dv_j - f(v_i) \right) \right) I[|v_i| \leq \gamma_{2n}]
\]

\[
+ \left( \frac{E_j p(v_j) - I[v_j > 0]}{f(v_j)} - \frac{E_j p(v_j) - I[v_j > 0]}{f(v_j)} \left( \int K_{h_n} (v_j - v_i) f(v_j) \, dv_j - f(v_i) \right) \right) I[|v_i| \leq \gamma_{2n}]
\]

\[
= \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_{2n}] - R_{ni}
\]

\[
+ \frac{E_j p(v_j) - I[v_j > 0]}{f(v_j)} I[|v_j| \leq \gamma_n] - \frac{E_j p(v_j) - I[v_j > 0]}{f(v_j)} \left( \int K_{h_n} (v_j - v_i) f(v_j) \, dv_j - f(v_i) \right) I[|v_j| \leq \gamma_n]
\]

\[
- \frac{E_j p(v_j) - I[v_j > 0]}{f(v_j)} I[|v_j| \leq \gamma_n]
\]

\[
= \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_{2n}] - R_{ni} - \left( \int \frac{p(v_j) - I[v_j > 0]}{f(v_j)} K_{h_n} (v_j - v_i) f(v_j) \, dv_j - f(v_i) \right) I[|v_j| \leq \gamma_n]
\]

\[
= \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_{2n}] - R_{ni} - \frac{p(v_i) - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_n] + R_{ni}
\]

where \( E_j \) denotes expectations with respect to variables subscripted \( j \), and \( R_{ni} \) denotes a generic negligible remainder term, satisfying, like (A.2),

\[
\frac{1}{n} \sum_{i=1}^{n} R_{ni} = o_p(n^{-1/2})
\]

What we need to do is to make sure that the projection by itself governs the large sample behavior of the estimator. So, we have (ignoring negligible remainder terms)

\[
\hat{\alpha} = U_n = \frac{2}{n(n-1)} \sum_{i,j} m(z_i, z_j) = U_n^* + (U_n - U_n^*)
\]

where \( U_n^* = \frac{1}{n} \sum r_n(z_i) \) which is what we will show is the statistic that governs the asymptotic distribution of \( \hat{\alpha} \). To show this, we need to prove that

\[
\frac{n}{v(\gamma_{2n})} E(U_n - U_n^*)^2 = o_p(1) \quad (B.5)
\]

A similar condition was shown in PSS. See their lemma 3.1 for the regular root \( n \) case. Following the proof of Lemma 3.1 in PSS, rewrite \( U_n - U_n^* \) as

\[
U_n - U_n^* = \frac{2}{n(n-1)} \sum_{i,j=i+1} q_n(z_i, z_j)
\]
where
\[ q_n(z_i, z_j) = [m_n(z_i, z_j) - r_n(z_i) - r_n(z_j) + E[m_n(z_i, z_j)]] \]

It is sufficient here to derive the order of
\[ E[(m_n(z_i, z_j))^2] \]
and then try to see whether the desired condition (A.3) holds. For this to hold, it is sufficient to compute the order of
\[
E \left[ \left( \frac{y_i - I[v_i > 0]}{f(v_i)} - \frac{y_i - I[v_i > 0]}{f(v_i)} \left( K_n \left( \frac{v_i - v_i}{\ell_n} \right) - f(v_i) \right) I[|v_i| \leq \gamma_n] \right)^2 \right]
\]
which is equal to
\[
E \left( \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_n] \right)^2 + E \left( \frac{y_i - I[v_i > 0]}{f(v_i)} \left( K_n \left( \frac{v_i - v_i}{\ell_n} \right) - f(v_i) \right) I[|v_i| \leq \gamma_n] \right)^2 - 2E \left( \frac{y_i - I[v_i > 0]}{f(v_i)} \right)^2 \left( K_n \left( \frac{v_i - v_i}{\ell_n} \right) - f(v_i) \right) I[|v_i| \leq \gamma_n]
\]
\[ = (1) + (2) + (3) \]

Consider (1).
\[ (1) = \int \frac{(y_i - I[v_i > 0])^2}{f(v_i)} I[|v_i| \leq \gamma_n] dv_i \]
\[ = \int_{-\gamma_n}^{0} \frac{p_i}{f(v_i)} dv_i + \int_{0}^{+\gamma_n} 1 - p_i \frac{1}{f(v_i)} dv_i \]

As a reminder, we have from before
\[ v(\gamma_2n) = \int \frac{p_i(1 - p_i)}{f(v_i)} I[|v_i| \leq \gamma_n] dv_i \]
\[ = \int_{-\gamma_2n}^{0} \frac{p_i(1 - p_i)}{f(v_i)} dv_i + \int_{0}^{\gamma_2n} \frac{p_i(1 - p_i)}{f(v_i)} dv_i \]
\[ \geq K_1 \int_{-\gamma_2n}^{0} \frac{p_i}{f(v_i)} dv_i + K_1 \int_{0}^{\gamma_2n} \frac{1 - p_i}{f(v_i)} dv_i = K_1 \times (1) \]
for some \( 0 < K_1 < \infty \). This is guaranteed to hold since when \( v \) is negative, \( 1 - p(v) \) has to be larger than say \( \epsilon \) for some \( \epsilon > 0 \). Hence
\[ (1) = O_p(v(\gamma_2n)) \]

Now, we find the order of the second term.
\[ (2) = \int \frac{(y_i - I[v_i > 0])^2}{f(v_i)} \left( K_n \left( \frac{v_i - v_i}{\ell_n} \right) - f(v_i) \right)^2 I[|v_i| \leq \gamma_n] dv_i \]
\[ = o_p(1) \]
The above is true by the results in Collomb and Hardle (1986) and Stoker (1991). Finally, it is easy to see that the third term is also of smaller order than (1). Finally,

$$\frac{n}{h_n v(\gamma_{2n})} (A.3) = \frac{n}{h_n v(\gamma_{2n})} O(\frac{1}{n^4}) n^2 O_p(E[(m_n(z_i, z_j))^2])$$

$$= O_p(1/nh_n) = o_p(1)$$

Combining results for these three term, we attain a linear representation for the $U$-statistic. Combining this with our remainder term $R_n$, we have a linear representation for the centered inverse weight estimator:

$$\hat{\alpha} - E[\hat{\alpha}] = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \hat{p}(v_i)}{f(v_i)^2} I[|v_i| \leq \gamma_{2n}] + o_p(\frac{\sqrt{v(\gamma_{2n})}}{n}) + o_p(n^{-1/2})$$

(B.6)

B.3 Proof of Theorem 3.1

We will work with the linear representation established in Section A.4:

$$\hat{\alpha} - E[\hat{\alpha}] = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \hat{p}(v_i)}{f(v_i)^2} I[|v_i| \leq \gamma_{2n}] + o_p(\frac{\sqrt{v(\gamma_{2n})}}{n})$$

(B.7)

We first derive the following result:

$$\hat{S}_n - v(\gamma_{2n}) = o_p(1)$$

(B.8)

Recall

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}(v_i)(1 - \hat{p}(v_i))}{f(v_i)^2} I[|v_i| \leq \gamma_{2n}]$$

(B.9)

which we will write for notational convenience as:

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{m}(v_i)}{f(v_i)^2}$$

(B.10)

which we can write as:

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{m(v_i)}{f(v_i)^2} + \frac{1}{n} \sum_{i=1}^{n} \frac{m(v_i)}{f(v_i)^2} \left( \frac{\hat{m}(v_i)}{m(v_i)} \frac{f(v_i)^2}{f(v_i)^2} - 1 \right)$$

(B.11)

Note that $m(v_i)$ and $f(v_i)$ are both positive and we can use Theorem A.3 to conclude that

$$\left( \frac{\hat{m}(v_i)}{m(v_i)} \frac{f(v_i)^2}{f(v_i)^2} - 1 \right)$$

(B.12)

is negligible in the sense that it averages to a term that is converging to zero, as in A.2. Therefore we have:

$$\hat{S}_n - v(\gamma_{2n}) = \frac{1}{n} \sum_{i=1}^{n} \frac{m(v_i)}{f(v_i)^2} - v(\gamma_{2n}) + o_p(1)$$

(B.13)
where we have used \( E[f^{n(v_i)}] = v(\gamma_{2n}) \) and the LLN. Collecting lead terms we get:

\[
\hat{S}_n = v(\gamma_{2n}) + o_p(1) \tag{B.14}
\]

Therefore we can now write:

\[
\sqrt{n} \hat{S}_n^{-1/2}(\hat{\alpha} - \alpha_0) = \hat{S}_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{y_i - p(v_i)}{f(v_i)} I[|v_i| \leq \gamma_{2n}] + \sqrt{n} \hat{S}_n^{-1/2} o_p(\sqrt{\frac{v(\gamma_{2n})}{n}}) \tag{B.15}
\]

So all we have to do is multiply the above expression by \( 1 = v(\gamma_{2n})^{-1/2}v(\gamma_{2n})^{1/2} \) to conclude that we express the right hand side of the above equation as:

\[
\frac{1}{\sqrt{n}v(\gamma_{2n})} \sum_{i=1}^{n} \frac{y_i - p(v_i)}{f(v_i)} I[|v_i| \leq \gamma_{2n}] + o_p(1) \tag{B.16}
\]

Under Assumption (ii) the desired result follows from the Lindeberg theorem.

### B.4 Optimal Rates for Feasible Estimators

In this section we show our derived optimal rates for the infeasible estimator (where the density function was assumed to be known) carry over to the feasible estimator, where the density is estimated nonparametrically. We do this by showing the bias and variance of the feasible estimator converge at the same rate as the infeasible. We will be using the notation \( E_i[] \) to denote expectation with respect to random variables subscripted by \( i \).

**Bias:** Let the bias be

\[
b(\gamma_{2n}) = E[\hat{\alpha}] - \alpha
\]

\[
= E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_{2n}]\right] - \alpha
\]

\[
= E[\hat{\alpha}] - \alpha - E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{(\hat{f}(v_i) - f(v_i))}{f(v_i)} I[|v_i| \leq \gamma_{2n}]\right]
\]

\[
= \text{Bias(Infeasible)} - E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} E[I[|v_i| \leq \gamma_{2n}]]\right] \]

\[
= \text{Bias(Infeasible)} - o_p(1) E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} E_{ij}[|v_i| \leq \gamma_{2n}]\right]
\]

\[
= \text{Bias(Infeasible)}(1 - o_p(1)) = O_p(\text{Bias(Infeasible)})
\]

where the \( o_p(1) \) term comes from using the same arguments to show \( (A.2) \), though now inside the expectation. See also Hardle and Stoker (1989) for a similar point.

**Variance:** For the variance, we have:

\[
\text{Var}(\hat{\alpha}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[|v_i| \leq \gamma_{2n}] - \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{(\hat{f}(v_i) - f(v_i))}{f(v_i)} I[|v_i| \leq \gamma_{2n}]\right)
\]

\[
= \text{Var}(\text{Infeasible}) + \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} \frac{(\hat{f}(v_i) - f(v_i))}{f(v_i)} I[|v_i| \leq \gamma_{2n}]\right) - 2\text{CovTerm}
\]

30
Let us start with the second term.

\[
\text{Var} \left( \frac{1}{n} \sum y_i - I[v_i > 0] \left( \hat{f}(v_i) - f(v_i) \right) \frac{I(|v_i| \leq \gamma_{2n})}{f(v_i)} \right) = o_p(1) \text{Var}(\text{Infeasible})
\]

This is using arguments similar to above. The leading term of the covariance is:

\[
E \left( \frac{y_i - I[v_i > 0]}{f(v_i)} \right) \frac{y_i - I[v_i > 0]}{f(v_i)} \left( \hat{f}(v_i) - f(v_i) \right) \frac{I(|v_i| \leq \gamma_{2n})}{f(v_i)}
\]

\[
= o_p(1) E \left( \frac{(y_i - I[v_i > 0])^2}{f^2(v_i)} \frac{I(|v_i| \leq \gamma_{2n})}{f(v_i)} \right)
\]

where the \(o_p(1)\) term again arises from using the same arguments we used to show (A.2)

The second term in the covariance is the squared of the “means” times a \(o_p(1)\) term which combined with the above leads to the overall covariance being of order

\[
o_p(1) \text{Var}(\text{Infeasible})
\]

**B.4.1 Interpretable sufficient conditions for (ii) and (iii) in Theorem 3.1**

**Condition (ii):** For a positive \(\delta\),

\[
|X_{ni}| > \epsilon \sqrt{\nu(\gamma_{2n})} \Rightarrow |X_{ni}|^\delta > \epsilon^\delta \sqrt{\nu(\gamma_{2n})}^\delta
\]

which means that (ii) can be bounded by

\[
\frac{E[X_{ni}^{2+\delta}]}{\nu(\gamma_{2n})^{1+\delta/2} \epsilon^{\delta/2}}
\]

so we will assume that \(E[X_{ni}^{2+\delta}]\) is finite.

Using Hospital’s rule as \(\gamma_{2n}\) gets large, we get (again focusing on the right tail)

\[
\frac{(1 - p(\gamma_{2n}))^{1+\delta}}{\hat{f}_V(\gamma_{2n})^\delta \nu(\gamma_{2n})^{\delta/2}}
\]

which by an expansion of \((1 - p(\gamma_{2n}))\) around 0, further simplifies to

\[
\frac{(\gamma_{2n} f_x(\gamma_{2n}))^{1+\delta}}{\nu(\gamma_{2n})^{1/2} \hat{f}_V(\gamma_{2n})^\delta}
\]

The above fraction converges to 0 as long as \(\gamma_{2n} f_x(\gamma_{2n}) \rightarrow 0\), which will happen with exponentially decaying densities such as logistic, normal, laplace, and will also happen with \(t\) distributions. The reason why this condition will suffice is that \(\delta\) can be arbitrarily close to 0 thus slowing the rate of decline to 0 of \(f_V(\gamma_{2n})\) in the above denominator. Therefore, condition (ii) will virtually always be satisfied.

**condition (iii):** This condition ensures that there is no limiting bias in the studentized estimator. Effectively, it ensures the bias of the estimator converges to 0 faster than the standard deviation, resulting in
a slightly sub-optimal rate, analogous to "undersmoothing" in density estimation. Interpretable sufficient conditions for (iii) to hold are that (in the tails) the density of \( \epsilon_i \) converges to 0 faster than the density of \( v_i \), and the density of \( \epsilon_i \) decline at an exponential rate in the tails.

To be more precise and illustrate why these arguments suffice for (iii), we will focus on the right tails of \( \epsilon_i \) and \( v_i \). Letting \( \gamma_n \) denote the trimming parameter, we wish to first evaluate the limit (as \( \gamma_n \to \infty \)) of \( \sqrt{v(\gamma_n)b(\gamma_n)} \). We will write this as \( \sqrt{\frac{v(\gamma_n)}{b(\gamma_n)}}b(\gamma_n)^{3/2} \). First, dealing with the term inside the square root, note both the numerator and denominator involve integrals with \( \gamma_n \) as a limit of integration; thus by Hopitale’s rule this term behaves like

\[
\frac{1}{f_V(\gamma_n)} \frac{p(\gamma_n)}{f(\gamma_n)}
\]

where recall \( p() \) denotes the propensity score, and \( f_V() \) denotes the density function of \( v_i \).

Turning attention to \( b(\gamma_n)^{3/2} \), recall we showed that \( b(\gamma_n) \) declines to 0 as \( \gamma_n f(\gamma_n) \) where \( f() \) denotes the density of \( \epsilon_i \).

So our condition(iii) becomes:

\[
\sqrt{\frac{f_x(\gamma_n)}{f_V(\gamma_n)}} \gamma_n f_x(\gamma_n)\gamma_n^{1/2}
\]

Note if \( \epsilon_i \) has thinner tails than \( v_i \) the term in the square root converges to 0, so we will focus on \( \gamma_n f_x(\gamma_n)\gamma_n^{1/2} \). Now suppose the density of \( \epsilon_i \) has exponential tails: \( f_x(\gamma_n) = \exp(-g(\gamma_n)) \) where \( g(\gamma_n) \) increases to infinity at a polynomial rate, i.e. \( g(\gamma_n) = \gamma_2^n \) where \( \gamma_2 > 0 \). Then, even if \( \gamma_n \) converges to infinity very slowly, i.e. a logarithmic rate (which recall was often close to the "optimal rate" in many of our examples), the term \( \gamma_n f_x(\gamma_n)\gamma_n^{1/2} \) will converge to 0. So we see that both exponential tails of \( \epsilon_i \) and relative thin tails of \( \epsilon_i \) to \( v_i \), will jointly suffice for (iii) to hold. But the exponential tail is grossly sufficient in the sense that with some members of the \( t \) family of distributions, condition (iii) can be satisfied but we need a more delicate argument so (i) is still satisfied. As we know from Lewbel(1997), we have to rule out the student \( t \) with 1 and 2 degrees of freedom anyways just for consistency, but for a student \( t \) with 3 degrees of freedom we could set \( \gamma_n = n^{1/6-\delta} \) for some arbitrarily small positive \( \delta \), and satisfy condition (iii).

Of course (iii) wont be satisfied when the tails of \( v_i \) are too thin relative to \( \epsilon_i \), for example when \( v_i \) has exponential tails and \( \epsilon_i \) is in the \( t \) family. But we note in such cases, as with nonparametric density estimation, an alternative to undersmoothing would be to estimate \( b(\gamma_n) \) and show the centred estimator is asymptotically norm with 0 mean.

### B.5 Outline of Proof of Theorem 4.1

Here, we just outline the proof for linear representation. The rest of the proof will follow using similar arguments as above. The estimator in this section is \( \alpha_n \) in (4.5):

\[
\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{d_i y_i}{p(x_i)} - \frac{(1 - d_i) y_i}{1 - p(x_i)} \right) I[|x_i| \leq \gamma_n]
\]
To get the linear representation below,

\[ X_{in} = \left[ \left( \frac{d_{iy_i}}{p(x_i)} - \frac{(1 - d_{iy_i})y_i}{1 - p(x_i)} - \alpha_0 \right) - \left( \frac{\mu_1(x_i)}{p(x_i)} + \frac{\mu_0(x_i)}{1 - p(x_i)} \right) (d_i - p(x_i)) \right] I[|x_i| \leq \gamma_{2n}] \]

We just consider the first piece:

\[ \frac{1}{n} \sum_{i=1}^{n} \frac{d_{iy_i}}{p(x_i)} I[|x_i| \leq \gamma_{2n}] \]

where

\[ p(x_i) = \frac{1}{n} \sum_{i=1}^{n} d_i K_h(x_i - \frac{x_i}{h_n}) \]

and \( a_i = p(x_i) f(x_i) \) and \( b_i = f(x_i) \). Linearizing the above we get:

\[ \frac{1}{n} \sum_{i=1}^{n} \frac{d_{iy_i}}{p(x_i)} I[|x_i| \leq \gamma_{2n}] = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{d_{iy_i}}{p(x_i)} + \frac{d_{iy_i}}{a_i} (\hat{b}_i - b_i) - \frac{d_{iy_i}b_i}{a_i^2} (\hat{a}_i - a_i) \right) + R(x_i) \]

The above is a U-statistic and so the projection of this U-statistic is as follows. The projection of \( \frac{d_{iy_i}}{a_i} (\hat{b}_i - b_i) \) is

\[ E_j \left[ \frac{d_{iy_j}}{a_j} (\hat{b}_j - b_j) \right] = E_j \left[ \frac{\mu_1(x_j)}{p(x_j)} f(x_j) (\hat{a}_j - a_j) \right] = \frac{\mu_1(x_j)}{p(x_j)} d_i + R_{2ni} \]

while the projection of \( \frac{d_{iy_i}b_i}{a_i^2} (\hat{a}_i - a_i) \) is

\[ E_j \left[ \frac{d_{iy_j}b_j}{a_j^2} (\hat{a}_j - a_j) \right] = E_j \left[ \frac{\mu_1(x_j)}{p(x_j)} f(x_j) (\hat{a}_j - a_j) \right] = \frac{\mu_1(x_j)}{p(x_j)} d_i + R_{2ni} \]

where analogous to \( (A.2) \),

\[ \frac{1}{n} \sum_{i=1}^{n} R_{1ni} = o_p(n^{-1/2}) = R_{2ni} \]

The argument is similar for computing the projection of \( \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - d_{iy_i})y_i}{1 - p(x_i)} I[|x_i| \leq \gamma_{2n}] \) Finally, given the trimming and hence boundedness of \( p(x_i) \) when \( |x_i| \leq \gamma_{2n} \), the remainder term \( R(x_i) \) will satisfy

\[ \frac{1}{n} \sum_{i=1}^{n} R(x_i) = o_p(n^{-1/2}) \]

by the same arguments we used to show \( (A.2) \).