Kernel Smoothing of the Treatment Effect CDF

Jonathan Levy, Mark van der Laan

December 31, 2018
Abstract

The strata-specific treatment effect or so-called blip for a randomly drawn strata of confounders defines a random variable and a corresponding cumulative distribution function. However, the CDF is not pathwise differentiable, necessitating a kernel smoothing approach to estimate it at a given point or perhaps many points. Assuming the CDF is continuous, we derive the efficient influence curve of the kernel smoothed version of the blip CDF and a CV-TMLE estimator. The estimator is asymptotically efficient under two conditions, one of which involves a second order remainder term which, in this case, shows us that knowledge of the treatment mechanism does not guarantee a consistent estimate. The remainder term also teaches us exactly how well we need to estimate the nuisance parameters (outcome model and treatment mechanism) to guarantee asymptotic efficiency. Through simulations we verify theoretical properties of the estimator and show the importance of machine learning over conventional regression approaches to fitting the nuisance parameters. We also derive the bias and variance of the estimator, the orders of which are analogous to a kernel density estimator. This estimator opens up the possibility of developing methodology for optimal choice of the kernel and bandwidth to form confidence bounds for the CDF itself.
1 background and Motivation

The stratum-specific treatment effect or blip (Gill and J. M. Robins 2001) function is defined as random variable given by the average treatment effect for a randomly drawn stratum of confounders. Estimating the cumulative distribution function or CDF of the blip function therefore estimates the proportion of the population that has an average treatment effect at or below a given level. Because clinicians treat patients based on confounders, such is of interest as to evaluating treatment. However, we will see the blip CDF is not pathwise differentiable so we need to estimate a kernel-smoothed version, which can be of interest in and of itself. Such might also provide a pathway to forming confidence intervals for the blip CDF.

Much consideration has been given to the distribution of \( Y_1 - Y_0 \), where \( Y_0 \) is the counterfactual outcome under the intervention to set treatment to \( a \), as per the neyman-rubin potential outcomes framework (Neyman 1923; Donald Rubin 1974). Neyman, 1923, realized that even in estimating the mean of \( Y_1 - Y_0 \), the impossibility of identifying the correlation of \( Y_1, Y_0 \) hampered variance estimation in small samples. Assumptions needed to estimate the joint distribution of \( Y_1 \) and \( Y_0 \) are hard to verify. Cox, 1958 assumes a constant treatment effect for pre-defined subgroups, while Fisher, 1951 suggests one can essentially view the counterfactual \( Y_1 \) estimate the joint distribution of \( Y_1, Y_0 \), as per the neyman-rubin potential outcomes framework (Neyman 1923; Donald Rubin 1974). We will impose the randomization assumption (J. Robins 1986; Greenland and J. Robins 1986), \( Y_a \perp A | W \) as well as positivity, \( 0 < E_P[A = a | W] < 1 \) for all \( a \) and \( W \). Defining \( b_P(W) = E_P[Y | A = 1, W] - E_P[Y | A = 0, W] \)
For a non-parametric model, \(T\) is not pathwise differentiable (van der Vaart [2000]) so instead we consider the smoothed version of the parameter mapping, using kernel, \(k\), with bandwidth, \(\delta\). Here we will suppress \(k\) in the notation for convenience:

\[
\Psi_{\delta,t}(P) = \mathbb{E}_W \int_x \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \mathbb{1}(b(W) \leq x) dx = \int_x \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) F(x) dx
\]

**NOTE:** We assume throughout this paper, \(Pr(b(W) = x) = 0\) for all values, \(x\). In other words, our blip distribution function is continuous.

### 4 Derivation of the Efficient Influence Curve of \(\Psi_{\delta,t}(P)\)

#### 4.0.1 Tangent Space for Nonparametric Model

The true data generating distribution, \(P\), has density, \(p\), which can be factorized as \(p(o) = p_Y(y|a,w)g(a|w)p_W(w)\). We consider the one dimensional set of submodels that pass through \(P\) at \(\epsilon = 0\) (van der Vaart [2000]):

\[
\{P_{\epsilon} \text{ s.t. } p_{\epsilon} = (1 + \epsilon S)p \mid \int SdP = 0, \int S^2dP < \infty\}
\]

The tangent space is the closure in \(L^2\) norm of the set of scores, \(S\), or directions for the paths defined above. We write:

\[
T = \{s(a) \mid \mathbb{E}S = 0, \mathbb{E}S^2 < \infty\} = \{S(y|a,w)\mathbb{E}P_a S = 0, \mathbb{E}S^2 < \infty\} \oplus \{s(a|w)\mathbb{E}P_w S = 0, \mathbb{E}S^2 < \infty\} \oplus \{S(w)\mathbb{E}P_w S = 0, \mathbb{E}S^2 < \infty\}
\]

For a non-parametric model, \(T = L^2(P)\) forms a Hilbert space with inner product defined as \(\langle f, g \rangle = \mathbb{E}f g\).

Our notion of orthogonality now is \(f \perp g\) if and only if \(\langle f, g \rangle = 0\) and, therefore, the above direct sum is valid. In other words, every score, \(S\), can be written as \(\frac{d}{dc} \log(p_{\epsilon})|_{\epsilon=0} = S_Y(y,a,w) + S_A(a|w) + S_W(w)\) where, due to the fact \(p_{\epsilon} = (1 + \epsilon S)p = p_{Y,P_A,P_W}\), it is easy to see \(\frac{d}{dc} \log(p_{\epsilon})|_{\epsilon=0} = S_Y(y,a,w)\).

\[
\frac{d}{dc} \log(p_{\epsilon})|_{\epsilon=0} = S_A(a|w) \text{ and } \frac{d}{dc} \log(p_{\epsilon})|_{\epsilon=0} = S_W(w)
\]

Furthermore we know that a projection of \(S\) on \(T_Y\) is:

\[
S_Y(y \mid w,a) = S(w,a,y) - E[S(W,A,Y) \mid W = w, A = a]
\]

\[
= S(w,a,y) - \int S(w,a,y)p_Y(y \mid w,a)d\nu(y)
\]

\[
= \frac{d}{dc} \log(p_{\epsilon})|_{\epsilon=0}
\]

#### 4.0.2 Efficiency Theory in Brief

The efficient influence curve at a distribution, \(P\), for the parameter mapping, \(\Psi_{\delta,t}\), is a function of the observed data, \(O \sim P\), noted as \(D_{\Psi_{\delta,t}}(P)(O)\). Its variance gives the generalized Cramer-Rao lower bound for the variance of any regular asymptotically linear estimator of \(\Psi_{\delta,t}\) (van der Vaart [2000]). For convenience we define the outcome model \(\bar{Q}(A,W) = E_P[Y \mid A,W]\) and the treatment mechanism as \(g(A \mid W) = E_P[A \mid W]\). We will simplify the notation for the blip here as well, leaving off the subscript.
for the distribution so that \( b(W) = E_P[Y \mid 1, W] - E_P[Y \mid 0, W] \). As in van der Vaart, 2000, we define the pathwise derivative at \( P \) along score, \( S \), as

\[
\lim_{\epsilon \to 0} \left( \frac{\Psi_{\delta,t}(P_\epsilon) - \Psi_{\delta,t}(P)}{\epsilon} \right) \to \Psi_{\delta,t}(S) \tag{1}
\]

We note to the reader, we imply a direction, \( S \), when we write \( P_\epsilon \), which has density \( p(1 + \epsilon S) \), but generally leave it off the notation as understood.

By the riesz representation theorem (Riesz [1909] for Hilbert Spaces, assuming the mapping in (1) is a bounded and linear functional on \( T \), it can be written in the form of an inner product \( \langle D^*(P), g \rangle \) where \( D^* \) is a unique element of \( T \), which we call the canonical gradient or efficient influence curve. Thus, in the case of a nonparametric model, the only gradient is the canonical gradient. It is notable that the efficient influence curve has a variance that is the lower bound for any regular asymptotically linear estimator (van der Vaart 2000). Since the TMLE, under conditions as discussed in this paper, asymptotically achieves variance equal to that of the efficient influence curve, the estimator is asymptotically efficient.

As a note to the reader: Our parameter mapping does not depend on the treatment mechanism, \( g \), and also \( T_A \perp T_Y \perp T_W \) which, means our efficient influence curve must therefore be in \( T_Y \perp T_W \) for the nonparametric model. Therefore, our efficient influence curve will have two orthogonal components in \( T_Y \) and \( T_W \) respectively. We have no component in \( T_A \), which is why we need not perform a TMLE update of the initial prediction, \( g_0 \), of \( g_0(A \mid W) \). Such also teaches us that for the semi-parametric model, where the treatment mechanism is known, the efficient influence function will remain the same.

**Theorem 4.1.** Assume \( k \) is lipschitz and smooth on \( \mathbb{R} \). The efficient influence curve for the parameter, \( \Psi_{t,\delta} \), is given by

\[
D^*_\Psi_{\delta,t}(P)(O) = \frac{-1}{\delta} k \left( \frac{b(W) - t}{\delta} \right) + 2A - \frac{1}{g(A \mid W)} (Y - \bar{Q}(A, W)) + \int \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) \mathbb{1}(b(W) \leq x) dx \]

**PROOF:** Define \( \Phi(x) = 1/(1 + \exp(x)) \). We also define \( b_\epsilon = E_{P_\epsilon}[Y \mid A = 1, W] - E_{P_\epsilon}[Y \mid A = 1, W] \), where \( P_\epsilon \) is defined via its density, \( p_\epsilon = (1 + \epsilon S(o))p(o) \), and \( p \) is the density of \( P \). \( S \) is the so-called score function in Hilbert Space, \( L^2_0(P) \), the completion (under the \( L^2 \) norm) of the space of mean 0 functions of finite variance. We remind the reader that since our model is nonparametric, the tangent space is \( L^2_0(P) \) (van der Vaart 2000). We will now compute the pathwise derivative functional on \( L^2_0(P) \), writing it as an inner product (covariance in the Hilbert Space \( L^2(P) \)), of the score, \( S \), and the efficient influence curve, a unique element of the tangent space, \( L^2_0(P) \). We notate the efficient influence curve as indexed by the distribution, \( P \), and as a function of the observed data, \( O \sim P \): \( D^*(P)(O) \). By dominated convergence we have

\[
\Psi_{\delta,t}(P) = \lim_{h \to 0} \mathbb{E}_w \int_{-h}^{1} \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) \Phi \left( \frac{b(W) - x}{h} \right) dx
\]
\[
\lim_{\epsilon \to 0} \frac{\Psi_{\delta,\epsilon}(P) - \Psi_{\delta,\epsilon}(P)}{\epsilon} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \lim_{h \to 0} \mathbb{E}_w \int \left[ \frac{x-t}{\delta} \right] \left( \Phi\left( b_\epsilon(W) - \frac{x}{h} \right) - \Phi\left( \frac{b(W) - x}{h} \right) \right) dx + \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \mathbb{I}\left( b(W) \leq x \right) - \Psi_{\delta,\epsilon}(P) \right) S(O) dx
\]

(2)

\[
\lim_{\epsilon \to 0} \frac{1}{\epsilon} \lim_{h \to 0} \mathbb{E}_w \int \left[ \frac{x-t}{\delta} \right] \left( \frac{1}{h} \Phi'\left( \frac{b(W) - x}{h} \right) \right) (b_\epsilon(W) - b(W)) dx + \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \left( \mathbb{I}\left( x - \frac{b(W)}{h} \right) \right) \right) (b_\epsilon(W) - b(W))^2 dx
\]

(3)

\[
\lim_{\epsilon \to 0} \frac{1}{\epsilon} \lim_{h \to 0} \mathbb{E}_w \int \left[ \frac{x-t}{\delta} \right] \left( \frac{1}{h} \Phi'\left( \frac{b(W) - x}{h} \right) \right) (b_\epsilon(W) - b(W)) dx + \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \left( \mathbb{I}\left( x - \frac{b(W)}{h} \right) \right) \right) (b_\epsilon(W) - b(W)) dx + R_{2,\delta}(b_\epsilon, b)
\]

(4)

We can note that for \( h(\epsilon) \) such that \( \frac{h(\epsilon)}{b'(\epsilon)} \to 0 \) as \( \epsilon \to 0 \), \( \frac{R_2}{\epsilon} \to 0 \) because \( R_2 \) is order \( \frac{\epsilon}{\epsilon^2} \). To see this, consider the convenient fact that \( \Phi''(x) \) is bounded.

Let’s now drop \( \lim_{\epsilon \to 0} \frac{1}{\epsilon} \) for now and use integration by parts to compute a part of the integrand in (5):

\[
\mathbb{E}_w \lim_{h \to 0} \mathbb{E}_w \left[ \int \left[ \frac{x-t}{\delta} \right] \left( \frac{1}{h} \Phi'\left( \frac{b(W) - x}{h} \right) \right) (b_\epsilon(W) - b(W)) \right] dx
\]

(5)

\[
\mathbb{E}_w \lim_{h \to 0} \mathbb{E}_w \left[ \int \left[ \frac{x-t}{\delta} \right] \left( \frac{1}{h} \Phi'\left( \frac{b(W) - x}{h} \right) \right) (b_\epsilon(W) - b(W)) \right] dx + \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \left( \mathbb{I}\left( x - \frac{b(W)}{h} \right) \right) \right) (b_\epsilon(W) - b(W)) dx
\]

(6)

We can summarize as follows:

\[
\lim_{\epsilon \to 0} \frac{\Psi_{\delta,\epsilon}(P) - \Psi_{\delta,\epsilon}(P)}{\epsilon} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}_w \left( \int \left[ \frac{x-t}{\delta} \right] \left( \frac{b(W) - x}{\delta} \right) (b_\epsilon(W) - b(W)) \right) + \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) \left( \mathbb{I}\left( x - \frac{b(W)}{h} \right) \right) \right) S(O) dx
\]

(7)

As previously stated, the second term disappears by easy choice of \( h \).

\[
\frac{d}{da} p_Y(y|a = 0, w)|_{a=0} = p_Y(y|a, w) S_Y(y|a, w)
\]

\[
= p_Y(y|a, w) (S(o) - \mathbb{E}[S(o)|a, w])
\]
We then compute the pathwise derivative along $S$ at $\epsilon = 0$:

$$
\lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}_w \left( \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) (b_\epsilon(W) - b(W)) \right) = \\
\int \left( \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) \right) \int p_\epsilon(y|a = 1, w) - \frac{d}{da} p_\epsilon(y|a = 0, w) \, dv(y) \, p_w(w) \, dv(w) = \\
\int \left( \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) \right) \int \frac{2a - 1}{g(a|w)} y p_\epsilon(y|a, w) S_\epsilon(o) \, dv(y) \, g(a|w) p_w(w) \, dv(a, w) = \\
\int \left( \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) \right) \int \frac{2a - 1}{g(a|w)} S(o) p(o) \, dv(o) - \int \frac{1}{\delta} Q(a, w) S(o) p(o) \, dv(o) = \\
\left\langle \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) \right\rangle \frac{2A - 1}{g(A|W)} (Y - Q(A, W)) , S(O) \right\rangle_{L_3^2(P)}
$$

Thus we finally get:

$$
\lim_{\epsilon \to 0} \frac{\Psi_{\delta, t}(P_\epsilon) - \Psi_{\delta, t}(P)}{\epsilon} = \mathbb{E}_w \left( \int \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) \, I(b(W) \leq x) - \Psi_{\delta, t}(P) \right) S(O) \, dx + \\
\left\langle \frac{-1}{\delta} k \left( \frac{b(w) - t}{\delta} \right) \right\rangle \frac{2A - 1}{g(A|W)} (Y - Q(A, W)) , S(O) \right\rangle_{L_3^2(P)}
$$

where

$$
D_{\Psi_{\delta, t}}(P)(O) = \frac{-1}{\delta} k \left( \frac{b(W) - t}{\delta} \right) \frac{2A - 1}{g(A|W)} (Y - Q(A, W)) + \int \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) \, I(b(W) \leq x) \, dx - \Psi_{\delta, t}
$$

And this is the efficient influence curve since the canonical gradient is the only gradient for a non-parametric model where the closure of the set of scores is all of $L_3^2(P)$.

**QED**

## 5 The Targeted Maximum Likelihood Estimator, TMLE

We will employ the notation, $P_\alpha f(O)$ to be the empirical average of function, $f(\cdot)$, and $P f(O)$ to be $\mathbb{E}_f f(O)$. Define a loss function, $L(P)(O)$, which is a function of the observed data, $O$, and indexed at the distribution on which it is defined, $P$, such that $E_{P_0} L(P)(O)$ is minimized at the true observed data distribution, $P = P_0$.

The targeted maximum likelihood (TML) estimating procedure maps an initial estimate, $P^0_n \in \mathcal{M}$, of the true data generating distribution to $P^*_n \in \mathcal{M}$ such that $P_n L(P^*_n) \leq P_n L(P^0_n)$ and such that $P_n D_{\Psi_{\delta, t}}(P^*_n) = 0_{d \times 1}$, where $d$, in this case, is the number of points on the CDF. $P^*_n$ is called the TMLE of the initial estimate $P^0_n$ (van der Laan and Daniel Rubin 2006; van der Laan and Rose 2011). For convenience, we define $Q_0(A, W) = E_{P_0}[Y | A, W]$ and its initial estimate, $Q^0_n(A, W)$ and we will use $g_0(A | W) = E_{P_0}[A | W]$ with corresponding initial estimate, $g_n$. The initial estimate of the distribution of $W$ is denoted $Q_{W,n}^0$, the empirical distribution of $W$, with density, $q_{W,n}$. For this paper, the TMLE procedure only adjusts the initial estimate of the outcome regression, leaving $g_n$ and $q_{W,n}$ as is. Thus we will only update $Q^0_n(A, W)$ to its TMLE, $Q^*_n(A, W)$. 

5
To perform the TMLE updating procedure, we may find an element of either a universal least favorable submodel (ulfm), a least favorable submodel (lfm), both defined in van der Laan and Gruber, 2016, or a canonical least favorable submodel (Levy 2018c). Both clfm and ulfm use a single dimensional submodel where as the lfm is of dimension, \( d \), and identical to a clfm if \( d = 1 \). The ulfm has the advantage of not relying on iteration as explained in van der Laan and Gruber, 2016, but here we did not notice an appreciable difference in performance so we used the faster clfm procedure. To construct a clfm, ulfm or lfm, one needs to know the efficient influence curve, which is given by

\[
D_{\Psi_{\delta,t}}(P)(O) = \frac{-1}{\delta} k\left(\frac{b(W) - t}{\delta}\right) * \frac{2A - 1}{g(A|W)} (Y - \bar{Q}(A,W)) + \int \frac{1}{\delta} k\left(\frac{x - t}{\delta}\right) I(b(W) \leq x) dx - \Psi_{\delta,t}
\]

where we estimate the CDF of blip at a given blip value, \( t \), using kernel, \( k \), and bandwidth \( \delta \) (Levy and van der Laan 2018). The CV-TMLE algorithm by the author (Levy 2018a) simplifies the originally formulated CV-TMLE algorithm by Zheng and van der Laan, 2010 and, in this case, turns out to be the same estimator if we use a pooled regression to fit the fluctuation parameter. The TMLE updating procedure is implemented in the software packages blipCDF (Levy 2018b) and (Coyle et al. 2018). Here we will provide for readers more familiar with TMLE, only the so-called clever covariate (van der Laan and Daniel Rubin 2006) for \( \Psi_{h,t} \), but the reader may consult Levy, 2018c for a detailed algorithm.

\[
H(A,W) = \frac{-1}{\delta} k\left(\frac{b(w) - t}{\delta}\right) \frac{2A - 1}{g(A|W)}
\]

If we are simultaneously estimating \( d \) points, \( t_1, t_2, ..., t_d \) on the CDF curve, we will have a \( d - \text{dimensional} \) clever covariate:

\[
(H_1(A,W), H_2(A,W), ..., H_d(A,W)) = \frac{1 - 2A}{\delta g(A|W)} \left(k\left(\frac{b(w) - t_1}{\delta}\right), k\left(\frac{b(w) - t_2}{\delta}\right), ..., k\left(\frac{b(w) - t_d}{\delta}\right)\right)
\]

The TMLE procedure yields \( \bar{Q}^*_n(A,W) \) and our estimator is then a plug-in, using the empirical distribution, \( Q_{W,n} \):

\[
\Psi_{\delta,t}(P^*_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\delta} \int k\left(\frac{x - t}{\delta}\right) I(b^*_n(W_i) \leq x) dx
\]

where \( b^*_n(W_i) = \bar{Q}^*_n(1,W_i) - \bar{Q}^*_n(0,W_i) \), the blip function estimate. For simultaneously estimating many points on the CDF of the blip, the TMLE procedure yields a common outcome model for all \( t \)-values, \( t_1, t_2, ..., t_d \), which has the advantage of preserving monotonicity.

5.0.1 Software

The TMLE is implemented in the software packages blipCDF (Levy 2018b) and (Coyle et al. 2018).

6 TMLE conditions

By solving the efficient influence curve equation with our TMLE update, we may then write a second order expansion, \( \Psi_\delta(P^*_n) - \Psi_\delta(P_0) = (P_n - P_0)D_{\Psi_\delta}(P^*_n) + R_2(P^*_n, P_0) \). We then arrive at the following three conditions (for fixed bandwidth, \( \delta \)) that guarantee asymptotic efficiency for this estimator (van der Laan and Daniel Rubin 2006, van der Laan and Rose 2011), the first of which is not required for CV-TMLE (Zheng and van der Laan 2010). Thus CV-TMLE is our preferred estimator, since it requires less conditions...
on our machine learning, enabling a more aggressive approach to fitting the treatment mechanism and the outcome model.

6.1 TMLE Conditions and Asymptotic Efficiency

We refer the reader to Targeted Learning Appendix (van der Laan and Rose 2011) as well as (van der Laan 2016) for a more detailed look at the theory of TMLE. For convenience, we will summarize some of the main results for the reader.

6.1.1 Conditions for Asymptotic Efficiency

Define the norm \( ||f||_{L^2(P)} = \sqrt{E_P f^2} \). Assume the following TMLE conditions:

1. \( D^\Psi_\delta,t(P^*_n) \) is in a P-Donsker class. This condition can be dropped in the case of using CV-TMLE (Zheng and van der Laan 2010). We show the advantages to CV-TMLE in our simulations.

2. \( R_2(P^*_n, P_0) \) is \( o_P(1/\sqrt{n}) \) for all \( j \).

3. \( D^\Psi_\delta,t(P^*_n) \overset{L^2(P)}{\to} D^\Psi_\delta,t(P_0) \)

then \( \sqrt{n}(\Psi_\delta,t(P^*_n) - \Psi_\delta,t(P_0)) \overset{D}{\to} N[0, \text{var}(D^\Psi_\delta,t(P_0))] \). Our plug-in TMLE’s and CI’s are given by

\[
\Psi_{\delta,t}(P^*_n) \pm z_{1-\alpha} \frac{\hat{\sigma}_n(D^\Psi_{\delta,t}(P^*_n))}{\sqrt{n}}
\]

Under the above conditions, these confidence bands will be as small as possible for any regular asymptotically linear estimator at significance level, \( 1 - \alpha \), where \( Pr(|Z| \leq z_{1-\alpha}) = \alpha \) for Z standard normal and \( \hat{\sigma}_n(D^\Psi_{\delta,t}(P^*_n)) \) is the sample standard deviation of \( \{D^\Psi_{\delta,t}(P^*_n)(O_i) \mid i \in [1:n]\} \) (van der Laan and Daniel Rubin 2006). Note, that if the TMLE conditions hold for the initial estimate, \( P^0_n \), then they will also hold for the updated model, \( P^*_n \) (van der Laan 2016), thereby placing importance on our ensemble machine learning in constructing \( P^0_n \).

For simultaneous confidence intervals, we refer the reader to Levy, van der Laan et al., 2018, for the method which leverages the efficient influence curve approximation to form confidence bounds that simultaneously cover the parameter values at a given significance level. Such inference is as tight as possible and certainly tighter than a standard bonferroni correction (Dunn 1961).

7 The Remainder Term for a TMLE Plug-in Estimator of \( \Psi_{\delta,t}(P_0) \)

In this section we will prove the remainder term of the previous section is \( \frac{1}{\delta} O\left(\|g - g_0\|_{L^2_{\bar{P}_0}} \|Q - \bar{Q}_0\|_{L^2_{\bar{P}_0}}\right) + \frac{1}{\delta} O\left(\|b - b_0\|_\infty\right) \), assuming WLOG that the support of the kernel is \([-1,1]\).

Lemma 7.1. Assume \( \text{lipschitz } F_0 = 1 - S \), where \( S(t) = \mathbb{E}I(b_0(W) > t) \) and assume WLOG the support of the kernel is \([-1,1]\). then
\[
P_0I(b_0(W) > t + \delta, b(W) < t + \delta) = O(||b_0 - b||_\infty)
\]

\[\text{proof:}
\]

\[
P_0I(b_0(W) > t + \delta, b(W) < t + \delta) = P_0I(b_0(W) > t + \delta, b(W) < t + \delta)I(b_0(W) - b(W) > b_0(W) - (t + \delta))
\]

\[
\leq P_0I(b_0(W) > t + \delta, b(W) < t + \delta)I(||b_0 - b||_\infty > b_0(W) - (t + \delta))
\]

\[
\leq Pr(t + \delta < b_0(W) < ||b_0 - b||_\infty + t + \delta)
\]

\[
\text{Lipschitz} \implies \leq L||b_0 - b||_\infty + O(||b_0 - b||^2_\infty)
\]

7
Theorem 7.2. The remainder term, $R_2(P_*, P) = P_0D^*(P) + \Psi(P) - \Psi(P_0)$, is $O(\|b_0 - b\|_\infty^2)$

Proof:

\[ R_2(P_0, P) = P_0D^*(P) + \Psi(P) - \Psi(P_0) \]

\[ = P_0 \left[ \frac{1}{\delta} k \left( \frac{b(W) - t}{\delta} \right) \frac{2A - 1}{g(A[W])} (Y - Q(A, W)) + \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) 1 (b(W) > x) dx - \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) 1 (b_0(W) > x) dx \right] \]

\[ = P_0 \frac{1}{\delta} k \left( \frac{b(W) - t}{\delta} \right) \left( \frac{2A - 1}{g(A[W])} (Y - Q(A, W)) \right) + P_0 \int \frac{1}{\delta} k \left( \frac{x-t}{\delta} \right) 1 (b(W) > x) - 1 (b_0(W) > x) dx \]

\[ = -\frac{1}{\delta} P_0 \left[ k \left( \frac{b(W) - t}{\delta} \right) \left( \left( \frac{g_0(1)|W|}{g(1)|W|} \right) (Q_0(1,W) - Q(1,W)) - \left( \frac{g_0(0)|W|}{g(0)|W|} \right) (Q_0(0,W) - Q(0,W)) \right) \right] + \]

\[ + \frac{1}{\delta} P_0 \left[ \int_{b_0(W)}^{b(W)} k \left( \frac{x-t}{\delta} \right) dx \right] \]

Clearly (12) will disappear if $g_0$ is known. Otherwise the term is $\frac{1}{\delta} \|g - g_0\|_{L_p^2} \|Q - Q_0\|_{L_p^2}$ by cauchy-schwarz.

Now let’s take a look at (13):

\[ \frac{1}{\delta} P_0 \left[ \int_{b_0(W)}^{b(W)} k \left( \frac{x-t}{\delta} \right) dx + k \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) \right] \]

We can divide the $W$ space into disjoint parts and integrate:

a) $t - \delta < b_0(W) < t + \delta$:

Assuming $F_0$ is lipschitz, we have as follows:

\[ \frac{1}{\delta} P_0 \left[ \int_{b_0(W)}^{b(W)} k \left( \frac{x-t}{\delta} \right) dx + k \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) \right] \]

Taylor expanding $k \left( \frac{x-t}{\delta} \right)$ about $\frac{b(W) - t}{\delta}$ we obtain

\[ \frac{1}{\delta} P_0 \left[ \int_{b_0(W)}^{b(W)} k \left( \frac{x-t}{\delta} \right) dx + k \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) \right] \]

Where $\gamma(x, b(W), \delta)$ is an intermediary point

\[ \leq P_0 \left[ \int_{b_0(W)}^{b(W)} k \left( \frac{x-t}{\delta} \right) dx \right] \left( b_0(W) - b(W) \right)^2 \]

Or

\[ \leq \frac{1}{\delta^2} \left( F_0(t + \delta) - F_0(t - \delta) \right) \cdot \|b_0(W) - b(W)\|_{\infty}^2 \]

Employing the Lipschitz condition for $F_0$ we arrive at

\[ \leq \frac{1}{\delta} O(\|b - b_0\|_\infty^2) \]

b) $b_0(W) > t + \delta, b(W) \leq t + \delta$
Theorem 7.3. The asymptotic variance of our TMLE estimator is of order 1/δ. Variance is of order 1 for the cases, b)

b) We will compute the variance of the efficient influence curve and show it is of order 1

\[ \int_{b_0(W)} k \left( \frac{x - t}{\delta} \right) dx + \kappa \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) \]

(19)

\[
\frac{1}{\delta} P_0 I(b_0(W) > t + \delta, b(W) \leq t + \delta) \left[ \int_{b_0(W)} k \left( \frac{x - t}{\delta} \right) dx + \kappa \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) \right]
\]

= \frac{1}{\delta} P_0 \left[ \int_{b_0(W)} k \left( \frac{b(W) - t}{h} \right) \left( \frac{b(W) - t}{h} \right) k' \left( \gamma (b(W), x, t, h) \right) dx + \kappa \left( \frac{b(W) - t}{h} \right) (b_0(W) - b(W)) \right] *

\int_{b_0(W)} I(b_0(W) > t + \delta, b(W) \leq t + \delta)

\leq \frac{1}{\delta} P_0 \int_{b_0(W)} I(t + \delta < b_0(W) < t + \delta + ||b_0 - b||_\infty) \times C(b_0(W) - b(W))^2

= \frac{C}{\delta} \int (b_0 - b)^2 I(t + \delta < b_0(W) < t + \delta + ||b_0 - b||_\infty) dP_{b,b_0}(b_0, b_0)

or \frac{1}{\delta} \int_{b_0(W)} I(b_0 - b)_2 \delta \int_{-\delta}^{\delta} (t + \delta < b_0(W) < t + \delta + ||b_0 - b||_\infty) dP_{b,b_0}(b_0, b_0)

\]

This region follows identically to b).

d) for the cases, b(W) and b_0(W) < t - \delta or b(W) and b_0(W) > t + \delta, we can notice

\[
\int_{b_0(W)} k \left( \frac{x - t}{\delta} \right) dx + \kappa \left( \frac{b(W) - t}{\delta} \right) (b_0(W) - b(W)) = 0.
\]

7.0.1 Variance is of order 1/δ:

Theorem 7.3. The asymptotic variance of our TMLE estimator is of order 1/δ if we satisfy the TMLE conditions of section 6.

We will compute the variance of the efficient influence curve and show it is of order 1/δ, thus proving the point.

\[
\begin{align*}
\mathbb{E} \left[ \frac{1}{\delta^2} k^2 \left( \frac{b(W) - t}{\delta} \right) \left[ \frac{2A - 1}{g(A|W)} (Y - \bar{Q}(A, W)) \right]^2 \right] & \leq \mathbb{E} \left[ \frac{C}{\delta^2} k^2 \left( \frac{b(W) - t}{\delta} \right) \right] \\
& = \mathbb{E} \left[ \frac{C}{\delta^2} \left( \frac{d}{db(W)} \right) \int_{-\infty}^{\infty} I(b(W) \leq x) k^2 \left( \frac{x - t}{\delta} \right) dx \right] \\
& = \mathbb{E} \left[ \frac{C}{\delta^2} \left( \frac{d}{db(W)} \right) \int_{-\infty}^{\infty} I \left( \frac{b(W) - t}{\delta} \leq y \right) k^2 (y) dy \right] \\
& = O \left( \frac{1}{\delta^2} \right)
\end{align*}
\]

Now, we assume k has finite support, WLOG, between [-1, 1] and that \(|k^2(x)| \leq M\)

\[
\begin{align*}
\mathbb{E} \left[ \frac{C}{\delta^2} k^2 \left( \frac{b(W) - t}{\delta} \right) \right] & \leq \mathbb{E} \left[ \frac{C_1}{\delta^2} I \left( -1 \leq \frac{b(W) - t}{\delta} \leq 1 \right) \right] \\
& = \mathbb{E} \left[ \frac{C_1}{\delta^2} \left( -\delta + t \leq b(W) \leq \delta + t \right) \right] \\
& \leq \frac{C_1}{\delta^2} |F(\delta + t) - F(-\delta + t)| \\
& \leq \text{Lipschitz } \frac{C_2}{\delta}
\end{align*}
\]

QED
7.0.2 Order of the bias

\[
\mathbb{E}_{P_{W}} I(b(W) \leq t) - \mathbb{E}_{P_{W}} \int \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) I(b(W) \leq x) dx = \\
F(t) - \int \frac{1}{\delta} k \left( \frac{x - t}{\delta} \right) F(x) dx = \\
F(t) - \int k(y) F(y\delta + t) dy = \\
\int k(y) [F(t) - F(y\delta + t)] dy = \\
\int k(y) \left[ \sum_{i=1}^{\infty} F^{(i)}(t)(y\delta)^{i}/i! \right] dy
\]

where (20) follows, assuming smoothness of the blip CDF function. This will make the order of the bias dependent on the order of the kernel, \( k \). Without smoothness, a lipschitz condition on the blip CDF assures that the bias is of order \( \delta \) from (19).

7.0.3 Generating Kernels

We generate a kernel of order \( K + 1 \) as follows. by generating symmetric polynomial kernels of finite support, the integration can be obtained via an explicit formula and is thus much faster and more accurate than numerical integration. We form polynomials of the form \( k(x) = \sum_{i=0}^{K+2} a_{i}x^{2i} \) where the support of the kernel is from \(-R\) to \(R\). The kernel \( k(\cdot) \) is of course orthogonal to any odd power. To make it order \( K + 1 \) for \( K \) an even positive number, we solve the following equations.

1. make sure the kernel is 0 at the end pts of the support:

\[
\sum_{i=0}^{K+2} a_{i}R^{2i} = 0
\]

2. make sure the kernel has derivative 0 at the end pts of the support in consideration of the remainder term analysis:

\[
\sum_{i=0}^{K+1} 2a_{i}R^{2i+1} = 0
\]

3. To enforce the necessary orthogonality, we solve for \( K > 0 \) and each \( r \) in the 2, 4, ..., 2\( K \)

\[
2 \sum_{i=0}^{K+2} a_{i} \frac{R^{2i+1+r}}{2i + 1 + r} = 0
\]

7.0.4 The Remainder Term Condition for Fixed Bandwidth

Considering section 5 remainder term results, if \( \text{Vert}Q_{n}^{*} - Q_{0}\|_{L^2(P_0)} = o_{P}(n^{r_Q}) \) and the second is \( \|g_{n} - g_{0}\|_{L^2(P_0)} = o_{P}(n^{r_g}) \), then \( r_{Q} + r_{g} \leq -0.5 \) will partially satisfy the TMLE remainder term condition 2 of section 6. However, we also need \( \frac{C}{n}\|b - b_{0}\|_{\infty} = o_{P}(n^{-0.25}) \), in order to guarantee the CV-TMLE estimator is unbiased and asymptotically efficient. Hence, CDF of the blip estimation is not doubly robust and the \( L^\infty \) norm sufficient condition is more demanding than if we have the same requirement but with the \( L^2(P_0) \) norm, as for the variance of the treatment effect (VTE) (Levy, Laan, et al. [2018]). The highly adaptive lasso
The consecutive bandwidths are found then we choose the bandwidth \( n \) consecutive bandwidths that are monotonic estimates with respect to the bandwidth. If no such \( n \) is the order of the first non-zero moment of the kernel (Levy and van der Laan 2018) or so-called order of the kernel. Thus our situation of kernel and bandwidth selection resembles the case for a standard kernel density estimator. However, we are additionally burdened with a second order remainder term that contains \( 1/\delta \) as a factor so we need to establish conditions under which this term does not blow up with increasing sample size. If we allow the bandwidth \( \delta_n \) to go to 0 so as to estimate the parameter, \( \Psi(P) = P(b(W) > t) \) for \( P \in \mathcal{M} \), then our remainder term teaches us that even if we know the true treatment mechanism, as in the case of a randomized trial, we require the sufficient condition

\[
\frac{C}{\delta_n} \| \bar{Q}_0 - \bar{Q}_n^0 \|_\infty = o_P(n^{-0.25})
\]

where we consider \( \delta_n \) going to 0 as \( n \) goes to infinity. Considering the order of the bias and variance, we see that to minimize the MSE, our optimal bandwidth is \( O(n^{-J/4}) \), where \( J \) is the degree of the 1st non-zero moment or so-called order of the kernel. The remainder condition would then require

\[
\| \bar{Q}_0 - \bar{Q}_n^0 \|_\infty = o_P \left( n^{-(2J+3)/(4J+7)} \right)
\]

Thus, perhaps higher order kernels can be useful in that they require a rate closer to \( n^{-0.25} \) for the sup norm bias. The highly adaptive lasso or so-called HAL (van der Laan 2016), guarantees \( \| \bar{Q}_0 - \bar{Q}_n^0 \|_{L^2(P_0)} = O_P(n^{-1/4-1/8(d+1)}) \), where \( d \) is the dimension of \((A, W)\), but the reader may notice this rate is for the less stringent \( L^2 \) norm. Even if we were to have a remainder condition in terms of the \( L^2 \) norm rather than the \( L_\infty \) norm, a high dimensional \((A, W)\) would necessitate a kernel of order greater than \( 4d+3 \).

### 7.0.5 Allowing the Bandwidth to Vanish and Conditions for Asymptotic Normality

The order of the variance is \( 1/\delta \) and the order of the bias is \( \delta^J \), where \( J \) is the degree of the first non-zero moment of the kernel (Levy and van der Laan 2018). We start with a largest trial, we require the sufficient condition

\[
\text{remainder term}
\]

To minimize the MSE, our optimal bandwidth is \( O(n^{-J/4}) \). Thus, perhaps higher order kernels can be useful in that they require a rate closer to \( n^{-0.25} \) for the sup norm bias. The highly adaptive lasso or so-called HAL (van der Laan 2016), guarantees

\[
\| \bar{Q}_0 - \bar{Q}_n^0 \|_{L^2(P_0)} = O_P(n^{-1/4-1/8(d+1)})
\]

where \( d \) is the dimension of \((A, W)\), but the reader may notice this rate is for the less stringent \( L^2 \) norm. Even if we were to have a remainder condition in terms of the \( L^2 \) norm rather than the \( L_\infty \) norm, a high dimensional \((A, W)\) would necessitate a kernel of order greater than \( 4d+3 \).

### 8 Simulations

#### 8.1 Well-specified Models

For well-specified logistic models where the data generating system is given by the following: \( W \) is a random normal, \( Pr(A = 1 \mid W) = g(A \mid W) = \text{expit}(2 + .2 * W) \) and \( E[Y \mid A, W] = \text{expit}(A + 2.5 * A * W + W) \). The TMLE’s using the MLE as an initial estimate performed very well, with normal sampling distributions, nominal coverage (93% or higher), as expected, and did so for all kernels if we used bandwidth \( n^{-1/(2J+1)} \) where \( J \) is the order of the kernel we let \( n \) attain values of 1000, 2500, 5000, 10000, 25000 and 50000. The MSE was lowest for the well-specified MLE plug-in, also as expected, but not appreciably. Hence TMLE is a fine estimator in the near impossible situation where we might get a parametric form correct.

#### 8.1.1 A Method for Choosing Bandwidth for a Given Kernel

We would like to form confidence bounds for the non-pathwise differentiable parameter, \( \Psi(P) = \text{Pr}(b(W) \leq t) \) for \( P \in \mathcal{M} \), and propose using some of the concepts in Chapter 25 of Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies(van der Laan and Rose 2018). We start with a largest bandwidth of size \( n^{1/2J+1} \) where \( J \) is the order of the kernel. Then we divide the bandwidth into 20 equal increments from \( n^{1/2J+1}/20, 2n^{1/2J+1}/20, ..., n^{1/2J+1} \). We then find the smallest set of 5 or more consecutive bandwidths that are monotonic estimates with respect to the bandwidth. If no such 5 or more consecutive bandwidths are found then we choose the bandwidth \( n^{1/2J+1} \). Let us call the consecutive
bandwidth sequence, $B_c = \{h_1, \ldots, h_c\}$, where $h_1$ is the smallest. We also monotonize the variance so as to force it to be increasing as the bandwidth gets smaller. We then form confidence intervals using the monotonized variance for each bandwidth in $B_c$. If the sequence of estimates is decreasing (increasing) as bandwidth decreases (for bandwidths in $B_c$), then we choose the confidence interval with the minimum (maximum) right (left) bound. The idea is that we are minimizing the MSE via this choice, assuming that our region, $B_c$ represents the monotonicity as the bandwidth approaches 0. We still need to refine the theory as our increments for the bandwidth (20 in this case) and definition of being monotonic (5 consecutive or more as described above) are somewhat arbitrary. On the positive side, we noticed coverage of the smoothed parameter maintained nominal levels as $n$ grew to 50,000 when applying our bandwidth selector. Figure 1 below displays the heuristic behind our bandwidth selector.

Table 1: coverage of smoothed parameter, kernel is order 10

| blip | n = 1000 | n = 2500 | n = 5000 | n = 10000 | n = 25000 | n = 50000 |
|------|----------|----------|----------|-----------|-----------|-----------|
| meth | fixed    | meth     | fixed    | meth      | fixed     | meth      |
| 0.145 | 0.907 0.947 | 0.920 0.949 | 0.916 0.941 | 0.935 0.948 | 0.944 0.949 | 0.944 0.948 |
| 0.085 | 0.911 0.953 | 0.950 0.946 | 0.939 0.934 | 0.958 0.962 | 0.942 0.947 | 0.955 0.950 |
| 0.025 | 0.925 0.944 | 0.950 0.960 | 0.958 0.948 | 0.949 0.948 | 0.947 0.941 | 0.948 0.945 |
| 0.005 | 0.916 0.940 | 0.929 0.949 | 0.942 0.966 | 0.949 0.959 | 0.952 0.954 | 0.939 0.937 |
| 0.095 | 0.934 0.951 | 0.934 0.949 | 0.946 0.942 | 0.943 0.943 | 0.944 0.948 | 0.944 0.947 |
| 0.155 | 0.933 0.952 | 0.942 0.946 | 0.936 0.948 | 0.944 0.952 | 0.942 0.946 | 0.948 0.942 |
| 0.215 | 0.927 0.958 | 0.927 0.951 | 0.932 0.941 | 0.934 0.942 | 0.953 0.954 | 0.954 0.939 |
| 0.275 | 0.893 0.955 | 0.913 0.955 | 0.905 0.951 | 0.914 0.935 | 0.926 0.942 | 0.938 0.949 |

meth means we applied the bandwidth selection method, fixed means we used bandwidth $n^{-1/(2^J+1)}$ where J is the kernel order.

Table 2: coverage for true parameter, kernel is order 10

| blip | n = 1000 | n = 2500 | n = 5000 | n = 10000 | n = 25000 | n = 50000 |
|------|----------|----------|----------|-----------|-----------|-----------|
| meth | fixed    | meth     | fixed    | meth      | fixed     | meth      |
| 0.145 | 0.671 0.001 | 0.436 0.298 | 0 0.294 | 0 0.338 | 0 0.325 | 0 |
| 0.085 | 0.612 0.136 | 0.593 0.304 | 0.743 0.001 | 0.836 | 0 0.870 | 0 0.878 | 0 |
| 0.025 | 0.770 0.166 | 0.615 0.019 | 0.405 0.001 | 0.207 | 0 0.076 | 0 0.032 | 0 |
| 0.035 | 0.850 0.071 | 0.924 0.001 | 0.927 | 0 0.938 | 0 0.903 | 0 0.830 | 0 |
| 0.095 | 0.747 0.070 | 0.895 0.000 | 0.912 | 0 0.924 | 0 0.906 | 0 0.801 | 0 |
| 0.155 | 0.750 0.251 | 0.859 0.020 | 0.903 | 0 0.907 | 0 0.911 | 0 0.945 | 0 |
| 0.215 | 0.695 0.947 | 0.717 0.861 | 0.793 0.692 | 0.855 0.370 | 0.867 0.099 | 0.877 | 0 |
| 0.275 | 0.858 0.008 | 0.817 0.707 | 0 0.493 | 0 0.147 | 0 0.010 | 0 |

meth means we applied the bandwidth selection method, fixed means we used bandwidth $n^{-1/(2^J+1)}$ where J is the kernel order.

Figure 1
It is important that our bandwidth selector allows us to still cover the smoothed parameter and we see by
the tables above, that if we apply the selector for kernel order 10, we greatly assist in covering the true
parameter and maintain nominal or very near nominal coverage of the smoothed parameter. Similar results
held for lesser order kernels as well.

8.2 Simulations for Mispecified Models
We call these simulations “mispecified” because we use the highly adaptive lasso or HAL (van der Laan and
Gruber [2016]) to recover the model without any specification on functional forms. The data generating system
consisted of the following functions in the order listed. \( W \) is a random normal, \( \Pr(A = 1 \mid W) = g(A \mid W) = \expit(-0.1 - 0.5 \sin(W) - 0.4 \cdot (|W| > 1) \cdot W^2) \) and \( E[Y \mid A, W] = \expit(3 \cdot A + 5 \cdot A \cdot \sin(W)^2 - A \cdot \cos(W)) \).
We simulated 1100 draws from the above data generating system and computed simultaneous TMLE’s for
the blip values -0.098, -0.018, 0.062, 0.142, 0.222, 0.302, 0.382 and 0.462 using bandwidth 2500\(^{-0.2} \) and an
order 1 polynomial kernel. Similar results held for the uniform kernel.

Here we show the huge advantage of data adaptive estimation in obtaining the initial estimates for TMLE_hal,
using the highly adaptive lasso. HAL will achieve \( n^{-0.25} \) \( L^2 \) rates of convergence to the true outcome
regression and propensity score model, assuming the truth is of bounded sectional variation norm and
CADLAG (continuous from the right and left hand limits) (van der Laan and Gruber [2016]). TMLE_glm
used initial estimates for the propensity score and outcome regression using logistic regression with main
terms and interactions. We can see it is catastrophic to do so here while HAL recovers very close to nominal
coverage and has very little bias. We can see that TMLE helped remove bias from the HAL initial estimates
as well. The results are displayed in Table 2 and Figure 2 below.

| blip     | MSE TMLE_hal | MSE TMLE_glm | coverage TMLE_hal | coverage TMLE_glm |
|----------|--------------|--------------|--------------------|-------------------|
| -0.145   | 0.00083      | 0.02003      | 0.91727            | 0.93            |
| -0.085   | 0.00089      | 0.01683      | 0.92545            | 0.01818          |
| -0.025   | 0.00087      | 0.00528      | 0.93727            | 0.58455          |
| 0.035    | 0.00071      | 0.00373      | 0.94909            | 0.81000          |
| 0.095    | 0.00061      | 0.02173      | 0.96182            | 0.10455          |
| 0.155    | 0.00065      | 0.04723      | 0.95182            | 0              |
| 0.215    | 0.00069      | 0.05803      | 0.94455            | 0              |
| 0.275    | 0.00067      | 0.04528      | 0.94091            | 0              |

simultaneous TMLE_hal coverage was 90%, TMLE_glm coverage was 3%
9 Conclusion

We can see we have developed an estimator with potential to efficiently estimate the kernel smoothed version of the CDF of the blip. We have shown such hinges on data adaptive estimation, in this case with the use of the highly adaptive lasso, to make our initial estimates in the targeted learning (van der Laan and Rose 2011) framework. Then the targeting helps eliminate bias and provides us with an avenue for immediate inference via the sample standard deviation of the efficient influence curve approximation. We have shown, in the basic case of one-dimensional $W$ and well-specified models, choosing the bandwidth of optimal order $n^{(1 - \frac{1}{2d+1})}$ (see section 10.1.1 with $d = 1$) provides normal and unbiased sampling distributions for the smoothed parameter. The next step is to develop a way to optimally (smallest MSE possible) select the bandwidth and kernel so that the estimator minus the truth blown up by $\sqrt{n\delta}$ is normally distributed. What we have shown in this paper is a first step but our method of determining monotonicity of the parameter as bandwidth vanishes is somewhat arbitrary. It also remains to be seen how monotonicity generally holds for small bandwidths. For instance, if the monotonicity changes direction for a small bandwidth, our proposed selector might be problematic.
References

Coyle, Jeremy et al. (2018). tmle3. URL: https://github.com/tlverse/tmle3

Dunn, Olive Jean (1961). “Multiple Comparisons Among Means”. In: Journal of the American Statistical Association 56.293, pp. 52–64.

Frechet, Maurice (1951). “Sur Les Tableaux de Correlation Dont Les Marges Sont Donnees”. In: Annals University Lyon A.14, pp. 53–77.

Gill, Richard D and James M Robins (2001). “Causal inference for complex longitudinal data : the continuous case.” In: Report Eurandom, Eindhoven: Eindhoven University of Technology 2001023.

Greenland, Sander and James Robins (1986). “Identifiability, Exchangeability, and Epidemiological Confounding”. In: International Journal of Epidemiology 15.3.

Heckman, James J. and Jeffrey Smith (1997). “Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts”. In: Review of Economic Studies 64, pp. 487–535.

Heckman, James J. and Jeffrey A. Smith (1998). “Evaluating the Welfare State”. In: National Bureau of Economic Research 6542.

van der Laan, Mark (2016). “A Generally Efficient Targeted Minimum Loss Based Estimator”. In: U.C. Berkeley Division of Biostatistics Working Paper Series 343. URL: http://biostats.bepress.com/ucbbiostat/paper343

van der Laan, Mark and Susan Gruber (2016). “One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels”. In: The International Journal of Biostatistics 12(1), pp. 351–378.

van der Laan, Mark and Sherri Rose (2011). Targeted Learning. New York: Springer.

— (2018). Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies (2018). Springer International Publishing AG.

van der Laan, Mark and Daniel Rubin (2006). “Targeted Maximum Likelihood Learning”. In: U.C. Berkeley Division of Biostatistics Working Paper Series 213. URL: http://biostats.bepress.com/ucbbiostat/paper213.

Levy, Jonathan (2018a). “An Easy Implementation of CV-TMLE”. In: arXiv:1811.04573 [stat.ME]. URL: arxiv.org/abs/1811.04573

— (2018b). blip CDF. URL: https://github.com/jlstiles/blipCDF

— (2018c). “Canonical Least Favorable Submodels: A New TMLE Procedure for Multidimensional Parameters”. In: ArXiv e-prints. URL: https://arxiv.org/abs/1811.01261.

Levy, Jonathan, Mark van der Laan, et al. (2018). “A Fundamental Measure of Treatment Effect Heterogeneity”. In: arXiv:1811.03745 [stat.ME]. URL: https://arxiv.org/abs/1811.03745

Levy, Jonathan and Mark van der Laan (2018). “Kernel Smoothing of the Treatment Effect CDF”. In: arXiv. URL: https://arxiv.org/abs/1811.06514

Neyman, Jerzy (1923). “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9”. In: Statistical Sciences 5.4. Translated and edited by D. M. Dubrowska and T. P. Speed from the Polish original which appeared in Roczniki Nauk Rolniczych Tom X, 1923, pp. 465–480.

Pearl, Judea (2000). Causality: Models, Reasoning and Inference. New York: Cambridge University Press, p. 484.

Riesz, Frigyes (1909). “Sur les opérations fonctionnelles linéaires”. In: C.R. Academy of Sciences Paris 149, pp. 974–977.

Robins, James (1986). “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period”. In: Journal of Mathematical Modeling 7, pp. 1393–512.

Rubin, Donald (1974). “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”. In: Journal of Educational Psychology 66.5, pp. 688–701.

Strotz, RH and HO Wold (1960). “Recursive vs. nonrecursive systems: an attempt at synthesis (part I of a triptych on causal chain systems)”. In: Econometrica 28, pp. 417–427.
van der Vaart, Aad (2000). *Asymptotic Statistics*. Vol. Chapter 25. Cambridge, UK: Cambridge University Press.
Wright, Sewall (1921). “Correlation and Causation”. In: *Journal of Agricultural Research* 20, pp. 557–585.
Zheng, Wenjing and Mark van der Laan (2010). “Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation”. In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 273.