On the multiply robust estimation of the mean of the g-functional

Andrea Rotnitzky\(^1\), James M. Robins\(^2\) and Lucia Babino\(^3\)

\(^1\)Di Tella University and CONICET, Buenos Aires, Argentina.

\(^2\)Departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

\(^3\)Instituto de Calculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina.

May 23, 2017

Contents

1 Introduction

2 Assumptions and the target of inference

\hspace{1cm} 2.1 Examples

\hspace{2cm} 2.1.1 Example 1.

\hspace{2cm} 2.1.2 Example 2.

\hspace{2cm} 2.1.3 Example 3.

\hspace{2cm} 2.1.4 Example 4.

3 Representations of the parameter of interest
3.1 Inverse probability weighted representation of the g-functional

3.2 Iterated conditional mean representation

   3.2.1 Interpretation of $\eta_k$ and $Y_k\left(\underline{g}_k\right)$ in Example 1.

   3.2.2 Interpretation of $\eta_k$ and $Y_k\left(\underline{g}_k\right)$ in Example 2.

   3.2.3 Interpretation of $\eta_k$ and $Y_k\left(\underline{g}_k\right)$ in Example 3.

   3.2.4 Interpretation of $\eta_k$ and $Y_k\left(\underline{g}_k\right)$ in Example 4.

4 Estimation based on parametric models for the nuisance functions

   4.1 Inverse probability weighting estimation

   4.2 Fully parametric maximum likelihood estimation

   4.3 Iterated regression estimation

   4.4 Weighted iterated regression

   4.5 Doubly robust estimation by iterated regression

   4.6 $K+1$ - multiply robust estimation

      4.6.1 The Bang and Robins estimator is $K+1$ - multiply robust

      4.6.2 The greedy iterated fit $K+1$ - multiply robust estimators

      4.6.3 The inverse probability weighted regression $K+1$ - MR estimators

   4.7 $2^K$ - multiply robust estimation

      4.7.1 Theoretical results background

      4.7.2 Iterated regressions of multiply robust outcomes

      4.7.3 Inverse probability weighted iterated regression.
Abstract

We study multiply robust (MR) estimators of the longitudinal g-computation formula of Robins (1986). In the first part of this paper we review and extend the recently proposed parametric multiply robust estimators of Tchetgen-Tchetgen (2009) and Molina, Rotnitzky, Sued and Robins (2017). In the second part of the paper we derive multiply and doubly robust estimators that use non-parametric machine-learning (ML) estimators of nuisance functions in lieu of parametric models. We use sample splitting to avoid the need for Donsker conditions, thereby allowing an analyst to select the ML algorithms of their choosing. We contrast the asymptotic behavior of
our non-parametric doubly robust and multiply robust estimators. In particular, we derive formulas for their asymptotic bias. Examining these formulas we conclude that although, under certain data generating laws, the rate at which the bias of the MR estimator converges to zero can exceed that of the DR estimator, nonetheless, under most laws, the bias of the DR and MR estimators converge to zero at the same rate.
1 Introduction

The goal of this paper is to construct multiply robust estimators of functionals defined by the longitudinal g-computation formula (aka g-formula) of Robins (1986) from \( n \) i.i.d. observations \( Z_i, i = 1, \ldots, n \). These g-functionals are widely studied in the causal inference literature, a leading special case being the functional corresponding to the expectation of a counterfactual response from longitudinal data under the assumption of no unmeasured confounding (Robins 1986, 1987, 1997). In this setting \( Z_i \) denotes all of subject \( i \)'s observed treatment, covariate, and outcome history over the study period. This is the third in a series of papers on multiply robust estimation that reports on results obtained by the authors and coworkers between 2012-2014. The first paper in the series (Molina, Rotnitzky, Sued and Robins, 2017) is to appear in Biometrika, the second (Babino, Rotnitzky and Robins, 2017) will hopefully appear in Biometrics. A fourth should be available later this year.

G-functionals depend on the observed data law only through the conditional distributions of outcome and covariate given past treatments, covariates and outcomes. Estimation of g-functionals requires the estimation of infinite dimensional nuisance parameters, such as a conditional mean or a conditional density. As such, g-functionals cannot be estimated consistently under the non-parametric model that includes all possible data laws. Therefore either finite-dimensional parametric models or non-parametric models with smoothness or sparsity constraints are often considered. Both parametric and nonparametric approaches have been used by different authors to estimate the nuisance functions.

We now provide a broad overview of the paper. Consider first the parametric case. Robins (2000, 2002) and Bang and Robins (2005) introduced a class of iterated conditional expectation estimators for g-functionals which they showed were doubly robust (DR), i.e. the estimators are asymptotically linear, and thus consistent and asymptotically normal (CAN), for the g-functional of interest provided either (i) parametric models for the conditional laws of treatment given past treatments, outcomes and covariates are correct for each treatment time \( k, k = 1, \ldots, K \), or (ii)
parametric models for certain iterated conditional expectations (ICEs) depending on the conditional distributions of outcome and covariates given past treatments, and covariates are correct at each time $k$. The estimators in Robins (2000) and and Bang and Robins (2005) were defined as the solutions to estimating equations, while those in Robins (2002) were plug-in estimators. Because of their similarity, we refer to all of these estimators as B&R estimators, although in this paper we consider only the plug-in form. Van der Laan and Gruber (2012) and Petersen, Schwab, Gruber, Blaser, Schomaker and van der Laan (2014) proposed DR estimators nearly identical to the plug-in version of the B&R estimator which they refer to as Targetted Maximum Likelihood Estimators (TMLEs). See Section 4.6.2 for additional discussion.

It has recently been shown by Molina et al. (2017) that the B&R estimator and thus the TMLE estimators confer more protection to model misspecification than had been thought. Specifically, Molina et al. proved that these estimators are asymptotically linear so long as the first $k$ conditional treatment models are correct and the last $K - k$ iterated expectation models are correct for any $k \in \{1, ..., K\}$. Thus, the aforementioned DR estimators are all actually $K + 1$ robust. In section 4.6 we review this result and several $K + 1$ robust estimators.

In fact, it is possible to construct so-called multiply robust (MR) estimators of the g-functional. MR estimators can be exponentially more robust to model misspecification than the $K + 1$ robust estimators. In particular these estimators are asymptotically linear and thus CAN for the g-functional of interest if a parametric model for either the time $k \in \{1, ..., K\}$ treatment probability or the time $k$ ICE is correct, thus providing $2^K$ opportunities to be CAN.

Tchetgen-Tchetgen (2009) constructed an iterated augmented inverse probability weighted (IAIPW-MR) estimator of a specific g-functional, namely the mean of an outcome at the end of a longitudinal study with monotone missing at random data. Molina, Rotnitzky, Sued and Robins (2017) derived a general theory for the existence of multiply robust estimators of functionals in non or semiparametric models whose likelihood factorizes as the product of variation independent factors with the functional depending on just one of these factors. Construction of an IAIPW-MR estimator of an arbitrary g-functional follows by application of the Molina et al. general theory. Both Tchetgen-Tchetgen (2009) and Molina et al. (2017) estimate the nuisance high dimensional functionals parametrically. Inverse augmented IPW multiple robust estimators that fit parametric models for the conditional treatment probabilities and ICEs are reviewed in section 4.7.2.
Iterated augmented IPW multiply robust estimators of g-functionals are not entirely satisfactorily because they do not respect bounds on the state space of the g-functional of interest. To address this problem, in sections 4.7.3 and 4.7.4 we derive two classes of multiply robust iterated conditional expectation plug-in estimators that fit parametric models for the conditional treatment laws and the ICEs, as did Tchetgen-Tchetgen (2009) and Molina et al. (2017).

Unfortunately, it is quite likely that all our parametric models for the $2K$ nuisance functions are misspecified. If so, parametric DR and MR estimators will be inconsistent, motivating the need for non-parametric estimators. Because the time-specific conditional treatment probabilities and the ICEs are infinite dimensional conditional densities and conditional expectations one would expect that nonparametric approaches to their estimation would be more robust than parametric approaches discussed above. In order to discuss the non-parametric approach we need to be more specific, as we now do.

For all parametric and non-parametric doubly and multiply robust estimators $\hat{\theta}$ of a g-functional $\theta$, the difference $\hat{\theta} - \theta$ can be decomposed as the sample average of a mean zero, finite variance, random variable $IF_\theta(P) = if_\theta(Z, P)$ plus a remainder $R$. If $R$ is $o_p(n^{-1/2})$ then $\hat{\theta}$ is asymptotically linear (and hence CAN) and $IF_\theta(P)$ is its influence function. Now, the remainder $R$ can be further decomposed as the sum $R_1 + R_2$ of two terms; $R_1$ is an empirical process term discussed below and $R_2$ is a drift term. It is well known (van der Vaart, 1998, ch. 25) that in a nonparametric model defined solely by smoothness or sparsity assumptions, all asymptotically linear estimators have the same influence function $IF_\theta(P)$. It follows that $\hat{\theta}$ will be an asymptotically linear estimator of $\theta$ if and only if both the empirical process term $R_1$ and the drift $R_2$ are $o_p(n^{-1/2})$.

The exact form of the drift of an MR estimator is given in equation (62) in Section 5.2 but is too complex to give here. For the purpose of our introduction it suffices to point that the drift has the following general form

$$E_P \left[ \sum_{k=1}^{K} \left\{ \frac{1}{h_k(Past_k)} - \frac{1}{\hat{h}_k(Past_k)} \right\} \{\eta_k(Past_k) - \hat{\eta}_k(Past_k)\} \hat{c}_k(Past_k) \right]$$

(1)

where (i) $h_k(past_k)$ and $\eta_k(past_k)$ are the true conditional treatment probability and ICE function at $k$, (ii) $Past_k$ is the random vector denoting the data recorded up to $k$ and $past_k$ denotes a possible realization of $Past_k$, (iii) $\hat{h}_k(past_k)$ and
\( \hat{\eta}_k(past_k) \) are estimates of \( h_k(past_k) \) and \( \eta_k(past_k) \). (iv) \( \hat{c}_k(past_k) \) is an order 1 random variable and (v) the functions \( \hat{\eta}_k(\cdot), \hat{h}_k(\cdot) \) and \( \hat{c}_k(\cdot) \) are considered as fixed functions when taking the expectation. Note if, for every \( k \), either \( \hat{h}_k(\cdot) = h_k(\cdot) \) or \( \hat{\eta}_k(\cdot) = \eta_k(\cdot) \), then the drift is zero. This fact underlies the asymptotics of our parametric MR estimators.

Now, non-parametric estimators of \( \eta_k(past) \) and \( h_k(past) \) cannot be \( n^{1/2} \)-consistent even under smoothness or sparsity constraints. Thus, our only hope for the drift to be \( o_p\left(n^{-1/2}\right) \) is that the functions \( \eta_k(past) \) and \( h_k(past) \) are sufficiently smooth or sparse in some basis so that, at each \( k \), \( \hat{h}_k(past) \) converges to \( h_k(past) \) and, \( \hat{\eta}_k(past) \) converges to \( \eta_k(past) \), at rates \( n^{-\alpha_k} \) and \( n^{-\beta_k} \) in such a way that the expectation (1) is \( o_p\left(n^{-1/2}\right) \); a sufficient condition is that \( \alpha_k + \beta_k > 1/2 \) for each \( k \).

The particular estimators \( \hat{h}_k(past) \) and \( \hat{\eta}_k(past) \) that obtain the best rates of convergence will vary depending on the unknown smoothness or sparsity of \( h_k(past) \) and \( \eta_k(past) \). Thus one would wish to implement various non-parametric estimators and use the data to adaptively choose those with the fastest rates of convergence. As is well known, this can be accomplished by using, say \( J \) machine learning algorithms to construct candidate estimators and then using cross validation to choose the best candidate for each of the \( 2K \) nuisance functions \( h_k(past) \) and \( \eta_k(past) \). (Dudoit and Van der Laan, 2003). Even for \( J \) polynomial in the sample size, this approach will generally achieve, for each nuisance function, a rate of convergence equal to the rate of the machine learning algorithm with the fastest convergence rate among the \( J \) algorithms. However, note that although this approach may give the best rate of convergence of the drift to 0 given the \( J \) machine learning algorithms, this rate could be slower than \( o_p\left(n^{-1/2}\right) \) because one or more of the functions \( h_k(past) \) and/or \( \eta_k(past) \) might not be smooth or sparse enough.

Even when the drift is \( o_p\left(n^{-1/2}\right) \), \( \hat{\theta} \) will an asymptotically linear estimator only if the empirical process term \( R_1 \) in the remainder \( R \) is also \( o_p\left(n^{-1/2}\right) \). To describe this term we need additional notation. Each of our DR and MR estimators \( \hat{\theta} \) are either exactly equal to, or asymptotically equivalent to, a sample average

\[ \mathbb{P}_n \left\{ m\left(Z, \hat{h}, \hat{\eta}\right) \right\} \]

of a random variable \( m\left(Z, \hat{h}, \hat{\eta}\right) \) that depends on subject \( i \)'s data and on the \( K \)-vectors of nuisance functions \( \hat{h} \equiv (\hat{h}_1, ..., \hat{h}_K) \) and \( \hat{\eta} \equiv (\hat{\eta}_1, ..., \hat{\eta}_K) \) obtained through the cross validation procedure described above. The empirical process term \( R \), also called the stochastic equicontinuity term, of each
of our DR and MR estimators is
\[
\left\{ \mathbb{P}_n[m(Z, \hat{h}, \hat{\eta})] - E_P[m(Z, h, \eta)] \right\} - \left\{ \mathbb{P}_n[m(Z, h, \eta)] - E_P[m(Z, h, \eta)] \right\}
\]
where the functions \( \hat{h} \) and \( \hat{\eta} \) are again treated as non-random when taking the expectation over \( Z \), even though they are actually random because estimated from the same data \( Z_i, i = 1, \ldots, n \). It is well known that if \( m(\cdot, \hat{h}, \hat{\eta}) \) and \( m(\cdot, h, \eta) \) lie in a Donsker class with probability one and \( \hat{h} \) and \( \hat{\eta} \) are \( L_2 \)-consistent for \( h \) and \( \eta \), then the stochastic equicontinuity term is \( o_p\left( \frac{n}{n-1/2} \right) \) as required.

However, for the outputs \( (\hat{h}, \hat{\eta}) \) of an arbitrary machine learning program, \( m(\cdot, \hat{h}, \hat{\eta}) \) cannot be assumed to lie in a Donsker class. In section 5.1 we describe how to overcome this problem by splitting the sample and using a cross-fit estimator, a name coined in Chernozhukov (2016). To obtain a cross-fit estimator we first randomly split the sample into \( U \), say 5, equal sized subsamples \( u = 1, \ldots, U \). For each split \( u \) we construct an estimator \( \hat{\theta}_u \) of \( \theta \). Then our cross-fit MR estimator is
\[
\hat{\theta}_{cf}^{MR} = U^{-1} \sum_{u=1}^{U} \hat{\theta}_u = U^{-1} \sum_{u=1}^{U} \mathbb{P}_{n/U} \left\{ m\left(Z, \hat{h}^u, \hat{\eta}^u\right) \right\}
\]
with \( \hat{\theta}_u = \mathbb{P}_{n/U} \left\{ m\left(Z, \hat{h}^u, \hat{\eta}^u\right) \right\} \), \( \mathbb{P}_{n/U} \) denotes the average over the \( n/U \) units in split \( u \) and the \( 2K \) estimated functions \( \hat{h}^u \equiv (\hat{h}_1^u, \ldots, \hat{h}_K^u), \hat{\eta}^u \equiv (\hat{\eta}_1^u, \ldots, \hat{\eta}_K^u) \) are obtained by machine learning as in the previous paragraph, but using data only on the \( n(U - 1)/U \) units not in the \( u \)th split. When \( \hat{h}^u \) and \( \hat{\eta}^u \) are \( L_2 \)-consistent for \( h \) and \( \eta \) then
\[
U^{-1} \sum_{u=1}^{U} \left[ \mathbb{P}_{n/U} \left\{ m\left(Z, \hat{h}^u, \hat{\eta}^u\right) \right\} - E \left\{ m\left(Z, \hat{h}^u, \hat{\eta}^u\right) \right\} \right] - \mathbb{P}_n \left\{ m\left(Z, h, \eta\right) - E_P \left\{ m\left(Z, h, \eta\right) \right\} \right\}
\]
is \( o_p\left( n^{-1/2} \right) \). This implies that \( \left( \hat{\theta}_{MR}^{cf} - \theta \right) = \mathbb{P}_n \{ IF_\theta (P) \} + R_2 + o_p\left( n^{-1/2} \right) \). Thus, if the drift is \( o_p\left( n^{-1/2} \right) \), our machine learning cross-fit estimator \( \hat{\theta}_{MR}^{cf} \) is an asymptotically linear estimator of \( \theta \), where from here on we treat the terms non-parametric and machine learning as synonyms.

Robins et al. (2008, p. 379) earlier used sample splitting to avoid the Donsker requirement in constructing efficient asymptotically linear estimators of functionals.
in non-parametric models, although their estimator did not use cross-fitting. Subsequently, Ayygari (2010) in his 2010 Harvard Phd. Thesis used a machine learning cross-fit estimator to obtain an asymptotically linear estimator of the parameter \( \theta \) in the semiparametric regression model
\[
E \left[ Y \mid A, X \right] = \theta A + \tau(X)
\]
thereby avoiding the Donsker requirement. This work was subsequently published as Robins et al. (2013). Zheng and van der Laan (2010) proposed a so-called cross-validated TMLE that used sample-splitting to avoid some, but perhaps not all, of the need for Donsker conditions. The Zheng and van der Laan estimator is quite similar, but not identical, to our doubly robust estimator of \( \theta \) of Section 5, called in that section \( \hat{\theta}_{DR,CF,mach,bang} \). Belloni et al. (2010) proposed a cross-fitting estimator to relax the degree of sparsity required to obtain an asymptotically linear instrumental variable estimator. The idea that sample-splitting and cross-fitting could be used to avoid the need for Donsker conditions long preceded any of the above references - for example, Van der Vaart (1998, page 391) - although the idea of explicitly combining cross-fitting with machine learning was not emphasized.

Recall that in the parametric setting MR estimators have \( 2^K \) opportunities to be CAN for \( \theta \) compared to \( K + 1 \) opportunities for DR estimators. In Section 5.2 we consider whether the marked advantage of MR over DR estimators carries over to the nonparametric setting by comparing their drifts. We find that although, under certain data generating laws, the advantage persists, nonetheless, under most laws, the drift of the DR and MR estimators converge to zero at the same rates and, thus, the MR estimators advantage does not persist.

In further detail, we can approximate the drift of a nonparametric cross-fit DR estimator \( \hat{\theta}^{cf}_{DR} \) given in (61) of Section 5.2 by the sum of the drift of the MR estimator \( \hat{\theta}^{cf}_{MR} \) given above plus the quantity

\[
E_P \left[ \sum_{1 \leq j < k \leq K} \left( \frac{1}{h_j(Past_j)} - \frac{1}{h_j(Past_j)} \right) (\eta_k(Past_k) - \hat{\eta}_k(Past_k)) \hat{c}_{j,k}(Past_{max\{k,j\}}) \right]
\]

where \( \hat{c}_{j,k}(past_{max\{k,j\}}) \) is an order 1 random variable.

It follows that the drift of \( \hat{\theta}^{cf}_{DR} \) has \( K(K - 1)/2 \) terms more than the drift than that of \( \hat{\theta}^{cf}_{MR} \). However, the rates of convergence of \( \hat{\theta}^{cf}_{MR} \) and \( \hat{\theta}^{cf}_{DR} \) to \( \theta \) are determined by the dominating term in their drifts, i.e. the term with the slowest rate of
convergence to zero. One would generally expect that term

\[ E_P \left[ \frac{1}{h_K(Past_K)} - \frac{1}{\hat{h}_K(Past_K)} \right] \left( \eta_K(Past_K) - \hat{\eta}_K(Past_K) \right) \left( \hat{c}_K(Past_K) \right) \]

be the dominating term in both the drift of the MR and the DR estimators because this term contains two regressions involving the entire history Past_K, which is a superset of the conditioning set in the regressions involved in all other terms. Thus, one would generally expect that \( \hat{\theta}^{cf}_{DR} \) and \( \hat{\theta}^{cf}_{MR} \) have drifts that converge to zero at identical rates.

However, it could happen that at the particular law \( P \) that generated the data, one of the \( K(K-1)/2 \) terms appearing in the drift of \( \hat{\theta}^{cf}_{DR} \) but not in the drift of \( \hat{\theta}^{cf}_{MR} \) converges to 0 slower than any of the terms in \( \hat{\theta}^{cf}_{MR} \). In such case, the drift of \( \hat{\theta}^{cf}_{MR} \) would have a faster rate of convergence to 0 than the drift of \( \hat{\theta}^{cf}_{DR} \). In particular, it could happen that \( \hat{\theta}^{cf}_{MR} \) is an asymptotically linear estimator of \( \theta \) even though \( \hat{\theta}^{cf}_{DR} \) is not. The frequency with which the law generating the data has the drift of \( \hat{\theta}^{cf}_{MR} \) converging to 0 faster than the drift of \( \hat{\theta}^{cf}_{DR} \) may be greater for \( K \) large, because the ratio of the number of terms in the drift of \( \hat{\theta}^{cf}_{DR} \) compared to the drift of \( \hat{\theta}^{cf}_{MR} \) increases linearly with \( K \), providing an increasing number of opportunities for the drift of \( \hat{\theta}^{cf}_{DR} \) to dominate the drift of \( \hat{\theta}^{cf}_{MR} \).

The paper is organized as follows. In section (3) we define the g-functional, review various representations for it and examples of its application. Section (4) discusses estimation of the g-functional based on parametric models for the nuisance functions. Sections 4.1, 4.2 and 4.3 reviews non-doubly robust IPW, parametric MLE and ICE plug-in estimators respectively. Section 4.4 and 4.5 give preliminary background on doubly-robust estimation. Section 4.6 considers three different DR plug-in estimators, and shows that they are, in fact, \( K + 1 \) robust. Section 4.7 discusses \( 2^K \) MR estimation, specifically, the theoretical background and three particular estimators, two of which are ICE plug-in estimators. Section 5 considers non-parametric DR and MR estimation. We propose a number of cross-fit machine learning DR and MR estimators and analyze their asymptotic properties.
2 Assumptions and the target of inference

Let \( Z = (Z_1, \ldots, Z_N, L_{K+1}) \) where \( Z_k = (A_k, L_k) \), \( k = 1, \ldots, K \), and \( A_k \) and \( L_k \) are, possibly multivariate, random vectors taking values in measurable spaces \( (A_k, A_k) \) and \( (L_k, L_k) \). Let \( Z = \otimes_{k=1}^K (A_k \times L_k) \times L_{k+1} \) and \( \mathcal{Z} = \otimes_{k=1}^K (A_k \times L_k) \times L_{K+1} \) and let \( \mathcal{P} \) be the collection of the densities of all probability measures on \( (Z, \mathcal{Z}) \) mutually absolutely continuous with respect to \( \mu = \otimes_{k=1}^K (\mu_k \times \mu'_k) \times \mu'_{k+1} \), where for each \( k \), \( \mu_k \) are measures on \( (A_k, A_k) \) and \( (L_k, L_k) \) respectively. For each \( P \in \mathcal{P} \), we write \( p = dP/d\mu \) and \( p(z) = \prod_{k=0}^K g_k (l_{k+1}|\overline{l}_k, \overline{a}_k) \prod_{k=1}^K h_k (a_k|\overline{l}_k, \overline{a}_{k-1}) \), or for short \( p = gh \), where \( g_k (l_{k+1}|\overline{l}_k, \overline{a}_k) \) and \( h_k (a_k|\overline{l}_k, \overline{a}_{k-1}) \) are (versions of) the conditional densities of \( L_{k+1} \) and \( A_k \) when \( Z \sim P \). Here and throughout for any vector \( w = (w_1, \ldots, w_s) \) and any \( r \leq t \leq s \), \( \overline{w}_r = (w_r, \ldots, w_t) \), \( \overline{w}_r = \overline{w}'_r \) and \( \overline{w}_r = \overline{w}_r \). Furthermore, \( A \perp\!\!\!\perp B \mid C \) denotes that \( A \) and \( B \) are conditionally independent given \( C \), \( [K] \) denotes the set \( \{1, \ldots, K\} \), \( gh^* \ll gh \) stands for \( p^* = gh^* \) is absolutely continuous with respect to \( p = gh \), \( E_{gh} (\cdot) \) stands for expectation under \( p = gh \) and, often we write \( (\overline{L}_{K+1}, \overline{A}_K) \) instead of \( Z \).

In this paper we are interested in inference about the parameter

\[
\theta (g) \equiv E_{gh^*} \{ \psi (\overline{L}_{K+1}) \}
\]

based on \( n \) i.i.d. copies of the random vector \( Z \) with unknown distribution \( p = gh \) assumed to belong to model \( \mathcal{P} \), where \( \psi \) is a given real valued measurable function on \( (Z, \mathcal{Z}) \) and \( h_k^* (a_k|\overline{l}_k, \overline{a}_{k-1}) \) is a given, i.e. known, conditional density for each \( k \in [K] \), such that \( p^* = gh^* \) is absolutely continuous with respect to \( p = gh \).

By definition, the parameter \( \theta (g) \) depends on the unknown data generating law \( p = gh \) only through \( g \equiv (g_0, \ldots, g_K) \). Explicitly,

\[
\theta (g) = \int \varphi (z) \prod_{k=0}^K g_k (l_{k+1}|\overline{l}_k, \overline{a}_k) \, d\mu (z)
\]

(2)

where \( (\overline{l}_0, \overline{a}_0) \equiv \text{nill} \) and

\[
\varphi (z) \equiv \left\{ \prod_{k=1}^K h_k^* (a_k|\overline{l}_k, \overline{a}_{k-1}) \right\} \psi (\overline{L}_{K+1})
\]

(3)
is a known, i.e., specified, function of \( z \).

The expression on the right hand side of (2) is often referred to as the g-computation formula (Robins, 1986), or, the g-formula for short. Our motivation for studying the functional \( \theta (g) \) is because special choices of \( h_\ast^k \) yield \( \theta (g) \) equal to parameters which are of interest in causal inference and in missing data analysis. Here we give some examples.

2.1 Examples

2.1.1 Example 1.

Mean of an outcome in a longitudinal study with ignorable drop-out. Consider a longitudinal study with drop-outs. Define \( L_k \) to be the data vector \( L_\ast^k \) that is recorded on a subject randomly selected from a target population if the subject is still on study at the \( k \)th study cycle and to be equal to an arbitrary vector in \( L_k \), say \( x_k \), otherwise. Assume no subject misses the first cycle. Then \( L_1 = L_\ast^1 \). Let \( A_k = 1 \) if the subject is on study at the \( (k + 1) \)th study cycle and \( A_k = 0 \) otherwise. Thus,

\[
L_k = A_{k-1} L_\ast^k + (1 - A_{k-1}) x_k.
\]

Let \( p = \prod_{j=0}^{K} g_j \prod_{j=1}^{K} h_j \) be the law of \( (\overline{A}_K, \overline{L}_{K+1}) \). Under the missing at random assumption that \( L_{K+1}^\ast \perp \perp A_k \mid (\overline{A}_{k-1} = 1, \overline{L}_{k-1}) \) for each \( k \in [K] \), and the positivity assumption that for all \( k \in [K] \), \( \Pr \{ h_k (1 | \overline{A}_{k-1} = 1, \overline{L}_{k-1}) > 0 \} = 1 \), the mean of the, potentially missing, last cycle outcome \( L_{K+1}^\ast \), i.e., of the outcome that would be recorded if the study did not suffer from drop-out, equals

\[
E_{g_0} [E_{g_1} \ldots E_{g_{K-1}} \{ E_{g_K} (L_{K+1} | \overline{A}_K = 1, \overline{L}_K) \mid \overline{A}_{K-1} = 1, \overline{L}_{K-1} \} \ldots | \overline{A}_1 = 1, \overline{L}_1]] .
\] (4)

In this display, as well as in the expression for the positivity assumption and throughout the rest of the paper, subscripts on \( E \) are used to indicate the sole conditional laws on which the expectation or probability depends on. For instance, in (4) the subscript \( g_{K+1} \) is a reminder that \( E_{g_K} (L_{K+1} | \overline{A}_K = 1, \overline{L}_K) \) depends only on \( g_K \).

The expression in (4) agrees with \( \theta (g) \) if we take \( h_k^\ast (a_k | \overline{l}_k, \overline{a}_{k-1}) = a_k \) and \( \psi (\overline{l}_{K+1}) = l_{K+1} \) (Robins, 1986, Robins, Rotnitzky and Zhao, 1995). Note that the
positivity assumption is the same as the assumption that \( gh^* \ll gh \). Note also that because \( A_k \) is a binary variable, \( \theta (g) \) actually involves only integrals over \( l_1, \ldots, l_{K+1} \) as, for each \( k \), the integral over \( a_k \) is indeed a sum with a single non-zero term.

### 2.1.2 Example 2.

**Outcome mean under a sequence of fixed treatments.** Suppose that in a longitudinal study \( L_k \) denotes the vector of variables to measured at the \( k\)th study cycle on a subject randomly selected from a target population. Assume that immediately after recording \( L_k \) the subject decides which of the available treatments in a set \( A_k \) he will take until the next study cycle. Let \( A_k \in A_k \) denote the subject’s treatment choice. Let \( p = \prod_{j=0}^{K} g_j \prod_{j=1}^{K} h_j \) be the law of \((A_K, L_{K+1})\). Also, let \( L_{K+1,:} \) be the counterfactual outcome at the end of follow-up if, possibly contrary to fact, the subject took treatment \( A_k = a^* \) for some fixed \( a^* = (a^*_1, \ldots, a^*_{K+1}) \). Contrasts of the mean of \( L_{K+1,:} \) involving different \( a^* \) quantify treatment effects. For instance, the average treatment effect (ATE) comparing the always on treatment vs never on treatment regimes is defined as the mean of \( L_{K+1,:} \) minus the mean of \( EL_{K+1,:} \).

Under the consistency assumption that \( A_k = a_k \Rightarrow L_k = L_{k+1,:} \) for all \( k \in [K] \), the no-unmeasured confounding assumption that for \( k \in [K] \), \( L_{K+1,:} \perp \perp A_k | (A_{k-1} = \overline{a^*}_{k-1}, L_{k-1}) \) and the positivity assumption that for \( k \in [K] \), \( \Pr \{ h_k (a^*_k | a_{k-1}^*, L_{k-1}) > 0 \} = 1 \), the mean of \( L_{K+1,:} \) equals (Robins, 1986)

\[
E_{g_0} \left[ \prod_{k=1}^{K+1} \left\{ E_{g_k} \left( L_{K+1,:} | A_k = \overline{a^*}_k, L_k \right) \right\} | A_{K-1} = \overline{a^*}_{K-1}, L_{K-1} \right] . \tag{5}
\]

This expression agrees with \( \theta (g) \) if we take \( h_k^* (a_k | \overline{a}_k, \overline{a}_{k-1}) = I_{\{a^*_k\}} (a_k) \) and \( \psi (L_{K+1}) = l_{K+1} \) where throughout, \( I_D (x) = 1 \) if \( x \in D \) and \( I_D (x) = 0 \) otherwise. Note that in Example 1 we could arrive at the formula (4) from the formula (5) if, in that example we regard \( A_k \) as a sequence of time-dependent treatments indexed by \( k \) and consider estimation of the mean of \( L_{K+1} \) had, contrary to fact, all subjects followed the treatment regime specified by \( a_k = 1 \) for \( k \in [K] \); that is, the regime in which no subject had dropped-out. Robins (1986, p. 1491; 1987a, sec. AD.5) provided additional discussion of the usefulness of regarding missing data indicators as time-dependent treatments.
2.1.3 Example 3.

**Outcome mean under a non-random dynamic treatment regime.** Assume that the recorded data $Z$ are as in the longitudinal study of Example 2. However, suppose that we are now interested in estimating the mean of $L_{K+1}$ if, contrary to fact, the entire study population followed a given non-random dynamic treatment regime which stipulates that right after study cycle $k$ and until just prior to study cycle $k+1$, a patient with covariate and treatment history $(\alpha_{k-1}, l_k)$ receives treatment $A_k = d_k(\alpha_{k-1}, l_k)$. Similarly to Example 2, the average treatment effect for comparing the two such regimes, say $d$ and $d'$ is defined as the mean of $L_{K+1}, d$ minus the mean of $L_{K+1}, d'$ where for any treatment regime $d = \{d_1, \ldots, d_K\}$, $L_{K+1, d}$ denotes the counterfactual outcome at the end of the study if, possibly contrary to fact, the subject had followed treatment regime $d$. Under the consistency assumption that $A_k = D_k \Rightarrow L_{k+1} = L_{k+1, d}$, where for any $j \in [K], D_j \equiv d_j(\alpha_{j-1}, l_j)$, the no-unmeasured confounding assumption that for $k \in [K], L_{K+1, d} \perp \perp A_k | (\alpha_{k-1} = D_{k-1}, l_{k-1})$, and the positivity assumption that for $k \in [K], \Pr \left( A_k = D_k | (\alpha_{k-1} = D_{k-1}, l_{k-1}) > 0 \right) = 1$, the mean of $L_{K+1, d}$ is

$$E_{g_0} \left[ E_{g_1} \left[ \cdots E_{g_{K-1}} \left( E_{g_K} (L_{K+1} | A_K = D_K, l_K) | A_{K-1} = D_{K-1}, l_{K-1} \right) \cdots | A_1 = D_1, l_1 \right] \right].$$

This expression agrees with $\theta(g)$ if we take $h^*_k (a_k | \alpha_{k-1}, l_{k-1}) = I_{\{d_k(\alpha_{k-1}, l_{k-1})\}}(a_k)$ and $\psi(l_{K+1}) = l_{K+1}$. Note also that the positivity assumption is the same as the assumption that $gh^* \ll gh$.

2.1.4 Example 4.

**Outcome mean under a random dynamic treatment regime.** Assume that the recorded data $Z$ are as in the longitudinal study of Example 2. Suppose that we are now interested in estimating the mean of $L_{K+1}$ if, contrary to fact, the entire population followed a random dynamic treatment regime which stipulates that at study cycle $k$ a patient with covariate and treatment history $(\alpha_k, l_{k+1})$ is randomized to receive treatment $A_{k+1} = a_k$ with probability $h^*_k (a_k | \alpha_{k-1}, l_k)$ where $a_k$ is in the set $A_k$ of treatments available at time $k$. Similarly to Example 2, the average treatment effect for comparing the two regimes, determined by $h^*$ and $h^{**}$, is defined as the mean of $L_{K+1, h^*}$ minus the mean of $L_{K+1, h^{**}}$ where for any
\( h^* \equiv \{ h^*_k : k \in \{ K \} \} \), \( L_{K+1,h^*} \) denotes the counterfactual outcome if, possibly contrary to fact, the subject had followed the random treatment regime \( h^* \). Under the consistency assumption that \( A_{h^*,k} = \overline{A}_k \) \( \Rightarrow \) \( L_{K+1} = L_{K+1,h^*} \) for all \( k \in \{ K \} \) where \( A_{h^*,k} \) is the treatment received at cycle \( k \) when the subject follows the random regime, the no-unmeasured confounding assumption that \( L_{K+1,h^*} \perp \perp A_k \mid (A_k - 1 = a_k, L_k = l_k) \) for all \( (a_k, l_k) \) such that \( \prod_{j=1}^{k} h^*_j (a_j \mid \overline{a}_{j-1}, \overline{l}_j) > 0 \) and the positivity assumption that \( gh^* << gh \), the average treatment effect is precisely equal to \( \theta (g) \) if we take \( \psi (\overline{l}_{K+1}) = l_{K+1} \).

3 Representations of the parameter of interest

Throughout the paper we assume that \( gh^* << gh \) where \( p = gh \) is the unknown law of \((L_{K+1}, \overline{A}_K)\). Robins (1993) noted that the parameter \( \theta (g) \) admits two representations, which we now review. These representations are important as they give rise to two distinct estimation strategies that we will review in the next section. Define the random variables

\[
\pi^*_j \equiv \prod_{r=j}^{k} h^*_r (A_r \mid \overline{L}_r, \overline{A}_{r-1})
\]

\[
\pi^*_j \equiv \prod_{r=j}^{k} h_r (A_r \mid \overline{L}_r, \overline{A}_{r-1}), \pi^k \equiv \pi^*_1 \text{ and } \pi^{*k} \equiv \pi^{*k}_1. \text{ Also, given } p = gh, g_k h_k \text{ stands for the } \overline{g}_k = \prod_{j=0}^{k} g_j \prod_{j=1}^{k} h_j.
\]

3.1 Inverse probability weighted representation of the g-functional

The Radon-Nykodim theorem implies that

\[
\theta (g) = E_{gh} \left\{ \psi (\overline{L}_{K+1}) \pi^{*K} / \pi^K \right\}
\]

This motivates the so-called inverse probability weighted estimators of \( \theta (g) \) discussed in section 4.1. Notice that when, as in Examples 1, 2 and 3, \( h^*_k (\cdot \mid \overline{a}_{k-1}, \overline{l}_k) \) is the indicator of following a given non-random treatment regime at study cycle \( k \),
\( \pi^{*K} \) is the indicator of having followed the regime through the entire study and \( \pi^K \) is the product of the conditional probabilities of following the regime at each study cycle for a subject with recorded data \((L_{K+1}, A_K)\). So, the right hand side of (7) is interpreted as the population weighted mean of \( \psi(L_{K+1}) \) among those subjects that follow the regime through the entire study weighted by the inverse of their probability of complying with the regime.

### 3.2 Iterated conditional mean representation

Robins (1986, 1997) derived another representation of \( \theta(g) \) in the form of an iterated conditional expectation.

\[
\eta_k(\bar{a}_k, \bar{l}_k; g_k) \equiv \int \psi(L_{K+1}) \prod_{j=k+1}^{K} h^*_j(a_j|l_j, a_{j-1}) \prod_{j=k}^{K} g_j(l_{j+1}|l_j, a_j) \left\{ \prod_{j=k+1}^{K+1} d\mu(j) \prod_{j=k+1}^{K} d\mu_j(a_j) \right\}
\]

for \( k \in [K] \), where \( \prod_{j=K+1}^{K} \equiv 1 \) and for any \( f = (f_1, \ldots, f_K), \), \( f_k \equiv (f_k, \ldots, f_K) \). Note that for any \((\bar{a}_k, \bar{l}_k)\) such that \( \prod_{j=1}^{k} h_j(a_j|l_j, a_{j-1}) \prod_{j=0}^{k-1} g_j(l_{j+1}|l_j, a_j) > 0 \), it holds that

\[
\eta_k(\bar{a}_k, \bar{l}_k; g_k) \equiv E_{\bar{a}_k, \bar{l}_k, g_k} \psi(L_{K+1}) | A_k = \bar{a}_k, L_k = \bar{l}_k.
\]

Define also, \( Y_{K+1}(g_{K+1}) \equiv \psi(L_{K+1}) \), and for \( k \in [K - 1] \), if \( \pi^{*k} > 0 \) define

\[
Y_{k+1}(g_{k+1}) \equiv y_{k+1, \pi_{k+1}} \left( A_{k+1}, L_{k+1}; g_{k+1} \right)
\]

\[
\equiv E_{h_{k+1}^*} \left\{ \eta_{k+1} \left( A_{k+1}, L_{k+1}; g_{k+1} \right) | A_k, L_{k+1} \right\}
\]

\[
= \int \eta_{k+1} (a_{k+1}, A_k, L_{k+1}; g_{k+1}) h_{k+1}^* (a_{k+1} | A_k, L_{k+1}) d\mu_{k+1} (a_{k+1})
\]

It immediately follows that for any \( k \in [K] \), if \( \pi^{*k} > 0 \) then

\[
\eta_k(\bar{a}_k, L_k; g_k) = E_{g_k} \left\{ Y_{k+1}(g_{k+1}) | A_k, L_k \right\},
\]

and

\[
\theta(g) = E_{g_0} \left\{ Y_1 (g_1) \right\}.
\]
3.2.1 Interpretation of $\eta_k$ and $Y_k \left( g_k \right)$ in Example 1.

Under the conditional independence and positivity assumptions made in this example, $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ evaluated at $\overline{a}_k = \overline{T}$ coincides with $E \left( L^*_k \mid \overline{A}_k = \overline{T}, \overline{L}_k = \overline{t}_k \right)$, i.e., the mean of the intended outcome $L^*_k$ among subjects that are still on study at study cycle $k+1$, i.e., with $\overline{A}_k = \overline{T}$, and that have recorded past $\overline{L}_k = \overline{t}_k$ up to time $t_k$ (Robins, 1986, 1997). Note that by the assumed conditional independence $L^*_{k+1} \perp \perp A_k \mid (\overline{A}_{k-1} = \overline{T}, \overline{L}_{k-1})$, this conditional mean is the same as $E \left( L^*_{k+1} \mid \overline{A}_{k-1} = \overline{T}, \overline{L}_k = \overline{t}_k \right)$. So, we can interpret $\eta_k \left( \overline{a}_k = \overline{T}, \overline{t}_k; g_k \right)$ as the best predictor of $L^*_k$ for subjects that are still on study at study cycle $k$ given the observed data $\overline{L}_k = \overline{t}_k$. On the other hand, $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ has no meaningful interpretation when $a_j = 0$ for some $j < k$. Nevertheless, we need not worry about this interpretation because the values taken by the function $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ when $\overline{a}_k \neq \overline{T}$ are irrelevant. This is because $\theta \left( g \right)$ does not depend on them. To interpret $Y_k \left( g_k \right)$ notice that this is only defined for units with $\pi^*_k > 0$, i.e., for units with $\overline{A}_{k-1} = \overline{T}$. For these units $Y_k \left( g_k \right)$ equals $\eta_k \left( \overline{a}_k = \overline{T}, \overline{L}_k; g_k \right)$ because in this example $h^*_k \left( a_k \mid \overline{A}_{k-1} = \overline{T}, \overline{L}_k \right) = a_k$.

3.2.2 Interpretation of $\eta_k$ and $Y_k \left( g_k \right)$ in Example 2.

Consider, for some fixed $\alpha^* = (a_1^*, \ldots, a_K^*)$, the mean of $L_{K+1 \mid \alpha^*}$. This equals $\theta \left( g \right)$ under the consistency, no-unmeasured confounding and positivity assumptions when we take $h^*_k \left( a_k \mid \overline{A}_{k-1} = \overline{T}, \overline{L}_k \right) = I \left\{ a_k^* \right\} \left( a_k \right)$. The interpretation of $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ for $\overline{a}_k = \overline{a}_k^*$ is identical to the one just given for Example 1, replacing $\overline{T}$ with $\overline{a}_k^*$ and $L^*_k$ with $L_{K+1 \mid \alpha^*}$. For $\overline{a}_k \neq \overline{a}_k^*$, the interpretation of $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ is irrelevant because, just as in Example 1, $\theta \left( g \right)$ does not depend on the values taken by $\eta_k \left( \overline{a}_k, \overline{t}_k; g_k \right)$ when $\overline{a}_k \neq \overline{a}_k^*$. Also, in analogy to Example 1, $Y_k \left( g_k \right)$ is defined only for units with $\overline{A}_{k-1} = \overline{a}_{k-1}^*$. For these units, $Y_k \left( g_k \right)$ equals to $\eta_k \left( \overline{a}_k = \overline{a}_k^*, \overline{L}_k; g_k \right)$.
3.2.3 Interpretation of $\eta_k$ and $Y_k(\underline{g}_k)$ in Example 3.

Under the consistency, conditional independence and positivity assumptions made in this example, the function $\eta_k(\underline{\pi}_k, \underline{\ell}_k; \underline{g}_k)$ evaluated at $\underline{\pi}_k = \underline{d}_k(\underline{\pi}_{k-1}, \underline{\ell}_k)$ where $\underline{d}_k(\underline{\pi}_{k-1}, \underline{\ell}_k) = [d_1(l_1), d_2(a_1, \ell_2), \ldots, d_k(\underline{\pi}_{k-1}, \underline{\ell}_k)]$, coincides with $E \{ L_{K+1,d_k|\underline{\pi}_k = \underline{d}_k(\underline{\pi}_{k-1}, \underline{\ell}_k), \underline{\ell}_k = \underline{l}_k} \}$, i.e. the mean of the counterfactual outcome $L_{K+1,d_k}$ among subjects that remain compliers with the treatment regime $d$ at the $k+1^{th}$ cycle, and that have recorded past $\underline{\ell}_k = \underline{l}_k$ (Robins, 1986, 1997). As in Examples 1 and 2, the interpretation of $\eta_k(\underline{A}_k, \underline{\ell}_k; \underline{g}_k)$ when $\underline{A}_k \neq \underline{d}_k(\underline{A}_{k-1}, \underline{\ell}_k)$ is irrelevant since $\theta(g)$ does not depend on it. Also, in analogy to Example 1, $Y_k(\underline{g}_k)$ is only defined for units with $\underline{A}_{k-1} = \underline{d}_{k-1}(\underline{A}_{k-2}, \underline{\ell}_{k-1})$. For these units, $Y_k(\underline{g}_k)$ equals to $\eta_k(\underline{A}_k = \underline{d}_k(\underline{A}_{k-1}, \underline{\ell}_k), \underline{\ell}_k; \underline{g}_k)$.

3.2.4 Interpretation of $\eta_k$ and $Y_k(\underline{g}_k)$ in Example 4.

Under the consistency, conditional independence and positivity assumptions made in this example, $\eta_k(\underline{\pi}_k, \underline{\ell}_k; \underline{g}_k)$ coincides with the $E \{ L_{K+1,h^*|\underline{\pi}_k = \underline{\pi}_k, \underline{\ell}_k = \underline{l}_k} \}$, i.e., the mean of the counterfactual outcome $L_{K+1,h^*}$ among subjects that received, in the real world, treatment $\underline{\pi}_k$ up to cycle $k$ and have recorded past outcomes $\underline{\ell}_k = \underline{l}_k$. Unlike the preceding examples, if $h^*$ assigns positive probability to all possible treatment values $a_k$, then $\theta(g)$ depends on the values $\eta_k(\underline{\pi}_k, \underline{\ell}_k; \underline{g}_k)$ for all $(\underline{\pi}_k, \underline{\ell}_k)$. This is because, unlike the preceding examples, here $\pi^*k > 0$ w.p.1. It follows from definition (6) that $\pi^{*k} \equiv \pi^{*1}_k$ is the product of the conditional probabilities given past $L$’s and treatments, the subject receives the treatments $A_j, j = 1, \ldots, k$, that he/she actually received when he/she follows regime $h^*$. Also, $Y_k(\underline{g}_k) \equiv E_{h^*_k+1} \{ \eta_{k+1}(\underline{A}_{k+1}, \underline{\ell}_{k+1}; \underline{g}_{k+1}) | \underline{A}_k, \underline{\ell}_{k+1} \} = E \{ L_{K+1,h^*|\underline{A}_{k-1} = \underline{a}_{k-1}, \underline{\ell}_k = \underline{l}_k} \}$, i.e. the mean of the counterfactual outcome $L_{K+1,h^*}$ among subjects that received, in the real world, treatment $\underline{A}_{k-1} = \underline{\pi}_{k-1} \neq \underline{\ell}_k$ up to cycle $k - 1$ and have recorded past outcomes $\underline{\ell}_k = \underline{l}_k$ up to an including cycle $k$. 

19
4 Estimation based on parametric models for the nuisance functions

4.1 Inverse probability weighting estimation

Uniform consistent estimation of $\theta(g)$ under the large non-parametric model $P$ cannot be carried out due to the curse of dimensionality. Both in theory and in practice one is forced to consider a dimension reducing plan. One such plan is motivated from display (7). Specifically, suppose that for each $k \in [K]$ we postulate a smooth parametric class for $h_k$, say,

$$C_k = \{h_k, \alpha_k \in \mathcal{V}_k : \alpha_k \in \Xi_k\}$$  \hspace{1cm} (10)

where $\Xi_k$ is a subset of a Euclidean space and $\mathcal{V}_k$ is the set of all conditional densities $p_{A_k|L_k, A_k^{-1}}$ for probability measures in $P$. We can then compute the estimator

$$\hat{\theta}_{IPW} \equiv \mathbb{P}_n \left\{ \psi \left( T_{K+1} \right) \frac{\pi^* K}{\hat{\pi} K} \right\}$$

where throughout $\hat{\pi}^j = \prod_{r=1}^{k} \hat{h}_r, \hat{\alpha}_{ML}, \hat{\pi}^k = \hat{\pi}_1, \hat{\alpha}_{ML} = (\hat{\alpha}_{1,ML}, \ldots, \hat{\alpha}_{K,ML}),$

$$\hat{\alpha}_{k,ML} = \arg \max_{\alpha_k \in \Xi_k} \mathbb{P}_n \{ \log h_{k,\alpha_k} \}$$ and $\mathbb{P}_n(\cdot)$ is the empirical mean operator, i.e.

$$\mathbb{P}_n(V) = n^{-1} \sum_{i=1}^{n} V_i.$$ Under regularity conditions, $\hat{\theta}_{IPW}$ is consistent and asymptotically normal, throughout CAN, i.e.

$$\sqrt{n} \left\{ \hat{\theta}_{IPW} - \theta(g) \right\}$$ converges to a mean zero normal random variable provided $p$ is in the submodel $\cap_{k=1}^{K} \mathcal{H}_k$ of $P$ where

$$\mathcal{H}_k \equiv \{p \in P : h_k \in C_k\}$$  \hspace{1cm} (11)

4.2 Fully parametric maximum likelihood estimation

Suppose that we postulate parametric models for each $g_k$, say $\{g_k, \xi_k \in \mathcal{U}_k : \xi_k \in \mathcal{F}_k\}$ where $\mathcal{F}_k$ is some Euclidean space, $k = 1, \ldots, K$ and compute the maximum likelihood estimator $\hat{\theta}(\hat{g}_{ML})$ of $\theta(g)$ where

$$\hat{g}_{ML} \equiv (g_{0,n}, g_{\xi_{ML}}), g_{\xi} \equiv (g_{1,\xi_1}, \ldots, g_{K,\xi_K}), \hat{\xi}_{ML} = (\hat{\xi}_{1,ML}, \ldots, \hat{\xi}_{K,ML}),$$

$$\hat{\xi}_{k,ML} = \arg \max_{\xi_k \in \mathcal{F}_k} \mathbb{P}_n \{ \log g_{k,\alpha_k} \}$$ and $g_{0,n}$ is the empirical law of $L_1$. The plug-in estimator $\hat{\theta}(\hat{g}_{ML})$ is CAN for $\theta(g)$ if the postulated parametric models are correct.
4.3 Iterated regression estimation

One can construct estimators of $\theta (g)$ that are CAN under semiparametric, rather than parametric models for $g$. The representation (9) of $\theta (g)$ and the recursion to arrive at $Y_1 (g_i)$ motivates a dimension reducing plan in which estimation of $\theta (g)$ is conducted assuming that, for each each $k \in [K]$, the map

$$ (\pi_k, \ell_k) \in \text{Posit}_k \rightarrow \eta_k \left( \pi_k, \ell_k; g_k \right), \tag{12} $$

with domain the set

$$ \text{Posit}_k \equiv \left\{ (\pi_{k-1}, \ell_k) : h_j^* (a_j | \pi_{j-1}, \ell_j) > 0, j = 1, \ldots, k-1 \right\}, $$

of possible histories $(\pi_{k-1}, \ell_k)$ under $h^*$, belongs to the parametric class

$$ \mathcal{R}_k = \left\{ \eta_{k, \tau_k} \in \mathcal{D}_k : \eta_{k, \tau_k} (\pi_k, \ell_k) = \Psi \left\{ \tau_k^T s_k (\pi_k, \ell_k) \right\} : \tau_k \in \mathcal{Y}_k \right\}, \tag{13} $$

where $\mathcal{D}_k$ is the set of all real valued functions with domain in $\text{Posit}_k$, $\Psi$ is a canonical link in a generalized linear model, $s_k$ is a known function and $\tau_k$ an unknown parameter, with $\mathcal{Y}_k$ a subset of a Euclidean space. Define

$$ \mathcal{G}_k \equiv \left\{ p \in \mathcal{P} : \eta_k \left( \cdot ; g_k \right) \in \mathcal{R}_k \right\}, k \in [K], \tag{14} $$

and the estimator

$$ \hat{\theta}_G \equiv \mathbb{P}_n \left( \tilde{Y}_1 \right), $$

where $\tilde{Y}_1$ is the output of the following recursive algorithm.

**Algorithm 1.** Set $\tilde{Y}_{K+1} \equiv \psi (\tilde{T}_{K+1})$ and recursively, for $k = K, K - 1, \ldots, 1$,

a) Estimate $\tau_k$ indexing the regression model

$$ \eta_{k, \tau_k} (\bar{A}_k, \bar{L}_k) \equiv \Psi \left\{ \tau_k^T s_k (\bar{A}_k, \bar{L}_k) \right\}, $$

for $E \left( \tilde{Y}_{k+1} | \bar{A}_k, \bar{L}_k \right)$ restricted to units verifying $\pi^{*k} > 0$ with $\hat{\tau}_{k,G}$ solving

$$ \mathbb{P}_n \left[ \pi^{*k} s_k (\bar{A}_k, \bar{L}_k) \left\{ \tilde{Y}_{k+1} - \Psi \left\{ \tau_k^T s_k (\bar{A}_k, \bar{L}_k) \right\} \right\} \right] = 0. \tag{15} $$
b) For units with \( \pi^{*k-1} > 0 \), compute

\[
\tilde{Y}_k \equiv y_{k,\tilde{\tau}_{k,\theta}}(\overline{A}_{k-1}, \overline{L}_k) \equiv \int h_k^*(a_k|\overline{A}_{k-1}, \overline{L}_k) \eta_{k,\tilde{\tau}_{k,\theta}}(a_k, \overline{A}_{k-1}, \overline{L}_k) \, d\mu_k(a_k).
\]

Note that if, as in Examples 1-3, for each \( k \in [K] \), \( h_k^* \) is an indicator function, then \( \pi^{*k} \) is also an indicator function. In such case, the factor \( \pi^{*k} \) ensures that estimation of \( \tau_k \) is based only on subjects with \( \pi^{*k} = 1 \). In Examples 1-3, subjects with \( \pi^{*k} = 1 \) are those that remain compliers at cycle \( k \). The estimator \( \hat{\tau}_{k,\theta} \) coincides with the estimator obtained from fitting, by iteratively reweighted least squares (IRLS), the regression model \( \Psi \{\tau^T s_k(\overline{A}_k, \overline{L}_k)\} \) restricted to those subjects. Note also that when, as in Examples 1-3, \( h_k^* \) is an indicator function, the integral in step (b) of the algorithm is equal to the function \( \eta_{k,\tilde{\tau}_{k,\theta}}(a_k, \overline{A}_{k-1}, \overline{L}_k) \) evaluated at the value of \( a_k \) for which \( h_k^*(a_k|\overline{A}_{k-1}, \overline{L}_k) = 1 \). Thus, for instance, in Examples 1 and 3, \( y_{k,\tilde{\tau}_{k,\theta}}(\overline{A}_{k-1}, \overline{L}_k) \) is equal to \( \eta_{k,\tilde{\tau}_{k,\theta}}(1, \overline{A}_{k-1}, \overline{L}_k) \) and \( \eta_{k,\tilde{\tau}_{k,\theta}}(d_k(\overline{A}_{k-1}, \overline{L}_k), \overline{A}_{k-1}, \overline{L}_k) \) respectively.

Whether or not \( \pi^{*k} \) is binary, we note that the equation (15) will have a unique solution when \( \psi(\overline{L}_{K+1}) \) falls in the range of \( \Psi(\cdot) \). For \( k = K \), the equation solved by the estimator \( \hat{\tau}_{K,\theta} \) agrees with the score equation from the fit of a generalized linear model with canonical link except that each individual contribution is weighted \( \pi^{*k} \) and should therefore be the maximizer of the weighted log-likelihood for the associated exponential family model with outcome \( \psi(\overline{L}_{K+1}) \). For \( k < K \), the estimating equation (15) is again a weighted score equation, under the same generalized linear model with the same canonical link function, but for the pseudo-outcome \( \tilde{Y}_{k+1} \). This pseudo-outcome falls in the range of \( \Psi(\cdot) \) because, by construction, it agrees with the conditional mean of \( \Psi(\tilde{\tau}_{k+1,\theta} s_{k+1}(\overline{A}_{k+1}, \overline{L}_{k+1})) \) given \( (\overline{A}_k, \overline{L}_{k+1}) \) under \( h_k^* \). Thus, for \( k < K \), the equation (15) has a unique solution.

We note that when one specifies parametric models \( R_k \) for \( \eta_{k,\theta}^2 \) there is the possibility that the resulting models \( G_k \) are incompatible. We do not discuss this issue in this paper. Molina et. al. (2017) give a careful discussion of the topic and Babino et. al. (2017) propose a modeling strategy which avoids model incompatibility.

To analyze the asymptotic behavior of \( \hat{\theta}_G \) and of several of the forthcoming
estimators, we define for any $\eta_k (A_k, L_k) , k \in [K] ,$

$$y_{k, \eta_k} (A_{k-1}, L_k) \equiv E_{h_k^*} \{ \eta_k (A_k, L_k) | A_{k-1}, L_k \}$$

$$= \int \eta_k (a_k, A_{k-1}, L_k) h_k^* (a_k | A_{k-1}, L_k) d\mu_k (a_k) ,$$

and

$$\Delta_k (\eta_k, \eta_{k+1}; g_k) \equiv \pi^* k \left[ \eta_k (A_k, L_k) - E_{g_k} \left\{ \eta_{k+1} (A_{k+1}, L_{k+1}) | A_k, L_k \right\} \right]$$

$$= \pi^* k \left[ \eta_k (A_k, L_k) - E_{g_k} \left\{ y_{k+1, \eta_{k+1}} (A_{k+1}, L_{k+1}) | A_k, L_k \right\} \right]$$

with $y_{K+1, \eta_{K+1}} (A_{K+1}, L_{K+1}) \equiv \psi (L_{K+1}) . \text{ Note that }$

$$\Delta_k (\eta_k, \eta_{k+1}; g_k) = 0 \text{ if } \eta_j = \eta_j^g \text{ for } j = k, k + 1. \quad (16)$$

where here, and sometimes in what follows, we write, for short, $\eta_j^g (\cdot, \cdot)$ instead of $\eta_j (\cdot, \cdot; g_j) . \text{ We further define for any } \eta = (\eta_1, \ldots, \eta_K) \text{ and any } p = gh,$

$$d^g (\eta) \equiv \sum_{k=1}^K E_{y_{k-1}, \bar{h}_k} \left\{ \frac{1}{\pi^* k} \Delta_k (\eta_k, \eta_{k+1}; g_k) \right\}$$

where, recall $\pi^* k \equiv \prod_{j=1}^k h_j (A_j | A_{j-1}, L_j) . \text{ Note that } d^g (\eta) \text{ does not depend on } h$ because each expectation $E_{y_{k-1}, \bar{h}_k} \left\{ \frac{1}{\pi^* k} \Delta_k (\eta_k, \eta_{k+1}; g_k) \right\}$ is not a function of $\bar{h}_k . \text{ In the Appendix we show the following result.}$

**Lemma 1:** For any $\eta_k (A_k, L_k) , k \in [K] , \text{ it holds that }$

$$E_{g_1} \left\{ y_{1, \eta_1} (L_1) \right\} - \theta (g) = d^g (\eta)$$

To facilitate the analysis of the limiting distribution of $\hat{\theta}_g \text{ we make the following notational conventions and definitions. For any function } t (Z; h, \eta) \text{ which depends on some or all the components of } h = (h_1, \ldots, h_K) \text{ and } \eta = (\eta_1, \ldots, \eta_K) , \text{ and any }$
data dependent functions \( \hat{h} \) and \( \hat{\eta} \), \( \hat{E}_{gh}\{ t (Z; \hat{h}, \hat{\eta}) \} \) stands for expectation under \( p = gh \) regarding \( \hat{h} \) and \( \hat{\eta} \) as non-random functions, that is
\[
\hat{E}_{gh}\{ t (Z; \hat{h}, \hat{\eta}) \} \equiv \int t (z; \hat{h}, \hat{\eta}) p (z) d\mu (z)
\]
With this definition
\[
d^g (\hat{\eta}) = \sum_{k=1}^{K} \hat{E}_{g_{k-1}, \hat{\eta}_k}\{ \frac{1}{\pi_k} \Delta_k (\hat{\eta}_k, \hat{\eta}_{k+1}; g_k) \}
\]
We are now ready to study the limiting behavior of \( \hat{\theta}_G \). Letting \( \hat{\eta}_G \equiv (\hat{\eta}_1, \ldots, \hat{\eta}_K) \), Lemma 1 immediately implies the following representation for \( \hat{\theta}_G \).
\[
\hat{\theta}_G - \theta (g) = \mathbb{P}_n \{ y_{1, \hat{\eta}_1, G} (L_1) \} - \hat{E}_{g_1}\{ y_{1, \hat{\eta}_1, G} (L_1) \} + d^g (\hat{\eta}_G).
\] (17)
To analyze the limiting distribution of \( \hat{\theta}_G \) we first note that the vector \( \hat{\tau}_G \equiv (\hat{\tau}_1, \ldots, \hat{\tau}_K) \) solves a joint system of estimating equations, so under regularity conditions, it has a probability limit under any \( p \in \mathcal{P} \) which we denote with \( \tau_{\lim, G} (p) \equiv (\tau_{1, \lim, G} (p), \ldots, \tau_{K, \lim, G} (p)) \). Furthermore, \( \{ \hat{\tau}_G - \tau_{\lim, G} (p) \} \) is asymptotically linear. In addition, under regularity conditions, the map \( \tau \to d^g (\eta_\tau) \) is differentiable. Then, writing \( \eta_{k, \lim, G} (p) \equiv \eta_{k, \tau_{\lim, G}(p)}, k \in [K] \), we conclude that
\[
d^g (\hat{\eta}_G) - d^g [\eta_{\lim, G} (p)] \text{ is asymptotically linear.}
\]
Furthermore, under regularity conditions, \( y_{1, \hat{\eta}_1, G} \) and \( y_{1, \eta_{1, \lim, G}} \) fall in a Donsker class, so
\[
\mathbb{P}_n \{ y_{1, \hat{\eta}_1, G} (L_1) \} - \hat{E}_{g_1}\{ y_{1, \hat{\eta}_1, G} (L_1) \} = \mathbb{P}_n \{ y_{1, \eta_{1, \lim, G}} (L_1) \} - \hat{E}_{g_1}\{ y_{1, \eta_{1, \lim, G}} (L_1) \} + o_p (n^{-1/2})
\]
is asymptotically linear. The representation (17) then implies that
\[
\hat{\theta}_G - \theta (g) - d^g [\eta_{\lim, G} (p)] \text{ is asymptotically linear.}
\]
To establish that $\hat{\theta}_G$ is CAN under model $\cap_{k=1}^{K} G_k$ it then suffices to show that
\[ d^g \left[ \eta_{\lim,G} (p) \right] = 0 \text{ if } p \in \cap_{k=1}^{K} G_k \] (18)

This fact is a consequence of the following result.

**Proposition 1:** Under regularity conditions,
\[ \eta_{k,\lim,G} (p) = \eta_k^g \text{ if } p \in \cap_{j=k}^{K} G_j \] (19)

Proposition 1 and (16) now imply that for all $k \in [K],$
\[ \Delta_k \left( \eta_{k,\lim,G} (p), \eta_{k+1,\lim,G} (p); g_k \right) = 0 \text{ if } p \in \cap_{j=1}^{K} G_j \]

and therefore that (18) holds.

**Proof of Proposition 1:** By reverse induction in $k$. Suppose first that $k = K$. Assume $p = gh \in G_K$. Then, $E_{g_k} \left\{ \psi (T_{k+1}) \mid A_{k}, L_{k} \right\} = \eta_{K,\tau_K (g_k)} (A_{k}, L_{k})$ for some $\tau_K (g_k)$ and therefore the equation (15) is an unbiased estimating equation for $\tau_K (g_k)$ since $\hat{\tau}_{K+1} = \psi (T_{K+1})$. Consequently, under standard regularity conditions for $M$-estimators, the probability limit $\tau_{K,\lim,G}$ of $\hat{\tau}_{K,G}$ is equal to $\tau_K (g_k)$ which, in turn, implies that (19) holds for $k = K$.

Suppose next that (19) holds for $k = K, \ldots, j + 1$. Noticing that, by construction, $\hat{Y}_{j+1} = y_{j+1,\tilde{\eta}_{j+1,G}} (A_{k}, L_{k+1})$, we conclude that $\hat{\tau}_{j,G}$ solves
\[ 0 = \mathbb{P}_n \left[ \pi^j s_j (A_{j}, L_{j}) \left\{ y_{j+1,\eta_{j+1,G}} (A_{k}, L_{k+1}) - \Psi \left\{ \tau_{j}^T s_j (A_{j}, L_{j}) \right\} \right\} \right] + o_p (1) \] (1)

Suppose $p = gh \in \cap_{j=1}^{K} G_k$. Then, by the inductive hypothesis
\[ y_{j+1,\eta_{j+1,G}} (A_{j}, L_{j+1}) = Y_{j+1} (\hat{g}_{j+1}) \]. Thus,
\[ E_{g_j} \left\{ y_{j+1,\eta_{j+1,G}} (A_{j}, L_{j+1}) \mid A_{j}, L_{j} \right\} = \eta_j^g (A_{j}, L_{j}) \]. Furthermore, since $p \in G_j$ then $\eta_j^g = \eta_j (g_j)$ and therefore the population equation
\[ E_{\gamma_{j}, \pi_j} \left[ \pi^j s_j (A_{j}, L_{j}) \left\{ y_{k+1,\eta_{k+1,G}} (A_{k}, L_{k+1}) - \Psi \left\{ \tau_{j}^T s_j (A_{j}, L_{j}) \right\} \right\} \right] = 0 \] (4.4)
is solved at $\tau_j = \tau_j (g_j)$. Then, under regularity conditions for the consistency of $M$-estimators, the probability limit $\tau_{j,\lim,G}$ of $\hat{\tau}_{j,G}$ is equal to $\tau_j (g_j)$, which shows (19) holds for $k = j$. 

25
4.4 Weighted iterated regression

Suppose that in Algorithm 1 we replace step (a) with a procedure that estimates, $\tau_k$ by weighted IRLS, i.e. with $\widehat{\tau}_{k,\omega}$ solving

$$\mathbb{P}_n \left[ \pi^k \omega_k(\bar{A}_k, \bar{L}_k) s_k(\bar{A}_k, \bar{L}_k) \left\{ \widehat{Y}_{k+1,\omega} - \Psi \left\{ \tau_k^T s_k(\bar{A}_k, \bar{L}_k) \right\} \right\} = 0 \right]$$

for some user specified scalar function $\omega_k(A_k, L_k)$, and where for each $k$, $\widehat{Y}_{k,\omega}$ is defined as $\widehat{Y}_k$ in step (b) of Algorithm 1 but with $\widehat{\tau}_{k,\omega}$ instead of $\widehat{\tau}_k$. The resulting estimator $\widehat{\theta}_{\omega}$ is also CAN for $\theta(g)$ under regularity conditions if $p \in \cap_{j=1}^K G_j$. In fact, the same holds even if $\omega_k(A_k, L_k) = \omega_k(A_k, L_k)_{ML}$ depends on the maximum likelihood estimator $\widehat{\alpha}_{ML}$ of $\alpha$ defined in section 4.1. Specifically, to analyze the limiting distribution of $\widehat{\theta}_{\omega}$ where we allow the possibility that $\omega_k = \omega_k(A_k, L_k)_{ML}$, note that regardless of the validity of any of the models $H_k$ or $G_k$, $(\widehat{\tau}_{\omega}, \widehat{\alpha}_{ML})$ is ultimately an $M$-estimator and as such, under regularity conditions, it has a probability limit $(\tau_{lim,\omega}(p), \alpha_{lim}(h))$. Furthermore, $\{\widehat{\tau}_{\omega} - \tau_{lim,\omega}(p)\}$ is asymptotically linear. Then, with $\eta_{k,lim,\omega}(p) = \eta_{k,\tau_{lim,\omega}(p)}$, $k \in [K]$, we reason as in the preceding section and conclude that

$$\widehat{\theta}_{\omega} - \theta(g) - d^g[\eta_{lim,\omega}(p)]$$

is asymptotically linear.

An argument essentially identical to that given for the proof of Proposition 1 shows that, under regularity conditions

$$\eta_{k,lim,\omega}(p) = \eta_k^g$$

if $p \in \cap_{j=k}^K G_j$

and consequently, that $d^g[\eta_{lim,\omega}(p)] = 0$ and thus, that $\widehat{\theta}_{\omega}$ is CAN for $\theta(g)$, if $p \in \cap_{k=1}^K G_k$.

We will argue in sections (4.6.3) and (4.7.3) that a particular choice of weights $\omega_k$, namely, $\omega_k = 1/\widehat{\pi}^k$ where

$$\widehat{\pi}^k \equiv \pi^k(\bar{A}_k, \bar{L}_k) \equiv \prod_{j=1}^k h_{j,\widehat{\alpha}_{j,ML}}(A_j | A_{j-1}, L_j),$$

yields estimators of $\theta$ that are CAN under a model larger than $\cap_{j=k}^K G_j$.

For ease of reference, we denote the estimators $\widehat{\tau}_{k,\omega}$ and $\widehat{\theta}_{\omega}$ using $\omega_k = 1/\widehat{\pi}^k$ as $\widehat{\tau}_{k,reg}$ and $\widehat{\theta}_{reg}$, and the pseudo outcome $\widehat{Y}_{k+1,\omega}$ in equation (20) as $\widehat{Y}_{k+1,reg}$.
4.5 Doubly robust estimation by iterated regression

When \( A_k \) is binary and \( h^*_k(a_k|\bar{t}_k, \bar{a}_{k-1}) = a_k \) as in Example 1, Bang and Robins (2005) (throughout B&R) proposed another iterated regression algorithm which, they argued, remarkably returns a so-called doubly robust estimator of \( \theta(g) \) in the union model \((\bigcap_{k=1}^{K} H_k) \cup (\bigcap_{k=1}^{K} G_k)\). This is an estimator that is CAN when \( p \) is in \((\bigcap_{k=1}^{K} H_k) \cup (\bigcap_{k=1}^{K} G_k)\), equivalently the estimator is CAN when either the models for all the \( h_k \) are correct, or the models for all the \( \eta_k \) are correct, but not necessarily both. Here we generalize the construction of the B&R estimator to arbitrary conditional densities \( h^*_k(a_k|\bar{t}_k, \bar{a}_{k-1}) \). The construction starts with the computation of the maximum likelihood estimator \( \hat{\alpha}_{ML} \) as above. Next, one considers the extended parametric class

\[
R^\text{ext}_k = \left\{ \eta_{k,v_k} \in \mathcal{D}_k : \eta_{k,v_k}(\bar{t}_k, \bar{l}_k) = \Psi \left\{ \tau_k^T s_k(\bar{u}_k, \bar{l}_k) + \lambda_k \pi^k(\bar{u}_k, \bar{l}_k)^{-1} \right\}, v_k \equiv (\tau_k, \lambda_k) \in \mathcal{Y}_k \times \mathbb{R} \right\}
\]

and subsequently applies Algorithm 1 to the extended model \( R^\text{ext}_k \). Specifically,

**Algorithm 2 (Robins, 2002, Bang and Robins, 2005)**

Set

\[
\tilde{Y}_{K+1} \equiv \psi(\bar{L}_{K+1})
\]

and recursively, for \( k = K, K-1, \ldots, 1 \),

a) Estimate \( v_k \equiv (\tau_k, \lambda_k) \) indexing the regression model

\[
\eta_{k,v_k}(\bar{A}_k, \bar{L}_k) \equiv \Psi \left\{ \tau_k^T s_k(\bar{u}_k, \bar{l}_k) + \lambda_k \pi^k(\bar{u}_k, \bar{l}_k)^{-1} \right\}
\]

for \( E(\tilde{Y}_{k+1}|\bar{A}_k, \bar{L}_k) \) restricted to units verifying \( \pi^k > 0 \) with \( \tilde{v}_k \equiv (\tilde{\tau}_k, \tilde{\lambda}_k) \) solving

\[
P_n \left[ \pi^k \left[ \frac{s_k(\bar{A}_k, \bar{L}_k)}{1/\pi^k} \right] \left[ \tilde{Y}_{k+1} - \psi \left\{ \tau_k^T s_k(\bar{A}_k, \bar{L}_k) + \lambda_k \left( 1/\pi^k \right) \right\} \right] = 0 \tag{22} \right]
\]

b) For units with \( \pi^{k-1} > 0 \), compute

\[
\tilde{Y}_k \equiv y_{k,\tilde{v}_k}(\bar{A}_{k-1}, \bar{L}_k) = \int h^*_k(a_k|\bar{A}_{k-1}, \bar{L}_k) \eta_{k,\tilde{v}_k}(a_k, \bar{A}_{k-1}, \bar{L}_k) \, d\mu_k(a_k).
\]

Finally, estimate \( \theta(g) \) with \( \hat{\theta}_{Bang} = P_n \left( \tilde{Y}_1 \right) \).
To analyze the limit distribution of $\hat{\theta}_{Bang}$ and argue that it is CAN under model $(\cap_{k=1}^{K} H_k) \cup (\cap_{k=1}^{K} G_k)$ we define $\tilde{\eta}_k \equiv \eta_k, \tilde{v}_k$ for $k \in [K]$ and $\tilde{\eta} \equiv (\tilde{\eta}_1, \ldots, \tilde{\eta}_K)$. Invoking Lemma 1 we obtain

\[
\hat{\theta}_{Bang} - \theta (g) \equiv \mathbb{P}_n \{ y_1, \tilde{\eta}_1 (L_1) \} - \mathbb{E}_g \{ y_1, \tilde{\eta}_1 (L_1) \} + d^g (\tilde{\eta}) \quad (23)
\]

As in our analysis of the distribution of $\hat{\theta}_g$, to analyze the limiting distribution of $\hat{\theta}_{Bang}$ we start by noting that the vectors $\tilde{v} \equiv (\tilde{v}_1, \ldots, \tilde{v}_K)$ and $\tilde{\alpha}_{ML}$ ultimately solve a joint system of estimating equations, so under regularity conditions, $\tilde{v}$ has a probability limit $v_{\lim} (p) \equiv (v_{1,\lim} (p), \ldots, v_{K,\lim} (p))$ under any $p \in \mathcal{P}$. Furthermore, $\{ \tilde{v} - v_{\lim} (p) \}$ is asymptotically linear. Then, under regularity conditions that imply the differentiability of the path $v \to d^g (\eta_v)$ where $\eta_v \equiv (\eta_{1,v}, \ldots, \eta_{K,v})$, letting $\eta_{k,\lim,Bang} (p) \equiv \eta_{k,v_k,\lim}(p)$, we have that

\[
d^g (\eta_v) - d^g [\eta_{\lim,Bang} (p)] \text{ is asymptotically linear.}
\]

Furthermore, under regularity conditions, if $y_1, \tilde{\eta}_1$ and $y_1, \eta_{\lim,Bang}(p)$ fall in a Donsker class, then

\[
\mathbb{P}_n \{ y_1, \tilde{\eta}_1 (L_1) \} - \mathbb{E}_g \{ y_1, \tilde{\eta}_1 (L_1) \} = \mathbb{P}_n \{ y_1, \eta_{\lim,Bang}(p) (L_1) \} - \mathbb{E}_g \{ y_1, \eta_{\lim,Bang}(p) (L_1) \} + o_p (n^{-1/2})
\]

is asymptotically linear. So, from expansion (23), we conclude that

\[
\hat{\theta}_{Bang} - \theta (g) - d^g \left[ \eta_{\lim,Bang} (p) \right] \text{ is asymptotically linear.}
\]

Now, because the limit values $v_{\lim} (p)$ and $\alpha_{\lim} (h) \equiv (\alpha_{1,\lim} (h_1), \ldots, \alpha_{K,\lim} (h_K))$ of $\tilde{v}$ and $\tilde{\alpha}_{ML}$ satisfy the population version of (22) (i.e. with $\mathbb{P}_n$ replaced by $E_{gh}$ and all the estimators replaced by their probability limits), then in particular, the second row of equation (22) implies that

\[
\mathbb{E}_{\pi^{k}_{h_k} \tilde{\eta}_k} \left\{ \frac{1}{\pi^{k}_{h_k}} \Delta_k \left( \eta_{k,\lim,Bang} (p), \eta_{k+1,\lim,Bang} (p) ; g_k \right) \right\} = 0, \quad (24)
\]

where $\pi^{k}_{h_k} (\tilde{\eta}_k) \equiv \prod_{j=1}^{k} h_{j,\alpha_{j,\lim}(h_j)} (A_j | A_{j-1}, \tilde{T}_j)$. Then, with $h_{\lim} (h) \equiv (h_{1,\alpha_{1,\lim}(h_1)}, \ldots, h_{K,\alpha_{K,\lim}(h_K)})$, $d^g \left[ \eta_{\lim,Bang} (p) \right] = c^p \left[ h_{\lim} (h) , \eta_{\lim,Bang} (p) \right]$, \quad (25)
where for any $h^\dagger = (h^\dagger_1, \ldots, h^\dagger_K)$ and $\eta^\dagger = (\eta^\dagger_1, \ldots, \eta^\dagger_K)$,

$$c^p (h^\dagger, \eta^\dagger) \equiv \sum_{k=1}^{K} E_{g_{k-1}, \eta_k} \left\{ \left\{ \frac{1}{\pi_k} - \frac{1}{\pi^\dagger_k} \right\} \Delta_k (\eta^\dagger_k, \eta^\dagger_{k+1}; g_k) \right\}$$

with $\pi^\dagger_k \equiv \prod_{j=1}^{k} h^\dagger_j (A_j | \overline{A}_{j-1}, \overline{L}_j)$. Note that, unlike $d^g (\eta^\dagger)$, $c^p (h^\dagger, \eta^\dagger)$ depends on $p = gh$ not only through $g$ but also through $h$.

From (25) we conclude that $\hat{\theta}_{Bang}$ is CAN under $(\cap_{k=1}^{K} \mathcal{H}_k) \cup (\cap_{k=1}^{K} \mathcal{G}_k)$ provided $c^p [h_{\lim} (h), \eta_{\lim, Bang} (p)] = 0$ for $p \in \cap_{k=1}^{K} \mathcal{H}_k$ and for $p \in \cap_{k=1}^{K} \mathcal{G}_k$.

That $c^p [h_{\lim} (h), \eta_{\lim, Bang} (p)] = 0$ for $p \in \cap_{k=1}^{K} \mathcal{H}_k$ follows immediately after recognizing that, under regularity conditions, the MLE $\hat{\alpha}_{j, ML}$ is consistent so when $p \in \cap_{k=1}^{K} \mathcal{H}_k$, $h_{j, \alpha_{j, lim}(h_j)} = h_j$ for all $j$.

On the other hand,

$$\eta_{k, \lim, Bang} (p) = \eta^\dagger_k \text{ if } p \in \cap_{j=1}^{K} \mathcal{G}_j$$  \hspace{1cm} (26)

This result follows essentially along the same lines of the proof of (19), upon noticing that when $p \in \mathcal{G}_k$ then $p$ also belongs to $\mathcal{G}^{ext}_k$ where $\mathcal{G}^{ext}_k$ is defined like $\mathcal{G}_k$ but with $\mathcal{R}^{ext}_k$ instead of $\mathcal{R}_k$.

We therefore conclude from (16) and (26) that if $p \in \cap_{j=1}^{K} \mathcal{G}_j$, $\Delta_k (\eta_{k, \lim} (p), \eta_{k+1, \lim} (p); g_k) = 0$ for all $k \in [K]$, and therefore $d^g [\eta_{\lim, Bang} (p)] = 0$.

4.6 $K+1$ - multiply robust estimation

4.6.1 The Bang and Robins estimator is $K+1$ - multiply robust

Surprisingly, a closer examination at the analysis of the asymptotic properties of the Bang and Robins estimator in the preceding subsection reveals, as we will argue next, that

$$c^p [h_{\lim} (h), \eta_{\lim, Bang} (p)] = 0 \text{ if } p = gh \in \cup_{j=1}^{K+1} [(\cap_{k=1}^{j-1} \mathcal{H}_k) \cap (\cap_{k=j}^{K} \mathcal{G}_k)]$$  \hspace{1cm} (27)

where $\cap_{k=1}^{0} \mathcal{H}_k \equiv \cap_{k=K+1}^{K} \mathcal{G}_k \equiv \mathcal{P}$. The assertion in (27) implies that under regularity conditions, $\hat{\theta}_{Bang}$ is CAN not just under model $(\cap_{k=1}^{K} \mathcal{H}_k) \cup (\cap_{k=1}^{K} \mathcal{G}_k)$ but also under...
the larger model $\cup_{j=1}^{K+1} \left[ \left( \cap_{k=1}^{j-1} H_k \right) \cap \left( \cap_{k=j}^{K} G_k \right) \right]$. This fact, that went unnoticed in B&R, is a special case of a general result on doubly robust estimation in factorized likelihood models discussed in Molina et al. (2017). Thus $\hat{\theta}_{\text{Bang}}$ confers even more robustness to model misspecification than that claimed in B&R, for it is CAN for $\theta (g)$ not only when one of the following occurs, (i) the models for all the $h_k$ are correct, or (ii) the models for all the $\eta_k$ are correct, but also when (iii) for some $j \in [K - 1]$ the models for $h_k, 1 \leq k \leq j$ and the models for $\eta_k, j + 1 \leq k \leq K$ are all correct. We designate an estimator that is CAN whenever (i), (ii) or (iii) holds, a $(K + 1)$ - multiply robust estimator.

To show (27), suppose that for some $j \in [K], p \in \left( \cap_{k=1}^{j-1} H_k \right) \cap \left( \cap_{k=j}^{K} G_k \right)$. Because $p \in \cap_{k=1}^{j-1} H_k$, $\pi_k \lim (T_k) = \pi_k$ for $k = 1, \ldots, j - 1$, so the first $j - 1$ terms in the sum involved in $e^p \left[ h_{\text{lim}} (h), \eta_{\text{lim,Bang}} (p) \right]$ are 0. On the other hand, when $p \in \cap_{k=j}^{K} G_k$, it follows from (26) and (16) that $\Delta_k (\eta_{k, \text{lim}} (p), \eta_{k+1, \text{lim}} (p); g_k) = 0$ for $k = j, \ldots, K$ so the last $K - j + 1$ terms of the summation involved in $e^p \left[ h_{\text{lim}} (h), \eta_{\text{lim,Bang}} (p) \right]$ also vanish.

4.6.2 The greedy iterated fit $K + 1$ - multiply robust estimators

The B&R estimator is not the only $K + 1$ multiply robust estimator in model $\cup_{j=1}^{K+1} \left[ \left( \cap_{k=1}^{j-1} H_k \right) \cap \left( \cap_{k=j}^{K} G_k \right) \right]$. In fact, an examination of the steps followed in the analysis of the preceding subsection reveals that any estimator, say $\hat{\theta}$, of $\theta$ that admits the expansion

$$
\hat{\theta} - \theta (g) = \mathbb{P}_n \left\{ y_{1, \hat{\eta}_1} (L_1) \right\} - \hat{E}_{g_1} \left\{ y_{1, \hat{\eta}_1} (L_1) \right\} + d^g (\hat{\eta})
$$

and verifies

1) $\mathbb{P}_n \left\{ y_{1, \hat{\eta}_1} (L_1) \right\} - \hat{E}_{g_1} \left\{ y_{1, \hat{\eta}_1} (L_1) \right\}$ is asymptotically linear and,

2) $d^g (\hat{\eta}) - e^p \left[ h_{\text{lim}} (h), \eta_{\text{lim}} (p) \right]$ is asymptotically linear for some $(h_{\text{lim}} (h), \eta_{\text{lim}} (p))$ satisfying

i) $\eta_{k, \text{lim}} (p) = \eta_k^g$ when $p \in \cap_{j=k}^{K} G_j$, and

ii) $h_{k, \text{lim}} (h) = h_k$ when $p \in H_k$
will be $K + 1$ multiply robust in model $\bigcup_{j=1}^{K+1} \left( \cap_{j=1}^{j-1} \mathcal{H}_k \right) \cap \left( \cap_{j=K}^{K} \mathcal{G}_k \right)$.

We now describe two estimators which satisfy these conditions. The first is the output of a slight modification of Algorithm 2, whereby the parameters $\tau_k$ and $\lambda_k$ of the extended model $\mathcal{R}_k^{ext}$ are estimated greedily: first $\tau_k$ is estimated under the original model $\mathcal{R}_k$ and next $\lambda_k$ is estimated under $\mathcal{R}_k^{ext}$ but assuming $\tau_k$ is fixed and known and equal to its estimated value. In the book Targeted Learning (2011), van der Laan and Rose, emphasize the utility of such a greedy version of the B&R plug-in estimator, as a greedy fit makes it easy to replace parametric estimators of $\eta_{k,\tau}(\mathcal{A}_k, \mathcal{L}_k)$ by more data adaptive machine learning estimators.

Algorithm 3. (Greedy iterated regression fit). Set $\tilde{Y}_{K+1} \equiv \psi(\mathcal{L}_{K+1})$ and for $k = K, K-1, \ldots, 1$,

a.1) Estimate $\tau_k$ indexing the regression model

$$
\eta_{k,\tau}(\mathcal{A}_k, \mathcal{L}_k) \equiv \Psi \left\{ \tau_k^T s_k(\mathcal{A}_k, \mathcal{L}_k) \right\}
$$

for $E \left( \tilde{Y}_{k+1} | \mathcal{A}_k, \mathcal{L}_k \right)$ restricted to units verifying $\pi^*_k > 0$ with $\tilde{\tau}_k$ solving

$$
\mathbb{P}_n \left[ \pi^*_k s_k(\mathcal{A}_k, \mathcal{L}_k) \left\{ \tilde{Y}_{k+1} - \Psi \left\{ \tau_k^T s_k(\mathcal{A}_k, \mathcal{L}_k) \right\} \right\} = 0
$$

a.2) Based on units with $\pi^*_k > 0$, estimate $\lambda_k$ indexing the regression model for $E \left( \tilde{Y}_{k+1} | \mathcal{A}_k, \mathcal{L}_k \right)$:

$$
\eta_{k,\tilde{\tau},\tilde{\lambda}}(\mathcal{A}_k, \mathcal{L}_k) \equiv \Psi \left\{ \tilde{\tau}_k^T s_k(\mathcal{A}_k, \mathcal{L}_k) + \lambda_k \left( 1/\hat{\pi}_k \right) \right\}
$$

which has offset $\tilde{\tau}_k s_k(\mathcal{A}_k, \mathcal{L}_k)$ with $\lambda_k$ solving

$$
\mathbb{P}_n \left[ \pi^*_k \left( 1/\hat{\pi}_k \right) \left\{ \tilde{Y}_{k+1} - \Psi \left\{ \tilde{\tau}_k^T s_k(\mathcal{A}_k, \mathcal{L}_k) + \lambda_k \left( 1/\hat{\pi}_k \right) \right\} \right\} = 0
$$

b) For units with $\pi^*_{k-1} > 0$, compute

$$
\tilde{Y}_k \equiv y_{k,\tilde{\tau},\tilde{\lambda}}(\mathcal{A}_{k-1}, \mathcal{L}_k) \equiv \int h_k^*(a_k | \mathcal{A}_{k-1}, \mathcal{L}_k) \Psi \left\{ \tilde{\tau}_k^T s_k(a_k, \mathcal{A}_{k-1}, \mathcal{L}_k) + \tilde{\lambda}_k \hat{\pi}_k(a_k, \mathcal{A}_{k-1}, \mathcal{L}_k)^{-1} \right\} d\mu_k(a_k).
$$
Finally, $\hat{\theta}_{\text{greed}} = \mathbb{P}_n \left( \bar{Y}_1 \right)$.

To analyze the limiting behavior of $\hat{\theta}_{\text{greed}}$, we define, for $k \in [K]$,

$$\tilde{\eta}_k (A_k, \tau_k) \equiv \tilde{\eta}_{k, \tau_k} (A_k, \tau_k) \equiv \Psi \left\{ \tilde{\tau}_k s_k (A_k, \tau_k) + \tilde{\lambda}_k \tilde{\pi}^k (A_k, \tau_k) \right\}$$

Then, by Lemma 1,

$$\hat{\theta}_{\text{greed}} - \theta (p) \equiv \mathbb{P}_n \left\{ y_{1, \eta_1} (L_1) \right\} - \tilde{E}_{1 \tau} \left\{ y_{1, \eta_1} (L_1) \right\} + \Delta \hat{\eta}_k (\tilde{\eta})$$  \hspace{1cm} (28)

As in our analyses of the distributions of $\hat{\theta}_{\text{G}}$ and $\hat{\theta}_{\text{Bang}}$, to analyze the limiting distribution of $\hat{\theta}_{\text{greed}}$ we start by noting that the vectors $\tilde{\tau} \equiv \left( \tilde{\tau}_1, \ldots, \tilde{\tau}_K \right)$, $\tilde{\lambda} \equiv \left( \tilde{\lambda}_1, \ldots, \tilde{\lambda}_K \right)$ and $\tilde{\alpha}_{ML}$ solve a joint system of estimating equations, so under regularity conditions, $(\tilde{\tau}, \tilde{\lambda})$ has a probability limit under any $p \in \mathcal{P}$ which we denote with $(\tau, \lambda) \lim (p)$. Letting $\eta_{k, \text{lim}, \text{greed}} (p) \equiv \eta_{k, (\tau, \lambda) \lim} (p)$ we conclude that if the map $(\tau, \lambda) \rightarrow d^q (\eta_{(\tau, \lambda)})$ is differentiable, then

$$d^q (\tilde{\eta}) - d^q [\eta_{k, \text{lim}, \text{greed}} (p)]$$

is asymptotically linear. So, from expansion (28), we conclude that

$$\hat{\theta}_{\text{greed}} - \theta (g) - d^q [\eta_{k, \text{lim}, \text{greed}} (p)]$$

is asymptotically linear. Just as for Algorithm 2, the inclusion of the covariate $1/\tilde{\tau}^k$ in the extended model fitted in step (a.2) of Algorithm 3 implies that

$$d^q \left[ \eta_{k, \text{lim}, \text{greed}} (p) \right] = c_p \left[ \eta_{\text{lim}} (h), \eta_{k, \text{lim}, \text{greed}} (p) \right]$$  \hspace{1cm} (29)

So the $K + 1$ multiply robustness of $\hat{\theta}_{\text{greed}}$ in model $\bigcup_{j=1}^{K+1} \left( (\cap_{k=1}^{j-1} \mathcal{H}_k) \cap (\cap_{k=j}^{K} \mathcal{G}_k) \right)$ follows because

$$\eta_{k, \text{lim}, \text{greed}} (p) = \eta_k^g$$

if $p \in \cap_{j=k}^{K} \mathcal{G}_j$

This result, whose proof we omit, follows essentially along the lines of the proof of 26.
4.6.3 The inverse probability weighted regression $K + 1$ - MR estimators

Here we will argue that the weighted-iterated regression estimators $\hat{\theta}_{\text{reg}}$ defined like the estimator $\hat{\theta}_{\omega}$ of section 4.4 using weights $\omega_k = 1/\hat{\pi}_k$, is also $K + 1$- multiply robust in model $\bigcup_{j=1}^{K+1} \left( \bigcap_{k=1}^{j-1} H_k \cap (\bigcap_{k=j}^{K} G_k) \right)$, provided one of the components of $s_k (A_k, L_k)$ is the constant 1.

Note that, unlike in Algorithms 2 or 3, to compute $\hat{\theta}_{\text{reg}}$ we do not include the covariate $1/\hat{\pi}_k$ in an extended regression model. However, by using weights $\omega_k = 1/\hat{\pi}_k$ in equation (20) and requiring that the vector $s_k (A_k, L_k)$ includes the component 1, we ensure that

$$d^g \left[ \eta_{\lim, \text{reg}} (p) \right] = c^p \left[ \eta_{\lim} (h) , \eta_{\lim, \text{reg}} (p) \right]$$

where $\eta_{\lim, \text{reg}} (p) = \eta_{k, \tau_{\lim, \text{reg}}(p)}$ with $\tau_{k, \lim, \text{reg}} (p)$ the probability limit of $\hat{\tau}_{k, \text{reg}}$.

Furthermore, the same argument as in the proof of (19) shows that $\eta_{k, \lim, \text{reg}} (p) = \eta^g_k$ when $p \in \cap_{j=k}^{K} G_j$. Thus, the requirement (2.i) in the conditions listed at the beginning of section 4.6.2. Since requirement (2.ii) holds as well, we conclude that under regularity conditions, $\hat{\theta}_{\text{reg}}$ is CAN in model $\bigcup_{j=1}^{K+1} \left( \bigcap_{k=1}^{j-1} H_k \cap (\bigcap_{k=j}^{K} G_k) \right)$.

4.7 $2^K$ - multiply robust estimation

4.7.1 Theoretical results background

Remarkably, it is possible to construct estimators of $\theta (g)$ which are CAN under the even larger model $\bigcap_{k=1}^{K} (H_k \cup G_k)$ than $\bigcup_{j=1}^{K+1} \left( \bigcap_{k=1}^{j-1} H_k \cap (\bigcap_{k=j}^{K} G_k) \right)$. Such estimators, which we designate as $2^K$-multiply robust, confer even more protection against model misspecification than $\hat{\theta}_{\text{bang}}, \hat{\theta}_{\text{greed}}, \hat{\theta}_{\text{reg}}$. Their construction is motivated by the following theoretical result in Molina et. al. (2017).

Given $h_k (A_k, L_k)$ and $\eta_k (A_k, L_k), k \in [K]$, define for each $j \in [K]$ the
random variable

\[ Q_j \left( \bar{h}_j^K, \bar{\eta}_j^K \right) = \frac{\pi_j^{*K}}{\pi_j} \psi(\bar{L}_{K+1}) - \sum_{k=1}^{K} \left\{ \frac{\pi_j}{\pi_j^{(k-1)}} \eta_k(\bar{A}_k, \bar{L}_k) - \frac{\pi_j^{(k-1)}}{\pi_j} y_k.\eta_k(\bar{A}_{k-1}, \bar{L}_k) \right\} \]

\[ = y_j.\eta_j(\bar{A}_{j-1}, \bar{L}_{j}) + \sum_{k=1}^{K} \frac{\pi_j^{*K}}{\pi_j} \left\{ y_{k+1}.\eta_{k+1}(\bar{A}_k, \bar{L}_{k+1}) - \eta_k(\bar{A}_k, \bar{L}_k) \right\} \]

\[ = \frac{h_j^*}{h_j} \left( A_j|\bar{A}_{j-1}, \bar{L}_{j} \right) \left\{ Q_{j+1} \left( \bar{h}_{j+1}^K, \bar{\eta}_{j+1}^K \right) - \eta_j(\bar{A}_j, \bar{L}_{j}) \right\} + y_j.\eta_j(\bar{A}_{j-1}, \bar{L}_{j}). \]

where \( y_{K+1}.\eta_K(\bar{A}_K, \bar{L}_{K+1}) \equiv \psi(\bar{L}_{K+1}). \)

Lemma 6 of Molina et al. (2017) implies that if for each \( k \in [K] \) either \( h_k^+ = h_k \) or \( \eta_k^+ = \eta_k^g \), then

\[ \theta(g) = E_{gh} \left[ Q_1 \left( \bar{h}_1^{K}, \bar{\eta}_1^{K} \right) \right] \]

and, if for each \( k \in \{j+1, \ldots, K\} \) either \( h_k^+ = h_k \) or \( \eta_k^+ = \eta_k^g \) then

\[ \eta_j^g(\bar{A}_j, \bar{L}_j) = E_{gh} \left\{ Q_{j+1} \left( \bar{h}_{j+1}^K, \bar{\eta}_{j+1}^K \right) | \bar{A}_j, \bar{L}_j \right\} \]

Now, define for arbitrary \( \eta_j^+ \equiv \eta_j^+(\bar{A}_k, \bar{L}_k), h_k^+ \equiv h_k^+(A_k|\bar{A}_{k-1}, \bar{L}_k) \) and \( p = gh, \)

\[ a^p(h^+, \eta^+) \equiv \sum_{k=1}^{K} E_{gh} \left\{ \frac{\pi_j^{(k-1)}}{\pi_j^{(k-1)}} \left( h_k^* - \frac{h_k^*}{h_k^+} \right) \left( \eta_k^+ - \eta_k^g \right) \right\} \]

and for any unit satisfying \( \pi_j^+ > 0 \) define

\[ a^p(\bar{h}_{j+1}^K, \eta_{j+1}^+, A_j, \bar{L}_j) \equiv \sum_{k=1}^{K} E_{gh} \left\{ \frac{\pi_j^{(k-1)}}{\pi_j^{(k-1)}} \left( h_k^* - \frac{h_k^*}{h_k} \right) \left( \eta_k^+ - \eta_k^g \right) | A_j, \bar{L}_j \right\} \]

where \( \pi_j^{(k-1)} \equiv \prod_{j=1}^{k-1} h_j^+ \left( A_j|\bar{A}_{j-1}, \bar{L}_j \right) \) and \( h_k \equiv h_k \left( A_k|\bar{A}_{k-1}, \bar{L}_k \right). \) In the Appendix we show the following Lemma.

**Lemma 2:** Define \( Q_K \left( \bar{h}_K^{K}, \bar{\eta}_K^{K} \right) \equiv \psi(\bar{L}_{K+1}) \) and \( \sum_{k=K+1}^{K} (\cdot) = 0. \) The following holds:
i) for arbitrary \( \eta^\dagger_k, h^\dagger_k \) and \( p = gh, k \in [K] \),

\[
E_{gh} \left\{ Q_1 \left( \bar{h}^\dagger_1, \bar{\eta}^\dagger_1 \right) \right\} - \theta (g) = a^p \left( h^\dagger, \eta^\dagger \right)
\]  

(34)

ii) for any \( j \in [K] \) and arbitrary \( h^\dagger_k \) and \( \eta^\dagger_k, k \in \{j + 1, \ldots, K\} \), if \( \pi^*j > 0 \) then

\[
E_{L_{j+1}} \left\{ Q_{j+1} \left( \bar{h}^\dagger_{j+1}, \bar{\eta}^\dagger_{j+1} \right)| \bar{A}_j, \bar{L}_j \right\} - \eta^\theta_j \left( \bar{A}_j, \bar{L}_j \right) = a^p_j \left( h^\dagger_{j+1}, \eta^\dagger_{j+1}, \bar{A}_j, \bar{L}_j \right)
\]  

(35)

By Lemma 2, the right hand side in display (34) is equal to the bias of

\[
P_n \left[ Q_1 \left( \bar{h}^\dagger_1, \bar{\eta}^\dagger_1 \right) \right]
\]

as an estimator of \( \theta (g) \) for fixed functions \( h^\dagger_k \) and \( \eta^\dagger_k, k \in [K] \).

We see that it is comprised by a sum of \( K \) terms. Each term is equal to 0 if either \( h^\dagger_k = h_k \) or \( \eta^\dagger_k = \eta^\theta_k \).

4.7.2 Iterated regressions of multiply robust outcomes

The theoretical results of the preceding subsection suggest that the estimator

\( \hat{\theta}_{MR} \equiv P_n \left( \hat{Q}_1 \right) \) where \( \hat{Q}_1 \) is the random variable returned by the following algorithm is, under regularity conditions, \( 2^K \)-multiply robust CAN for \( \theta (g) \) in model \( \cap_{k=1}^K (H_k \cup G_k) \).

**Algorithm 4. (Iterated regression of multiply robust outcomes)** Set

\( \hat{Q}_{K+1} \equiv \psi \left( \bar{L}_{K+1} \right) \) and for \( k = K, K - 1, \ldots, 1 \),

a) Estimate \( \tau_k \) indexing the regression model

\[
\eta_{k, \tau_k} \left( \bar{A}_k, \bar{L}_k \right) \equiv \Psi \left\{ \tau^{T}_k s_k \left( \bar{A}_k, \bar{L}_k \right) \right\}
\]

for \( E \left( \hat{Q}_{k+1}| \bar{A}_k, \bar{L}_k \right) \) restricted to units verifying \( \pi^*k > 0 \) with \( \hat{r}_{k, MR} \)

solving

\[
P_n \left[ \pi^*s_k \left( \bar{A}_k, \bar{L}_k \right) \left\{ \hat{Q}_{k+1} - \Psi \left\{ \tau^{T}_k s_k \left( \bar{A}_k, \bar{L}_k \right) \right\} \right\} \right] = 0
\]  

(36)

35
b) For units with $\pi^{*k-1} > 0$, compute

$$\hat{Y}_{k,MR} \equiv \hat{y}_{k,MR} (A_{k-1}, L_k) \equiv \int h_k^* (a_k|A_{k-1}, L_k) \eta_{k,\hat{\tau}_{k,MR}} (a_k, A_{k-1}, L_k) \, d\mu_k (a_k)$$

and

$$\hat{Q}_k \equiv \frac{h_k^*}{h_k} [\hat{Q}_{k+1} - \eta_{k,\hat{\tau}_{k,MR}} (A_k, L_k)] + \hat{Y}_{k,MR}.$$ 

Tchetgen-Tchetgen (2009) proposed Algorithm 4 for estimation of the mean of an outcome at the end of longitudinal study with monotone missing at random data, i.e. for the target parameter $\theta (g)$ of example 1. The estimator $\hat{\theta}_{MR}$ from Algorithm 4 for the mean of $\theta (g)$ an arbitrary g-functional follows by applying the general theory for constructing multiply robust estimating functions discussed in Molina et al. (2017).

To analyze the limiting distribution of $\hat{\theta}_{MR}$ and that of several estimators that we shall introduce later, define for any $\eta_j^\dagger$ and $h_j^\dagger, j \in [K], \text{ and all } k \in [K],$

$$\Gamma_k \left( \hat{h}_{k+1}^{l K}, \hat{\eta}_k^{l K} ; g_k \right) \equiv \pi^{*k} \left[ \eta_k^\dagger (A_k, L_k) - E_{g_k} \left\{ Q_{k+1} \left( \hat{h}_{k+1}^{l K}, \hat{\eta}_{k+1}^{l K} | \hat{A}_k, \hat{L}_k \right) \right\} \right] \tag{37}$$

where, recall, $Q_K \left( \hat{h}_{K+1}^{l K}, \hat{\eta}_{K+1}^{l K} \right) \equiv \psi (T_{K+1}).$

From Lemma 2 we have that if $\pi^{*k} > 0$ then

$$\Gamma_k \left( \hat{h}_{k+1}^{l K}, \hat{\eta}_k^{l K} ; g_k \right) = 0 \text{ if } \eta_k^\dagger = \eta_k^g \text{ and for } j > k, \text{ either } \eta_j^\dagger = \eta_j^g \text{ or } h_k^\dagger = h_k^g \tag{38}$$

We further define

$$b^g (h^\dagger, \eta^\dagger) \equiv \sum_{k=1}^{K} E_{gh} \left[ \frac{1}{\pi_1^{(k-1)}} \left( \frac{1}{h_k} - \frac{1}{h_k^\dagger} \right) \Gamma_k \left( \hat{h}_{k+1}^{l K}, \hat{\eta}_k^{l K} ; g_k \right) \right].$$

In the Appendix we show the following useful decompositions:

**Lemma 3:** For any $\eta_j^\dagger, h_j^\dagger, j \in [K], \text{ it holds that } a^g (h^\dagger, \eta^\dagger) = b^g (h^\dagger, \eta^\dagger) = c^g (h^\dagger, \eta^\dagger)$
The identity $a^{p} (h^t, \eta^t) = c^{p} (h^t, \eta^t)$ will be helpful in our analysis, in section 5, of machine learning doubly and multiply robust estimators. Aside from this, it is interesting to note that we could have arrived at the identities (25), (29) and (30) by noticing that indeed because of the special way in which the iterated regression functions $\eta_k$ are estimated, it just happens that the doubly robust estimators defined earlier $\hat{\theta}_{Bang}, \hat{\theta}_{greed}$ and $\hat{\theta}_{reg}$ satisfy

$$
\hat{\theta}_{Bang} = P_n \left[ Q_1 \left( \hat{\eta}_{1}^{K}, \hat{\eta}_{Bang,1}^{K} \right) \right], \hat{\theta}_{greed} = P_n \left[ Q_1 \left( \hat{\eta}_{1}^{K}, \hat{\eta}_{greed,1}^{K} \right) \right]
$$

and

$$
\hat{\theta}_{reg} = P_n \left[ Q_1 \left( \hat{\eta}_{1}^{K}, \hat{\eta}_{reg,1}^{K} \right) \right].
$$

The identity $a^{p} (h^t, \eta^t) = b^{p} (h^t, \eta^t)$ and Lemma 2 immediately imply the following representation for $\hat{\theta}_{MR}$.

$$
\hat{\theta}_{MR} - \theta (g) = P_n \left\{ Q_1 \left( \hat{\eta}_{1}^{MR,1}, \hat{\eta}_{MR,1}^{K} \right) \right\} - E_{g_1} \left\{ Q_1 \left( \hat{\eta}_{1}^{MR,1}, \hat{\eta}_{MR,1}^{K} \right) \right\} + b^{p} \left( h, \hat{\eta}_{MR} \right) \tag{39}
$$

where $\hat{\eta}_{MR} \equiv (\hat{\eta}_{1,MR}, \ldots, \hat{\eta}_{K,MR})$ with $\hat{\eta}_{k,MR} \equiv \eta_{k,\hat{\tau}_{k,MR}}$.

Just as we reasoned earlier, to analyze the limiting distribution of $\hat{\theta}_{MR}$ we first note that the vectors $\hat{\tau}_{MR} \equiv (\hat{\tau}_{1,MR}, \ldots, \hat{\tau}_{K,MR})$ and $\hat{\alpha}_{ML}$ ultimately solve a joint system of estimating equations, so under regularity conditions, they have a probability limit under any $p \in P$ which we denote with $\tau_{lim,MR} (p)$ and $\alpha_{lim} (h)$. Furthermore, $\{ \hat{\tau}_{MR} - \tau_{lim,MR} (p) \}$ is asymptotically linear. Then, letting $\eta_{lim,MR} (p) \equiv \eta_{k,\tau_{lim,MR} (p)}$, if the map $(\alpha, \tau) \rightarrow b^{p} (h_{\alpha}, \eta_{\tau})$ is differentiable we have that

$$
b^{p} \left( h, \hat{\eta}_{MR} \right) - b^{p} \left[ h_{\lim}, \eta_{lim,MR} (p) \right] \text{ is asymptotically linear.}
$$

Furthermore, if $\hat{\eta}_{1}^{K}, \hat{\eta}_{MR,1}^{K}, h_{\lim} (h)$ and $\eta_{lim,MR} (p)$ fall in a Donsker class, then

$$
P_n \left\{ Q_1 \left( \hat{\eta}_{1}^{K}, \hat{\eta}_{MR,1}^{K} \right) \right\} - E_{g_1} \left\{ Q_1 \left( \hat{\eta}_{1}^{K}, \hat{\eta}_{MR,1}^{K} \right) \right\} =
$$

$$
P_n \left\{ Q_1 \left( h_{lim,1} (h), \hat{\eta}_{lim,MR,1} (p) \right) \right\} - E_{g_1} \left\{ Q_1 \left( h_{lim,1} (h), \hat{\eta}_{lim,MR,1} (p) \right) \right\} + o_p \left( n^{-1/2} \right)
$$

is asymptotically linear. 37
The representation (39) then implies that
\[ \hat{\theta}_{MR} - \theta (g) - b^p [h_{lim}, \eta_{lim,MR} (p)] \]
is asymptotically linear
Below we show that, under regularity conditions,
\[ b^p [h_{lim}, \eta_{lim,MR} (p)] = 0 \text{ if } p \in \cap_{k=1}^K (\mathcal{H}_k \cup \mathcal{G}_k) \] (40)
which then establishes that, under regularity conditions, \( \hat{\theta}_{MR} \) is CAN under model
\( \cap_{k=1}^K (\mathcal{H}_k \cup \mathcal{G}_k) \).
The assertion (40) is essentially a consequence of the following proposition.

**Proposition 2:**
\[ \eta_{k,lim,MR} (p) = \eta^g_k \text{ if } p \in \mathcal{G}_k \cap \bigcap_{k+j=K}^K (\mathcal{H}_k \cup \mathcal{G}_k) \] (41)
which we now show by induction.

**Proof of Proposition 2.** By reverse induction in \( k \). For \( k = K \), step (a) of Algorithm 4 is the same as step (a) of Algorithm 1, so
\[ \eta_{K,lim,MR} (p) = \eta_{K,lim,G} (p) \] and consequently, under regularity conditions, (41) holds for \( k = K \) as was already established in the proof of Proposition 1.
Next, suppose that (41) holds for \( k = K, \ldots, j + 1 \). Noticing that, by construction, \( \hat{Q}_{j+1} = Q_{j+1} \left( \hat{h}_{j+1}, \hat{\eta}_{MR,j+1} \right) \), we conclude that \( \hat{\tau}_{j,MR} \) solves
\[
0 = \mathbb{P}_n \left[ \pi^j s_j (\bar{A}_j, \bar{L}_j) \left\{ Q_{j+1} \left( \hat{h}_{lim,j+1}, \hat{\eta}_{lim,j+1} \right) - \Psi \left\{ \tau^T s_j (\bar{A}_j, \bar{L}_j) \right\} \right\} \right] + o_p (1)
\]
Suppose \( p \in \mathcal{G}_j \cap \bigcap_{k=j+1}^K (\mathcal{H}_k \cup \mathcal{G}_k) \). Then, for each \( k = j + 1, \ldots, K \), either \( p \in \mathcal{H}_k \) or \( p \in \mathcal{G}_k \). If \( p \in \mathcal{G}_k \) then since \( p \) also belongs to \( \bigcap_{r=k+1}^K (\mathcal{H}_r \cup \mathcal{G}_r) \) we have by inductive hypothesis that \( \eta_{k,lim,MR} (p) = \eta^g_k \). If \( p \in \mathcal{H}_k \), then \( h_{k,lim} = h_k \). Thus, for every \( k = j + 1, \ldots, K \), \( h_{k,lim} = h_k \) or \( \eta_{k,lim,MR} (p) = \eta^g_k \).
Consequently, by part (ii) of Lemma 2,
\[
E_{\eta_j} \left\{ Q_{j+1} \left( \hat{h}_{lim,j+1}, \hat{\eta}_{lim,j+1} \right) \left| \bar{A}_j, \bar{L}_j \right\} \eta^g_j (\bar{A}_j, \bar{L}_j) \right. = \eta^g_j (\bar{A}_j, \bar{L}_j) \). Furthermore, since \( p \in \hat{\mathcal{G}}_j \), \( \eta^g_j = \eta_{j,\tau_j (g_j)} \) for some \( \tau_j (g_j) \) and therefore the equation
\[
E_{\bar{A}_j, \bar{L}_j} \left[ \pi^j s_j (\bar{A}_j, \bar{L}_j) \left\{ Q_{j+1} \left( \hat{h}_{lim,j+1}, \hat{\eta}_{lim,j+1} \right) - \Psi \left\{ \tau^T s_j (\bar{A}_j, \bar{L}_j) \right\} \right\} \right] = 0
\]
is solved at $\tau_j = \tau_j (g_j)$. Then, under regularity conditions for the consistency of $M$--estimators, the probability limit $\tau_{j,\text{lim,MR}}$ of $\hat{\tau}_{j,\text{MR}}$ is equal to $\tau_j (g_j)$ which shows (41) holds for $k = j$.

Having shown (41) we now show that (40) holds by proving that for $p \in \cap_{r=k}^{K} (\mathcal{H}_r \cup \mathcal{G}_r)$ it holds that

$$E_{\mathcal{G}_{k-1} \mid h_k} \left[ \frac{1}{\pi (j-1)} \left( \frac{1}{h_k} - \frac{1}{h_{k,\text{lim}}} \right) \Gamma_k \left( \hat{\eta}_{\text{lim},k+1}^K, \hat{\eta}_{\text{lim},\text{MR},k} (p) ; g_k \right) \right] = 0 \quad (42)$$

Suppose then that $p \in \cap_{r=k}^{K} (\mathcal{H}_r \cup \mathcal{G}_r)$. If $p \in \mathcal{H}_k$ then $h_{k,\text{lim}} = h_k$ and thus (42) holds. If $p \notin \mathcal{H}_k$ then $p \in \mathcal{G}_k \cap [\cap_{r=k+1}^{K} (\mathcal{H}_r \cup \mathcal{G}_r)]$. Then, by (41), $\eta_{k,\text{lim,MR}} (p) = \eta_k^q$. In addition, for $r = k + 1, \ldots, K$, either $p \in \mathcal{H}_r$ in which case $h_{r,\text{lim}} = h_r$ or $p \in \mathcal{G}_r \cap [\cap_{s=r+1}^{K} (\mathcal{H}_s \cup \mathcal{G}_s)]$ in which case, again by (41), $\eta_{r,\text{lim,MR}} (p) = \eta_r^q$ and for $r = k + 1, \ldots, K$, either $\eta_{r,\text{lim,MR}} (p) = \eta_r^q$ or $h_{r,\text{lim}} = h_r$. Thus, by (38),

$$\Gamma_k \left( \hat{\eta}_{\text{lim},k+1}^K, \hat{\eta}_{\text{lim,MR},k} (p) ; g_k \right) = 0,$$

which then implies that (42) holds.

**Remark 1. (Another $K+1$ - multiply robust estimator)** By estimating in Algorithm 4 each $\tau_k$ regressing $\widehat{Q}_{k+1}$ on $s_k (\overline{A}_k, \overline{L}_k)$ we ensure that our estimator $\hat{\eta}_{k,\text{MR}}$ converges to $\eta_{k,\text{lim,MR}} (p)$ satisfying (41). Suppose instead that we estimate $\tau_k$ with the estimator $\hat{\tau}_{k,\mathcal{G}}$ of Algorithm 1, but we estimate $\theta (g)$ with

$$\hat{\theta}_{\text{DR}} = \mathbb{P}_n \left[ Q_1 \left( \hat{h}_1^K, \hat{\eta}_{1,\mathcal{G}} \right) \right] \quad (43)$$

where, recall from section 4.3, $\hat{\eta}_{\mathcal{G}} \equiv (\hat{\eta}_{1,\mathcal{G}}, \ldots, \hat{\eta}_{K,\mathcal{G}})$ and $\hat{\eta}_{k,\mathcal{G}} \equiv \hat{\eta}_{k,\hat{\tau}_{k,\mathcal{G}}}$. Then, with $\eta_{\text{lim,MR},p} (p) \equiv \eta_{k,\tau_{k,\text{lim},\mathcal{G}}}$ defined as in section 4.3, we have that under regularity conditions,

$$\hat{\theta}_{\text{DR}} - \theta (g) - b^p \left[ h_{\text{lim}}, \eta_{\text{lim,MR},p} (p) \right] \text{ is asymptotically linear.}$$

However, unlike $b^p \left[ h_{\text{lim}}, \eta_{\text{lim,MR},p} (p) \right]$, $b^p \left[ h_{\text{lim}}, \eta_{\text{lim,MR},p} (p) \right]$ is not equal to 0 for all $p \in \cap_{k=1}^{K} (\mathcal{H}_k \cup \mathcal{G}_k)$ because by estimating $\tau_k$ with $\hat{\tau}_{k,\mathcal{G}}$ we only ensure that $\eta_{k,\text{lim,MR},p} (p) = \eta_k^q$ if $p \in \cap_{j=k}^{K} \mathcal{G}_k$, as established in (19), but not necessarily for $p$ in the bigger model $\mathcal{G}_k \cap [\cap_{j=k+1}^{K} (\mathcal{H}_k \cup \mathcal{G}_k)]$. Yet, (19) does imply that $b^p \left[ h_{\text{lim}}, \eta_{\text{lim,MR},p} (p) \right] = 0$ for $p \in \cup_{j=1}^{K+1} [(\cap_{j=1}^{K} \mathcal{H}_k) \cap (\cap_{k=j}^{K} \mathcal{G}_k)]$. This implies that, under regularity conditions, $\hat{\theta}_{\text{DR}}$ is another $K+1$ multiply robust estimator.
Algorithm 4 may not always be feasible. Specifically, if the link function \( \Psi \) takes values in a bounded space, it may happen that the equation (36) does not have a solution as the values of \( \hat{Q}_{k+1,MR} \) can be arbitrarily large. Such will be the case whenever \( \hat{\pi}_{k+1,MR} \) is very close to 0 for some sample units. In fact, even if we had succeeded in computing \( \hat{\tau}_{k,MR} \) for all \( k \), we may still face the possibility that \( \hat{\theta}_{MR} \) falls outside the parameter space for \( \theta (g) \). For instance, if the parameter space for \( \theta (g) \) is the interval \((−\sigma, \sigma)\) for some \( \sigma > 0 \) (a situation which occurs when \( |\varphi(z)| < \sigma \) for all \( z \) where \( \varphi(z) \) is defined in 3), \( \hat{\theta}_{MR} \) may fall outside the interval \((−\sigma, \sigma)\) if for some units in the sample \( \hat{\pi}_K \) is very close to 0. In the next subsection we discuss a number of ways in which this problem can be overcome.

4.7.3 Inverse probability weighted iterated regression.

There exist a number of ways to overcome the issues with unbounded outcomes in Algorithm 4. In fact, remarkably, whenever for each \( j \in [K] \), \( s_j(\bar{a}_j, \bar{t}_j) \) can be decomposed as

\[
s_j(\bar{a}_j, \bar{t}_j) = \begin{bmatrix} b_j(\bar{a}_j, \bar{t}_j) \\ s_{j-1}(\bar{a}_{j-1}, \bar{t}_{j-1}) \end{bmatrix}
\]

(44)

for some known, possibly vector valued, function \( b_j \), where \( s_0(\bar{a}_0, \bar{t}_0) \equiv 1 \), it just happens that the weighted iterated regression estimator \( \hat{\theta}_{reg} \) of section 4.4 using weights \( \omega_k(\bar{A}_k, \bar{L}_k) = 1/\hat{\pi}_k \) is \( 2^K \)-multiply robust, i.e. it is CAN for \( \theta (g) \) in model \( \cap_{k=1}^K (H_k \cup G_k) \). Note that if (44) does not hold it it is always possible to enlarge the parametric class \( R_j \) by adding to the covariate vector \( s_j(\bar{A}_j, \bar{t}_j) \) the components of \( s_{j-1}(\bar{A}_{j-1}, \bar{t}_{j-1}) \) that are not in \( s_j(\bar{A}_j, \bar{t}_j) \) so as to ensure that (44) holds. Unlike the outcomes in step (a) of Algorithm 4, by construction, the outcomes \( \hat{Y}_{k+1,reg} \equiv \hat{Y}_{k+1,\omega}, k \in [K-1] \), being regressed to obtain the estimator \( \hat{\tau}_{k,reg} \) are guaranteed to fall in the range of the link function \( \Psi (\cdot) \). Thus, so long as the sample space of \( \psi(\bar{L}_{K+1}) \) falls in the range of \( \Psi (\cdot) \), the equation (20) with \( \omega_k(\bar{A}_k, \bar{L}_k) = 1/\hat{\pi}_k \) and with \( \hat{Y}_{k+1,reg} \) replacing \( \hat{Y}_{k+1,\omega} \), is guaranteed to have a solution for all \( k \in [K] \). Furthermore, if the range of \( \Psi (\cdot) \) and sample space of \( \psi(\bar{L}_{K+1}) \) agree, then the estimator \( \theta_{reg} \) is guaranteed to fall in the sample space of \( \theta (g) \).

To see why \( \hat{\theta}_{reg} \) is \( 2^K \)-multiply robust when \( s_j(\bar{A}_j, \bar{t}_j) \) is a sub-vector of \( s_k(\bar{A}_k, \bar{L}_k) \)
for any $k > j$, notice that in such case \( \tilde{\tau}^T_{k,\text{reg}} \) satisfies

\[
\mathbb{P}_n \left[ \frac{\pi^j_{s_j}}{\hat{\pi}^j} s_j (\bar{A}_j, \bar{L}_j) \sum_{k=j+1}^{K} \left\{ \left( \tilde{Y}_{k+1,\text{reg}} - \hat{\eta}_{k,\text{reg}} \right) \frac{\pi^{j+1}_k}{\hat{\pi}^{j+1}_k} \right\} \right] = 0
\]

where \( \hat{\eta}_{k,\text{reg}} \equiv \Psi \{ \tilde{\tau}^T_{k,\text{reg}} s_k (\bar{A}_k, \bar{L}_k) \} \). Thus, for any \( j \in [K] \), \( \tilde{\tau}_{j,\text{reg}} \) solves

\[
0 = \mathbb{P}_n \left\{ \frac{\pi^j_{s_j}}{\hat{\pi}^j} s_j (\bar{A}_j, \bar{L}_j) \left[ \tilde{Y}_{j+1,\text{reg}} + \sum_{k=j+1}^{K} \left( \tilde{Y}_{k+1,\text{reg}} - \hat{\eta}_{k,\text{reg}} \right) \frac{\pi^{j+1}_k}{\hat{\pi}^{j+1}_k} \right] - \Psi \left( \tau^T_{j} s_j (\bar{A}_j, \bar{L}_j) \right) \right\}
\]

where for any \( j \in [K-1] \), \( \tilde{Q}_{j+1,\text{reg}} \equiv Q_{j+1} \left( \frac{\pi^K_{h_{j+1}}}{\pi^K_{\hat{h}_{j+1}}}, \frac{\pi^K_{\hat{\eta}_{\text{reg},j+1}}}{\pi^K_{\hat{\eta}_{\text{lim},\text{reg}}}} \right) \). Also,

\[
\hat{\theta}_{\text{reg}} = \mathbb{P}_n \left[ \tilde{Y}_{1,\text{reg}} + \sum_{k=1}^{K} \frac{\pi^{j+1}_k}{\hat{\pi}^{j+1}_k} \left( \tilde{Y}_{k+1,\text{reg}} - \hat{\eta}_{k,\text{reg}} \right) \right]
\]

\[
= \mathbb{P}_n \left( \tilde{Q}_{1,\text{reg}} \right)
\]

Thus, \( \hat{\theta}_{\text{reg}} \) is indeed the result of fitting Algorithm 4 except that in equation (36) \( \pi^k s_k (\bar{A}_k, \bar{L}_k) / \hat{\pi}^k \) instead of \( s_k (\bar{A}_k, \bar{L}_k) \) multiplies the difference \( \left\{ \tilde{Q}_{k+1} - \Psi \left( \tau^T_{k} s_k (\bar{A}_k, \bar{L}_k) \right) \right\} \). The proof that \( \hat{\theta}_{\text{reg}} \) is CAN for \( \theta (g) \) under \( \cap_{k=1}^{K} (H_k \cup G_k) \) is essentially the same as the proof that \( \hat{\theta}_{\text{MR}} \) is CAN under the same model. Note, however, that the variances of the limiting mean zero normal distributions of \( \hat{\theta}_{\text{MR}} \) and \( \hat{\theta}_{\text{reg}} \) are not the same because

\[
n^{1/2} \left\{ b^p \left( \hat{h}, \hat{\eta}_{\text{MR}} \right) - b^p \left[ h_{\text{lim}}, \eta_{\text{lim},\text{MR}} (p) \right] \right\}
\]

and

\[
n^{1/2} \left\{ b^p \left( \hat{h}, \hat{\eta}_{\text{reg}} \right) - b^p \left[ h_{\text{lim}}, \eta_{\text{lim},\text{reg}} (p) \right] \right\}
\]

converge to different mean zero normal distributions.

### 4.7.4 Greedy-fit multiply robust iterated regression.

The estimator \( \hat{\theta}_{\text{reg}} \) of section 4.7.3 requires that one fits models \( R_k \) given in (13) whose dimension grow with \( k \). When \( K \) is large, step (a) of the algorithm would
then require the fit by (weighted) IRLS of a large model and thus may result in numerical instability problems. The following extension of the greedy fit Algorithm 3, results in a \(2^K\)-multiply robust estimator of \(\theta (g)\) which does not require that the models \(R_k\) be of growing dimension. Furthermore, the estimation procedure never requires the fit of a model whose parameter has dimension larger than \(\max \{ \dim (\tau_k) : k \in [K] \}\), where \(\tau_k\) is the parameter indexing model \(R_k\). In what follows \(s_0 (A_0, L_0) \equiv 1\).

**Algorithm 5. (Multiply robust estimation by greedy fit iterated regression)**

For \(j \in [K]\) set \(\hat{Y}_{k+1}^{(j)} = \psi (L_{k+1})\), define \(s_0 (A_0, L_0) \equiv 1\), and for any \(k = K, K - 1, \ldots, 1\), repeat

\[\hat{\pi} \equiv \hat{\tau \tau}, \text{ for } E (\hat{Y}^{(k)} | A_k, L_k) \text{ restricted to units verifying } \hat{\pi} > 0 \text{ with } \hat{\pi} \text{ solving } \]

\[\mathbb{P} \left\{ \hat{\pi} \left( A_k, L_k \right) \left\{ \hat{Y}^{(k)} - \Psi \{ \hat{\tau \tau}, s_k (A_k, L_k) \} \right\} \right\} = 0 \quad (45)\]

**b)** For \(j = k - 1, k - 2, \ldots, 0\), repeat \{  

\[i) \text{ Estimate the parameter } \lambda^{(j)} \theta \text{ indexing the regression model} \]

\[\eta^{(j)} \theta (A_k, L_k) \equiv \Psi \left\{ \tau_k^{T} s_k (A_k, L_k) \right\} \]

\[\text{for } E (\hat{Y}^{(k)} | A_k, L_k) \text{ restricted to units verifying } \hat{\pi} > 0 \text{ with } \hat{\pi} \text{ solving } \]

\[\mathbb{P} \left\{ \hat{\pi} \left( A_k, L_k \right) \left\{ \hat{Y}^{(k)} - \Psi \{ \lambda^{(j)} \theta, s_k (A_k, L_k) \} \right\} \right\} = 0 \quad (46)\]

Let \(\hat{\eta}^{(j)} (A_k, L_k) \equiv \eta^{(j)} \theta \left( A_k, L_k \right) \) .
ii) For units with \( \pi^{*k-1} > 0 \), compute
\[
\hat{Y}_k^{(j)} \equiv y_{k,\hat{\pi}_k^{(j)}}(\mathbf{A}_{k-1}, L_k)
\]
\[
\equiv \int h_k^* (a_k | \mathbf{A}_{k-1}, L_k) \eta_{k, \hat{\pi}_k^{(j)}}(a_k, \mathbf{A}_{k-1}, L_k) \, d\mu_k(a_k)
\]

Finally, \( \hat{\theta}_{MR,\text{greed}} = \mathbb{P}_n \left( \hat{Y}_1^{(0)} \right) \).

By construction, each \( \hat{\tau}_{j,\text{greed}} \) is the estimated coefficient in a regression on \( (\mathbf{A}_j, L_j) \) of the outcome \( \hat{Y}_j^{(j)} \) with weights \( \pi^{*j} \). On the other hand, step (b) of the algorithm (the fit of the extended model) ensures precisely that \( \hat{\tau}_{j,\text{greed}} \) is also the estimated coefficient in a regression on \( (\mathbf{A}_j, L_j) \) of the pseudo outcome \( Q_{j+1} (\hat{\pi}^K_{h_{j+1}}, \hat{\pi}_{j+1}^{(j)}) \) with weights \( \pi^{*j} \). Specifically, by step (a) \( \hat{\tau}_{j,\text{greed}} \) satisfies
\[
0 = \mathbb{P}_n \left[ \pi^{*j} S_j (\mathbf{A}_j, L_j) \left\{ \hat{Y}_j^{(j)} - \eta_{j, \hat{\tau}_{j,\text{greed}}} \right\} \right]
\]

Because, by step (b.i), for all \( j < k \leq K \),
\[
\mathbb{P}_n \left[ \pi^{*j} S_j (\mathbf{A}_j, L_j) \left\{ \left( \hat{Y}_{k+1}^{(j)} - \hat{\eta}_k^{(j)} \right) \left( \pi^{*k}_{j+1} / \hat{\pi}_{j+1}^{(k)} \right) \right\} \right] = 0
\]

where, recall, \( \hat{\eta}_k^{(j)} \equiv \eta_{k, \hat{\pi}_k^{(j)}} \), then (47) implies that
\[
0 = \mathbb{P}_n \left[ \pi^{*j} S_j (\mathbf{A}_j, L_j) \left\{ \hat{Y}_{j+1}^{(j)} - \eta_{j, \hat{\tau}_{j,\text{greed}}} \right\} \right] + \mathbb{P}_n \left[ \sum_{k=j+1}^{K} \pi^{*j} S_j (\mathbf{A}_j, L_j) \left\{ \left( \hat{Y}_{k+1}^{(j)} - \hat{\eta}_k^{(j)} \right) \left( \pi^{*k}_{j+1} / \hat{\pi}_{j+1}^{(k)} \right) \right\} \right]
\]

This last equality, in turn, is equal to
\[
0 = \mathbb{P}_n \left[ \pi^{*j} S_j (\mathbf{A}_j, L_j) \left\{ Q_{j+1} \left( \pi^K_{h_{j+1}}, \hat{\pi}_{j+1}^{(j)} \right) - \eta_{j, \hat{\tau}_{j,\text{greed}}} (\mathbf{A}_j, L_j) \right\} \right] \quad (48)
\]

Furthermore,
\[
\hat{\theta}_{MR,\text{greed}} = \mathbb{P}_n \left( \hat{Y}_1^{(0)} \right) = \mathbb{P}_n \left[ Q_{j+1} \left( \pi^K_{h_{j+1}}, \hat{\pi}_1^{(0)} \right) \right] \quad (49)
\]

An analysis similar to that conducted for \( \hat{\theta}_{MR} \) now shows that \( \hat{\theta}_{MR,\text{greed}} \) is \( 2^K \)-multiply robust CAN for \( \theta(g) \) in model \( \bigcap_{k=1}^{K} (G_k \cup H_k) \).
We now illustrate the implementation of the estimators $\hat{\theta}_{reg}, \hat{\theta}_{MR}$ and $\hat{\theta}_{MR,\text{greed}}$ for $K = 2$ in Example 1, assuming $\psi (L_{K+1}) = L_3$ and $L_3$ is a binary outcome. In model $R_k, k = 1, 2$, we need only consider a parametric class for $\eta_k (\vec{I}_k; g_k) \equiv \eta_k (\vec{a}_k = 1, \vec{I}_k; g_k), k = 1, 2$, because in the algorithms that compute of $\hat{\theta}_{reg}, \hat{\theta}_{MR}$ and $\hat{\theta}_{MR,\text{greed}}$, the regression in step (a) is restricted to subjects with $\pi^{*k} = 1$, i.e. with $A_k = \overline{\text{I}}$. As indicated in section 2.1.1, under the assumptions of Example 1, $\eta_k (\vec{I}_k; g_k)$ coincides with $E (L_3 | \overline{A}_k = \overline{\text{I}}, \vec{L}_k = \vec{I}_k), k = 1, 2$, where, recall, $L_3^*$ is the intended, possibly unobserved outcome. If $L_3^*$ is binary, a natural parametric model $\eta_{k, \tau_k} (\vec{I}_k)$ for $\eta_k (\vec{I}_k; g_k)$ is then a logistic regression model

$$\eta_{k, \tau_k} (\vec{I}_k) = \exp \{ \tau_k s_k (\vec{I}_k) \}$$

for a vector $s_k (\vec{I}_k)$ of given functions of $\vec{I}_k$ which includes one entry with the constant 1.

The calculation of $\hat{\theta}_{reg}, \hat{\theta}_{MR}$ and $\hat{\theta}_{MR,\text{greed}}$ requires that we first fit by maximum likelihood parametric models for each $h_k$. Since $A_k$ is binary and by assumption, $A_k = 0 \Rightarrow A_j = 0$ for $j > k$, then a natural parametric model for $h_k (a_k | \overline{\alpha}_{k-1}, \vec{I}_k)$ is

$$h_{k, \alpha_k} (a_k | \overline{\alpha}_{k-1}, \vec{I}_k) = a_{k-1} \exp \{ a_k \alpha_k^T r_k (\vec{I}_k) \} / \{ 1 + \exp \{ \alpha_k^T r_k (\vec{I}_k) \} \}$$

where $r_k (\vec{I}_k)$ is a vector of specified functions of $\vec{I}_k$. That is, we assume that the probability of response at cycle $k + 1$ among those still in study at cycle $k$ follows a logistic regression with covariate vector $r_k (\vec{I}_k)$. Because by definition, $A_0 = 1$ the estimator $h_{1, \overline{\alpha}_{ML,1}} (1 | \overline{A}_0, \overline{\text{I}}_1)$ is the fitted value from the logistic regression of the binary outcome $A_1$ on the covariate vector $r_1 (\overline{\text{I}}_1)$ among the entire study participants. On the other hand, $h_{2, \overline{\alpha}_{ML,2}} (1 | \overline{A}_1, \overline{\text{I}}_2)$ is equal to 0 for subjects for whom $L_2^*$ is missing, i.e. for which $A_1 = 0$, and it is equal to the fitted value from the logistic regression of outcome $A_2$ on covariates $r_2 (\overline{\text{I}}_2)$ among subjects for which $L_2^*$ is observed. In what follows we describe the three algorithms. To simplify notation, we use the shortcuts $\hat{h}_1 \equiv h_{1, \overline{\alpha}_{ML,1}} (1 | \overline{\text{I}}_1)$ and $\hat{h}_2 \equiv h_{2, \overline{\alpha}_{ML,2}} (1 | A_1 = 1, \overline{\text{I}}_2)$. In what follows we explain in detail the algorithm to compute $\hat{\theta}_{MR,\text{greed}}$. To avoid repetition, the algorithms for computing $\hat{\theta}_{reg}, \hat{\theta}_{MR}$ are given with less detail.
Calculation of $\hat{\theta}_{\text{greed}}$. \textit{(Greedy fit multiply robust estimation)}

Steps for $k = 2$

(a) This step requires that we use subjects with $\pi^{*2} > 0$. These are precisely the subjects with $A_2 = 1$, i.e. the subjects that did not dropped from the study. This step of the algorithm requires that we fit model

$$\eta_{2,\tau_2} (\overline{A}_2, \overline{T}_2) \equiv \expit \left\{ \tau_2^T s_2 (\overline{A}_2, \overline{T}_2) \right\}$$

just using subjects with $A_2 = 1$. Because subjects with $A_2 = 1$ must necessarily have $A_1 = 1$, then the relevant model that we need to estimate is

$$\eta_{2,\tau_2} (\overline{A}_2 = 1, \overline{T}_2) \equiv \expit \left\{ \tau_2^T s_2 (\overline{A}_2 = 1, \overline{T}_2) \right\}$$

If, as we indicated at the start of this section, in a slight abuse of notation we write $s_2 (\overline{T}_2) \equiv s_2 (\overline{A}_2 = 1, \overline{T}_2)$, then this step of the algorithm boils down to computing the logistic regression estimator $\hat{\tau}_{2,\text{greed}}$ for the outcome $\hat{Y}_3^{(2)} \equiv L_3$ on the covariate vector $s_2 (\overline{T}_2)$ just using subjects $A_2 = 1$. That is, $\hat{\tau}_{2,\text{greed}}$ satisfies

$$\mathbb{P}_n \left( A_2 s_2 (\overline{T}_2) \left[ \hat{Y}_3^{(2)} - \expit \left\{ \hat{\tau}_{2,\text{greed}}^T s_2 (\overline{T}_2) \right\} \right] \right) = 0 \quad (50)$$

Step (b) for $k = 2, j = 1$ (i) In this step we are required to use again only subjects with $\pi^{*2} > 0$, so we continue to restrict the calculations to subjects with $A_2 = 1$. Using these subjects we are required to fit the logistic regression model $\eta_{2,\lambda_2}^{(1)} (\overline{A}_2, \overline{T}_2)$ for $E \left( \hat{Y}_3^{(1)} | \overline{A}_2, \overline{T}_2 \right)$ where $\hat{Y}_3^{(1)} \equiv L_3$,

$$\eta_{2,\lambda_2}^{(1)} (\overline{A}_2, \overline{T}_2) \equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^T s_2 (\overline{A}_2, \overline{T}_2) + \lambda_2^{(1)} \frac{s_1 (\overline{A}_1, \overline{L}_1)}{\hat{\pi}_2^2} \right\}$$

$$\equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^T s_2 (\overline{A}_2, \overline{T}_2) + \lambda_2^{(1)} \frac{s_1 (\overline{A}_1, \overline{L}_1)}{\hat{h}_2} \right\},$$

$\hat{\tau}_{2,\text{greed}}^T s_2 (\overline{A}_2, \overline{T}_2)$ is an offset and $\lambda_2^{(1)}$ is the unknown parameter. Once again, because we are only using subjects with $\overline{A}_2 = 1$, then we only care
to fit the model for \( E \left( \hat{Y}^{(1)}_3 | \bar{A}_2 = 1, \bar{L}_2 \right) \):

\[
\eta^{(1)}_{2, \lambda^{(i)}_2} (\bar{A}_2 = 1, \bar{L}_2) \equiv \expit \left( \hat{\tau}^T \text{greed}_2 s_2 (\bar{L}_2) + \lambda^{(i)}_2 \frac{s_1 (L_1)}{\bar{h}_2} \right)
\]

where \( s_k (L_k) \equiv s_k (A_k = 1, L_k), k = 1, 2 \). Thus, the estimator \( \hat{\lambda}^{(i)}_2 \) satisfies

\[
P_n \left( A_2 \frac{s_1 (L_1)}{\bar{h}_2} \left[ \hat{Y}^{(1)}_3 - \expit \left( \hat{\tau}^T \text{greed}_2 s_2 (\bar{L}_2) + \hat{\lambda}^{(i)}_2 \frac{s_1 (L_1)}{\bar{h}_2} \right) \right] \right) = 0
\]

(ii) This step is calculated using only subjects with \( \pi^{i_1} > 0 \), i.e. with \( A_1 = 1 \). For these subjects we must compute

\[
\hat{Y}^{(1)}_2 \equiv \eta^{(1)}_{2, \lambda^{(i)}_2} (A_1, \bar{L}_2)
\]

\[
\equiv \int \eta^{(1)}_{2, \lambda^{(i)}_2} (A_1, a_2, \bar{L}_2) h^*_2 (a_2 | A_1, \bar{L}_2) \, d\mu (a_2)
\]

\[
= \sum_{a_2=0}^1 \eta^{(1)}_{2, \lambda^{(i)}_2} (A_1, a_2, \bar{L}_2) h^*_2 (a_2 | A_1, \bar{L}_2)
\]

Now, because we are only doing the calculation for subjects with \( A_1 = 1 \), and because \( h^*_2 (a_2 | A_1 = 1, \bar{L}_2) = a_2 \), the last display simplifies to

\[
\hat{Y}^{(1)}_2 = \eta^{(1)}_{2, \lambda^{(i)}_2} (A_1 = 1, A_2 = 1, \bar{L}_2)
\]

\[
= \expit \left( \hat{\tau}^T \text{greed}_2 s_2 (\bar{L}_2) + \hat{\lambda}^{(i)}_2 \frac{s_1 (L_1)}{\bar{h}_2} \right)
\]

**Step (b) for \( k = 2, j = 0 \) (i)** In this step we are required to use again only subjects with \( \pi^{i_2} > 0 \), so we continue to restrict the calculations to subjects with \( A_2 = 1 \). Using these subjects we are now required to fit the logistic regression model \( \eta^{(0)}_{2, \lambda^{(j)}_2} (\bar{A}_2, \bar{L}_2) \) for \( E \left( \hat{Y}^{(0)}_3 | \bar{A}_2, \bar{L}_2 \right) \) where
This step is calculated using only subjects with $$\pi^* > 0$$, i.e. with $$A_1 = 1$$. For these subjects we must compute

$$\hat{Y}_2^{(0)} = \eta_{2,\lambda_2^{(0)}}^{(0)}(A_1, \overline{L}_2)$$

$$\equiv \int \eta_{2,\lambda_2^{(0)}}^{(0)}(A_1, a_2, \overline{L}_2) h_2^*(a_2 | A_1, \overline{L}_2) \, d\mu(a_2)$$

$$= \sum_{a_2=0}^1 \eta_{2,\lambda_2^{(0)}}^{(0)}(A_1, a_2, \overline{L}_2) h_2^*(a_2 | A_1, \overline{L}_2)$$

Now, because we are only doing the calculation for subjects with $$A_1 = 1$$, and because $$h_2^*(a_2 | A_1 = 1, \overline{L}_2) = a_2$$, the last display simplifies to

$$\hat{Y}_2^{(0)} = \eta_{2,\lambda_2^{(0)}}^{(0)}(A_1 = 1, A_2 = 1, \overline{L}_2)$$

$$\equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^T \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} + \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} \right\}$$

$$= \expit \left\{ \hat{\tau}_{2,\text{greed}}^T \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} + \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} \right\}$$

where $$s_k(\overline{L}_k) \equiv s_k(\overline{L}_k), k = 1, 2$$. Thus, the estimator $$\hat{\lambda}_2$$ satisfies

$$\mathbb{P}_n \left( A_2 \frac{1}{\hat{h}_1 \hat{h}_2} \left[ \hat{Y}_3^{(0)} - \expit \left\{ \hat{\tau}_{2,\text{greed}}^T \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} + \hat{\lambda}_2 \frac{s_1}{\hat{h}_2} \right\} \right] \right) = 0 \quad (52)$$

(ii) This step is calculated using only subjects with $$\pi^* > 0$$, i.e. with $$A_1 = 1$$. For these subjects we must compute

$$\hat{Y}_3^{(0)} \equiv L_3,$$

$$\eta_{2,\lambda_2^{(0)}}^{(0)}(\overline{A}_2, \overline{L}_2) \equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^{(0)}(\overline{A}_2, \overline{L}_2) + \hat{\lambda}_2^{(0)} \frac{s_1}{\hat{h}_2} + \lambda_2 \frac{1}{\hat{h}_2} \right\}$$

$$\equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^{(0)}(\overline{A}_2, \overline{L}_2) + \hat{\lambda}_2^{(0)} \frac{s_1}{\hat{h}_2} + \lambda_2 \frac{1}{\hat{h}_2} \right\},$$

$$\hat{\tau}_{2,\text{greed}}^{(0)}(\overline{A}_2, \overline{L}_2) + \hat{\lambda}_2^{(0)} \frac{s_1}{\hat{h}_2}$$ is an offset and $$\lambda_2^{(0)}$$ is the unknown parameter. Once again, because we are only using subjects with $$\overline{A}_2 = \overline{T}$$, then we only care to fit the model for $$E(\hat{Y}_3^{(0)} | \overline{A}_2 = 1, \overline{T}_2):$$

$$\eta_{2,\lambda_2^{(0)}}^{(0)}(\overline{A}_2 = 1, \overline{T}_2) \equiv \expit \left\{ \hat{\tau}_{2,\text{greed}}^{(0)}(\overline{T}_2) + \hat{\lambda}_2^{(0)} \frac{s_1}{\hat{h}_2} + \lambda_2 \frac{1}{\hat{h}_2} \right\}$$

where $$s_k(\overline{T}_k) \equiv s_k(\overline{A}_k = 1, \overline{T}_k), k = 1, 2$$. Thus, the estimator $$\hat{\lambda}_2$$ satisfies
Steps for $k = 1$

(a) This step requires that we use subjects with $\pi^* > 0$. These are precisely the subjects with $A_1 = 1$. This step of the algorithm requires that we fit the model for $E(\hat{Y}_2^{(1)} | \overline{A}_1, \overline{L}_1)$:

$$\eta_{1, \tau_1} (\overline{A}_1, \overline{L}_1) \equiv \exp \{ \tau_1^T s_1 (\overline{A}_1, \overline{L}_1) \}$$

just using subjects with $A_1 = 1$. Then the model we care to estimate is actually

$$\eta_{1, \tau_1} (A_1 = 1, \overline{L}_1) \equiv \exp \{ \tau_1^T s_1 (A_1 = 1, \overline{L}_1) \}$$

Writing $s_1 (\overline{L}_1) \equiv s_1 (A_1 = 1, \overline{L}_1)$, then this step of the algorithm boils down to computing the logistic regression estimator $\hat{\tau}_{1, \text{greed}}$ for the outcome $\hat{Y}_2^{(1)} \equiv L_3$ on the covariate vector $s_1 (\overline{L}_1)$ just using subjects $A_1 = 1$. Then, the estimator $\hat{\tau}_{1, \text{greed}}$ satisfies

$$\mathbb{P}_n \left[ A_1 s_1(\overline{L}_1) \left\{ \hat{Y}_2^{(1)} - \exp \{ \hat{\tau}_{1, \text{greed}} s_1 (\overline{L}_1) \} \right\} \right] = 0 \quad (54)$$

Step (b) for $k = 1$, $j = 0$ (i) In this step we are required to use again only subjects with $\pi^* > 0$, so we continue to restrict the calculations to subjects with $A_1 = 1$. Using these subjects we are required to fit the logistic regression model $\eta_{1, \lambda_1}^{(0)} (\overline{A}_1, \overline{L}_1)$ for $E(\hat{Y}_2^{(0)} | \overline{A}_1, \overline{L}_1)$ where $\hat{Y}_2^{(0)}$ was calculated in (53) and

$$\eta_{1, \lambda_1}^{(0)} (\overline{A}_1, \overline{L}_1) \equiv \exp \left\{ \hat{\tau}_{1, \text{greed}} s_1 (\overline{A}_1, \overline{L}_1) + \lambda_1^{(0)} \frac{1}{\hat{h}_1} \right\}$$

$$\equiv \exp \left\{ \hat{\tau}_{1, \text{greed}} s_1 (\overline{A}_1, \overline{L}_1) + \lambda_1^{(0)} \frac{1}{\hat{h}_1} \right\}$$

$\hat{\tau}_{1, \text{greed}} s_1 (\overline{A}_1, \overline{L}_1)$ is an offset and $\lambda_1^{(0)}$ is the unknown parameter. Once again, because we are only using subjects with $\overline{A}_1 = \overline{1}$, then we only care to fit the model for $E(\hat{Y}_2^{(0)} | \overline{A}_1 = 1, \overline{L}_1)$:

$$\eta_{1, \lambda_1}^{(0)} (\overline{A}_1 = 1, \overline{L}_1) \equiv \exp \left\{ \hat{\tau}_{1, \text{greed}} s_1 (\overline{L}_1) + \lambda_1^{(0)} \frac{1}{\hat{h}_1} \right\}$$
where \( s_1(L_1) \equiv s_1(A_1 = 1, \overline{L}_1) \). Thus, the estimator \( \hat{\lambda}_1^{(0)} \) satisfies

\[
\mathbb{P}_n \left( A_1 \frac{1}{h_1} \left[ \hat{Y}_2^{(0)} - \expit \left\{ \hat{\tau}_{1, \text{greed}}^{T} s_1(\overline{L}_1) \hat{\lambda}_1^{(0)} \frac{1}{h_1} \right\} \right] \right) = 0 \tag{55}
\]

**(ii)** This step is calculated using only subjects with \( \pi^{*0} > 0 \), i.e. all subjects in the sample because by definition, \( \pi^{*0} = 1 \). For all subjects we must then compute

\[
\hat{Y}_1^{(0)} \equiv y_{1, \hat{\lambda}_1^{(0)}}^{(0)} (L_1) = \int \eta_{1, \hat{\lambda}_1^{(0)}}^{(0)} (a_1, L_1) h_1^{*} (a_1 | L_1) \, d\mu(a_1)
\]

\[
= \sum_{a_2=0}^{1} \eta_{1, \hat{\lambda}_1^{(0)}}^{(0)} (a_1, L_1) h_1^{*} (a_1 | L_1)
\]

Now, because \( h_1^{*} (a_1 | L_1) = a_1 \), the last display simplifies to

\[
\hat{Y}_1^{(0)} = \eta_{1, \hat{\lambda}_1^{(0)}}^{(0)} (A_1 = 1, \overline{L}_1)
\]

\[
= \expit \left\{ \hat{\tau}_{1, \text{greed}}^{T} s_1(\overline{L}_1) \hat{\lambda}_1^{(0)} \frac{1}{h_1} \right\}
\]

Finally, the estimator \( \hat{\theta}_{MR, \text{greed}} \) is \( \mathbb{P}_n \left( \hat{Y}_1^{(0)} \right) \). That is, \( \hat{\theta}_{MR, \text{greed}} \) satisfies

\[
\mathbb{P}_n \left( \hat{Y}_1^{(0)} - \hat{\theta}_{MR, \text{greed}} \right) = 0 \tag{56}
\]

Note that the outcomes \( \hat{Y}_j^{(k)} \) being regressed at each iteration of the algorithm are bounded between 0 and 1. But, unlike for the computation of \( \hat{\theta}_{\text{reg}} \) given below, to compute \( \hat{\theta}_{MR, \text{greed}} \) we do require that \( s_1(\overline{L}_1) \) be a subvector of \( s_2(\overline{L}_2) \) nor that 1 be a component of \( s_1(\overline{L}_1) \) and \( s_2(\overline{L}_2) \).

We now derive equations (48) and (49) for this example. To arrive at (49) we sum the equations (52), (55) and (56), i.e. the equations in which the outcome has a
superscript \((0)\), and obtain

\[
0 = \mathbb{P}_n \left[ \left( \hat{Y}_1^{(0)} - \hat{\theta}_{MR,greed} \right) + \frac{A_1}{h_1} \left[ \hat{Y}_2^{(0)} - \expit \left\{ \hat{\tau}_{1,greed}s_1 \left( \mathcal{T}_1 \right) + \hat{\lambda}_1^{(0)} \frac{1}{h_1} \right\} \right] \right.
\]
\[
\quad + \frac{A_2}{h_1 h_2} \left[ \hat{Y}_3^{(0)} - \expit \left\{ \hat{\tau}_{2,greed}s_2 \left( \mathcal{T}_2 \right) + \hat{\lambda}_2^{(0)} \frac{1}{h_2} + \hat{\lambda}_2^{(0)} \frac{1}{h_1 h_2} \right\} \right] \right) 
\]
\[
= \mathbb{P}_n \left[ \left( y_1,\eta^{(0)}_1,\hat{\eta}^{(0)}_1 \right) (\mathcal{T}_1) - \hat{\theta}_{MR,greed} \right] + \frac{A_1}{h_1} \left[ y_2,\eta^{(0)}_2,\hat{\eta}^{(0)}_2 \right] (\mathcal{T}_2) - \eta^{(0)}_1 (\mathcal{T}_1) \right) 
\]
\[
+ \frac{A_1 A_2}{h_1 h_2} \left[ L_3 - \eta^{(0)}_2,\hat{\eta}^{(0)}_2 \right] (\mathcal{T}_2) \right) \right) 
\]
\[
= \mathbb{P}_n \left[ \frac{A_1 A_2}{h_1 h_2} \hat{Y}_3 \right] - \left\{ \frac{A_1}{h_1} \eta^{(0)}_1,\hat{\lambda}^{(0)}_1 \right\} (\mathcal{T}_1, \mathcal{T}_2) \right) \right) 
\]
\[
- \left\{ \frac{A_1 A_2}{h_1 h_2} \eta^{(0)}_2,\hat{\lambda}^{(0)}_2 \right\} (\mathcal{T}_2) - \frac{A_1}{h_1} y_2,\eta^{(0)}_2,\hat{\eta}^{(0)}_2 \right\] (\mathcal{T}_2) \right) \right) \right) \right) \right) \right) 
\]
\[
= \mathbb{P}_n \left[ \frac{\pi^{*}}{\pi} L_3 - \sum_{k=1}^{2} \left\{ \frac{\pi^k}{\pi} \eta^{(0)}_k,\hat{\lambda}^{(0)}_k \right\} (\mathcal{T}_k) - \frac{\pi^{*(k-1)}}{\pi} y_k,\eta^{(0)}_k,\hat{\eta}^{(0)}_k \right\] (\mathcal{T}_k) \right) \right) \right) \right) 
\]
\[
= \mathbb{P}_n \left[ Q_1 \left( \mathcal{T}, \hat{\eta}^{(0)} \right) \right] - \hat{\theta}_{MR,greed} 
\]

where \(\hat{\eta}^{(0)} \equiv \left( \eta^{(0)}_1,\hat{\lambda}^{(0)}_1,\eta^{(0)}_2,\hat{\lambda}^{(0)}_2 \right)\).
Likewise, to arrive at equation (48) we sum equations (51) and (54)

\[
0 = P_n \left( A_1 s_1 (L_1) \left[ \hat{Y}_2^{(1)} - \expit \{ \hat{\tau}_{1, \text{greed}} s_1 (L_1) \} \right] + A_2 \frac{s_1 (L_1)}{\hat{h}_2} \left[ \hat{Y}_3^{(1)} - \expit \{ \hat{\tau}_{2, \text{greed}} s_2 (L_2) + \hat{\lambda}_2^{(1)} s_1 (L_1) \hat{h}_2 \} \right] \right)
\]

\[
= P_n \left( A_1 s_1 (L_1) \left[ y_{2, \eta_{1, \lambda_2}^{(1)}} (A_1 = 1, L_2) - \expit \{ \hat{\tau}_{1, \text{greed}} s_1 (L_1) \} \right] + A_2 \frac{s_1 (L_1)}{\hat{h}_2} \left[ L_3 - \eta_{2, \lambda_2}^{(1)} (A_2 = 1, L_2) \right] \right)
\]

\[
= P_n \left( A_1 s_1 (A_1 = 1, L_1) \left[ A_2 \frac{s_2 (L_2)}{\hat{h}_2} L_3 - \left\{ A_2 \frac{s_2 (L_2)}{\hat{h}_2} \eta_{2, \lambda_2}^{(1)} (A_2, L_2) - y_{2, \eta_{1, \lambda_2}^{(1)}} (A_1, L_2) \right\} \right] - \eta_{1, \text{greed}} (A_1, L_1) \right)
\]

\[
= P_n \left( \pi_1 s_1 (A_1, L_1) \left[ \frac{s_2 (L_2)}{s_2} L_3 - \left\{ \frac{s_2 (L_2)}{s_2} \eta_{2, \lambda_2}^{(1)} (A_2, L_2) - \frac{s_2 (L_2)}{s_2} y_{2, \eta_{1, \lambda_2}^{(1)}} (A_1, L_2) \right\} \right] - \eta_{1, \text{greed}} (A_1, L_1) \right)
\]

\[
= P_n \left( \pi_1 s_1 (A_1, L_1) \left\{ Q_2 \left( \frac{s_2 (L_2)}{s_2}, \frac{s_2 (L_2)}{s_2} \right) - \eta_{1, \text{greed}} (A_1, L_1) \right\} \right)
\]

where \( \frac{s_2}{s_2} = \hat{h}_2 \) and \( \frac{s_2}{s_2} \) \( \equiv \eta_{2, \lambda_2}^{(1)} \).

**Calculation of** \( \hat{\theta}_{\text{reg}} \). (**Weighted iterated regression**).

**Steps for** \( k = 2 \)

(a) Using subjects with \( A_2 = 1 \), compute the weighted logistic regression estimator \( \hat{\tau}_{2, \text{reg}} \) for the outcome \( L_3 \) on the covariate vector \( s_2 (L_2) \) with weight \( 1/\pi_2^2 = 1/\left( \hat{h}_1 \hat{h}_2 \right) \). That is, \( \hat{\tau}_{2, \text{reg}} \) solves

\[
P_n \left\{ A_2 \frac{s_2 (L_2)}{\hat{h}_1 \hat{h}_2} \left[ L_3 - \expit \{ \tau_2 s_2 (L_2) \} \right] \right\} = 0 \quad (57)
\]
(b) For each subject with $A_1 = 1$ compute $\hat{Y}_{2,\text{reg}} \equiv \expit\{\hat{\tau}_{2,\text{reg}}^T s_2 (L_2)\}$.

Steps for $k = 1$

(a) Using subjects with $A_1 = 1$, compute the weighted logistic regression estimator $\hat{\tau}_{1,\text{reg}}$ for the outcome $\hat{Y}_{2,\text{reg}}$ on the covariate vector $s_1 (L_1)$ with weight $1/\hat{\pi}_1 = 1/\hat{h}_1$. That is, $\hat{\tau}_{1,\text{reg}}$ solves

$$\mathbb{P}_n \left[ A_1 \frac{s_1 (L_1)}{\hat{h}_1} \left\{ \hat{Y}_{2,\text{reg}} - \expit \left\{ \hat{\tau}_{1,\text{reg}}^T s_1 (L_1) \right\} \right\} \right] = 0 \quad (58)$$

(b) For all study subjects compute $\hat{Y}_{1,\text{reg}} \equiv \expit\{\hat{\tau}_{1,\text{reg}}^T s_1 (L_1)\}$.

The estimator $\hat{\theta}_{\text{reg}}$ is $\mathbb{P}_n \left( \hat{Y}_{1,\text{reg}} \right)$. As indicated earlier, the estimator $\hat{\theta}_{\text{reg}}$ is multiply robust so long as $s_1 (L_1)$ is a subvector of $s_2 (L_2)$ and 1 is an entry of both $s_1 (L_1)$ and $s_2 (L_2)$. Note that the outcomes $L_3$ and $\hat{Y}_{2,\text{reg}}$ in step (a) of each iteration are bounded between 0 and 1 and consequently, the equations (57) and (58) always have a solution. In addition, because $\hat{Y}_{1,\text{reg}}$ is also bounded between 0 and 1, the estimator $\hat{\theta}_{\text{reg}}$ is guaranteed to fall between 0 and 1.

We now turn to the application of Algorithm 4. As indicated earlier, the algorithm applied to the present example returns precisely the estimator of $\theta (g)$ derived by by Tchetgen-Tchetgen (2009).

Calculation of $\hat{\theta}_{MR}$. (Iterated regression of multiply robust outcomes)

Steps for $k = 2$

(a) Using subjects with $A_2 = 1$, compute the logistic regression estimator $\hat{\tau}_{2,MR}$ for the outcome $L_3$ on the covariate vector $s_2 (L_2)$. That is, $\hat{\tau}_{2,MR}$ solves

$$\mathbb{P}_n \left[ A_2 s_2 (L_2) \left\{ L_3 - \expit \left\{ \hat{\tau}_{2,MR}^T s_2 (L_2) \right\} \right\} \right] = 0 \quad (59)$$

(b) For each subject with $A_1 = 1$ compute

$$\hat{Q}_2 \equiv \frac{A_2}{\hat{h}_2} \left\{ L_3 - \expit \left( \hat{\tau}_{2,MR}^T s_2 (L_2) \right) \right\} + \expit \left( \hat{\tau}_{2,MR}^T s_2 (L_2) \right)$$
Steps for $k = 1$

(a) Using subjects with $A_1 = 1$, compute the logistic regression estimator $\hat{\tau}_{1,MR}$ for the outcome $\hat{Q}_2$ on the covariate vector $s_1(L_1)$. That is, $\hat{\tau}_{1,MR}$ solves

$$\mathbb{P}_n \left[ A_1 s_1 (L_1) \left\{ \hat{Q}_2 - \expit \{ \tau_{1,MR}^T s_1 (L_1) \} \right\} \right] = 0 \quad (60)$$

(b) For all subjects compute

$$\hat{Q}_1 \equiv \frac{A_1}{h_1} \left\{ \hat{Q}_2 - \expit \left( \hat{\tau}_{1,MR}^T s_1 (L_1) \right) \right\} + \expit \left( \hat{\tau}_{1,MR}^T s_1 (L_1) \right)$$

Finally $\hat{\theta}_{MR} = \mathbb{P}_n \left( \hat{Q}_1 \right)$. Note that the outcome $\hat{Q}_2$ in step (a) of the second iteration (i.e. corresponding to $k = 1$), unlike the outcome $\hat{Y}_{2,\text{reg}}$ of the preceding, is not guaranteed to be between 0 and 1 since $1/\hat{h}_2$ can be arbitrarily large. Consequently, it is possible that equation (60) would not have a solution. Even if a solution $\hat{\tau}_{1,MR}$ is found, there is no guarantee that the estimator $\hat{\theta}_{MR}$ would fall between 0 and 1 since $\hat{Q}_1$ can be arbitrarily large.

5 Machine learning $K + 1$ and $2^K$ multiply robust estimators

So far we have considered estimation of the functions $\eta_k^\theta$ and $h_k$ under parametric working models. We will now consider extending some of the estimators in the preceding sections to allow for more flexible estimation of $\eta_k$ and $h_k$ by machine learning algorithms.

Our machine learning estimators will use sample splitting because the true functions $\eta_k^\theta$ and $h_k$, and/or the machine learning estimators of any of them may fail to fall in a Donsker class. We thus randomly split the sample into $U$ equal sized subsamples indexed $u = 1, \ldots, U$, where $U$ is a small fixed number, say 5. We refer to the set of sample units in the $u^{th}$ sample split as the $u^{th}$ validation sample and the set comprised by the remaining sample units as the $u^{th}$ training sample. We let $\mathbb{P}_{n,u}^{v}$ be the empirical distribution of the subjects in the $u^{th}$ validation sample and $\mathbb{P}_{n,u}^{t}$ be the empirical distribution of the subjects in the $u^{th}$ training sample. We consider
machine learning generalizations of earlier doubly robust and multiple robust estimators of

$$\theta (g) \equiv E_{gh^*} \left\{ \psi (L_{K+1}) \right\}$$

$$= \int \psi (L_{K+1}) \prod_{k=1}^{K} h_k^* (a_k | \ell_k, \alpha_{k-1}) \prod_{k=0}^{K} g_k (l_{k+1} | \ell_k, \alpha_k) \, d\mu (z)$$

defined as

$$\hat{\theta}_{DR,CF,mach} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left\{ Q_1 \left( \frac{\pi (t, u)}{h_{mach}}, \frac{\pi (t, u)}{h_{mach}} \right) \right\},$$

$$\hat{\theta}_{DR,CF,mach,bang} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left( y^u_1 \tilde{\eta}_{1, mach,bang}^u (L) \right),$$

$$\hat{\theta}_{DR,CF,mach,reg} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left( y^u_1 \tilde{\eta}_{1, mach, reg}^u (L) \right),$$

$$\hat{\theta}_{MR,CF,mach} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left\{ Q_1 \left( \frac{\pi (t, u)}{h_{mach}}, \frac{\pi (t, u)}{h_{mach}} \right) \right\},$$

$$\hat{\theta}_{MR,CF,mach,bang} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left( y^u_1 \tilde{\eta}_{1, mach,bang}^u (L) \right),$$

$$\hat{\theta}_{MR,CF,mach,reg} = U^{-1} \sum_{u=1}^{U} \mathbb{P}^u \left( y^u_1 \tilde{\eta}_{1, mach, reg}^u (L) \right)$$

where, recall, for any $\overline{h} \equiv (h_1, \ldots, h_K)$ and $\overline{\eta} \equiv (\eta_1, \ldots, \eta_K),$

$$Q_1 (\overline{h}, \overline{\eta}) \equiv \frac{\pi^{*K}}{\pi^K} \psi (L_{K+1}) - \sum_{k=1}^{K} \left\{ \frac{\pi^{*k}}{\pi^k} \eta_k (\overline{A}_k, L_k) - \frac{\pi^{*(k-1)}}{\pi^{(k-1)}} y_{k, \eta_k} (\overline{A}_{k-1}, L_k) \right\},$$

with

$$y_{k, \eta_k} (\overline{A}_{k-1}, L_k) \equiv \int h^*_k (a_k | \overline{A}_{k-1}, L_k) \, d\mu_k (a_k),$$

$$\pi^{*k} \equiv \prod_{j=1}^{k} h^*_j \text{ and } \pi^k \equiv \prod_{j=1}^{k} h_j.$$
and where \( \hat{\eta}_{mach}^t \equiv \left( \hat{\eta}_{1,mach}^{t,u}, \ldots, \hat{\eta}_{K,mach}^{t,u} \right) \) is the output of part (a), \( y_{1,\hat{\eta}_{1,mach}^{t,u}}^u \) and 

\( \hat{\eta}_{mach}^t \equiv \left( \hat{\eta}_{1,mach}^{t,u}, \ldots, \hat{\eta}_{K,mach}^{t,u} \right) \) are outputs of part (b), 

\( \hat{\eta}_{mach}^t \equiv \left( \hat{\eta}_{1,mach}^{t,u}, \ldots, \hat{\eta}_{K,mach}^{t,u} \right) \) is the output of part (c), \( y_{1,\hat{\eta}_{1,mach,bang}^{t,u}}^u \) and \( y_{1,\hat{\eta}_{1,mach,reg}^{t,u}}^u \) are the outputs of step (d), and \( y_{1,\hat{\eta}_{1,mach,bang}^{t,u}}^u \) and \( y_{1,\hat{\eta}_{1,mach,reg}^{t,u}}^u \) are the outputs of step (e) of the following algorithm:

**Algorithm 6. (Cross–Fitting Machine Learning Multiple Robust estimation)**

For \( u = 1, \ldots, U \), in the \( u^{th} \) training sample run steps a)-c) and in the validation sample run steps d) and e)

a) For \( k = K, K - 1, \ldots, 1 \), a preferred machine learning algorithm to estimate \( h_k \).

Let \( \hat{h}_k^{t,u} \) be the output of the algorithm and let \( \hat{\eta}_{mach}^t \equiv \hat{h}_1^{t,u} \times \ldots \times \hat{h}_K^{t,u} \).

b) Set \( \hat{Y}_{K+1,mach}^u = \psi \left( \mathcal{T}_{K+1} \right) \) and for \( k = K, K - 1, \ldots, 1 \), repeat,

b.1) Compute \( \tilde{\eta}_{k,mach}^t (\cdot, \cdot) \), the output of a preferred machine learning algorithm for estimating \( E \left( \hat{Y}_{K+1,mach}^u \mid \mathcal{A}_k, \mathcal{L}_k \right) \).

b.2) For units with \( \pi^{t,k} > 0 \), compute

\[
\hat{Y}_{k,mach}^u = y_{k,\hat{\eta}_{1,mach}^{t,u}}^u (\mathcal{A}_{k-1}, \mathcal{L}_k) \equiv \int h_k^* (a_k | \mathcal{A}_{k-1}, \mathcal{L}_k) \tilde{\eta}_{k,mach}^t (a_k, \mathcal{A}_{k-1}, \mathcal{L}_k) d\mu_k (a_k).
\]

c) Set \( \tilde{Q}_{K+1,mach}^u = \psi \left( \mathcal{L}_{K+1} \right) \) and for \( k = K, K - 1, \ldots, 1 \), repeat

C.1) Compute \( \tilde{\eta}_{k,mach}^t (\cdot, \cdot) \), the output of a preferred machine learning algorithm for estimating \( E \left( \tilde{Q}_{K+1,mach}^u \mid \mathcal{A}_k, \mathcal{L}_k \right) \).

c.2) For units with \( \pi^{t,k} > 0 \), compute

\[
\hat{Y}_{k,mach}^u = y_{k,\tilde{\eta}_{1,mach}^{t,u}}^u (\mathcal{A}_{k-1}, \mathcal{L}_k) \equiv \int h_k^* (a_k | \mathcal{A}_{k-1}, \mathcal{L}_k) \tilde{\eta}_{k,mach}^t (a_k, \mathcal{A}_{k-1}, \mathcal{L}_k) d\mu_k (a_k).
\]

and

\[
\tilde{Q}_{k,mach}^u = \frac{h_k^*}{h_k^{t,u}} \left[ \tilde{Q}_{k+1,mach}^u - \tilde{\eta}_{k,mach}^t (\mathcal{A}_k, \mathcal{L}_k) \right] + \hat{Y}_{k,mach}^u.
\]
d. For units in the validation sample, set $\tilde{Y}_{K+1,mach}^u \equiv \psi \left( T_{K+1} \right)$ and for $k = K, K-1, \ldots, 1$, repeat,

d.1 Based on validation units with $\pi^{*k} > 0$, estimate $\lambda_k$ and $\beta_k$ indexing the regression models $\Psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \lambda_k \left( 1/\pi_{mach}^{(t,u),k} \right) \right\}$ and $\Psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \beta_k \right\}$ for $E \left( \tilde{Y}_{k+1,mach,bang}^u | \tilde{A}_k, \tilde{T}_k \right)$ and $E \left( \tilde{Y}_{k+1,mach,reg}^u | \tilde{A}_k, \tilde{T}_k \right)$, which have offset $\psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right]$, with $\lambda_{k,mach}$ and $\beta_{k,mach}$ solving

$$P_{n}^{u} \left[ \pi^{*k} \left( 1/\pi_{mach}^{(t,u),k} \right) \left[ \tilde{Y}_{k+1,mach,bang}^u - \psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \lambda_k \left( 1/\pi_{mach}^{(t,u),k} \right) \right\} \right] \right] = 0$$

$$P_{n}^{u} \left[ \pi^{*k} \left( 1/\pi_{mach}^{(t,u),k} \right) \left[ \tilde{Y}_{k+1,mach,reg}^u - \psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \beta_k \right\} \right] \right] = 0$$

and set $\tilde{\eta}_{L,mach,bang}^u (\tilde{A}_k, \tilde{T}_k) = \psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \lambda_{k,mach} \left( 1/\pi_{mach}^{(t,u),k} \right) \right\}$ and $\tilde{\eta}_{L,mach,reg}^u (\tilde{A}_k, \tilde{T}_k) = \psi \left\{ \psi^{-1} \left[ \tilde{\eta}_{L,mach}^t \right] + \beta_{k,mach} \right\}$

d.2 For validation sample units with $\pi^{*k-1} > 0$, compute

$$\tilde{Y}_{k,mach,bang}^u \equiv \tilde{y}_{k,mach,bang}^u (\tilde{A}_{k-1}, \tilde{T}_k) \equiv \int h_k^* (a_k | \tilde{A}_{k-1}, \tilde{T}_k) \tilde{\eta}_{k,mach,bang}^u (a_k, \tilde{A}_{k-1}, \tilde{T}_k) d\mu_k (a_k)$$

and

$$\tilde{Y}_{k,mach,reg}^u \equiv \tilde{y}_{k,mach,reg}^u (\tilde{A}_{k-1}, \tilde{T}_k) \equiv \int h_k^* (a_k | \tilde{A}_{k-1}, \tilde{T}_k) \tilde{\eta}_{k,mach,reg}^u (a_k, \tilde{A}_{k-1}, \tilde{T}_k) d\mu_k (a_k)$$

e. For units in the validation sample set $\tilde{Y}_{K+1,mach}^u \equiv \psi \left( T_{K+1} \right)$ and for $k = K, K-1, \ldots, 1$, repeat steps d.1) and d.2) but with $\tilde{\eta}_{L,mach}^t$ replacing $\tilde{\eta}_{L,mach}^t$ and renaming $\tilde{\eta}_{L,mach,bang}$ as $\tilde{\eta}_{L,mach,bang}^u$, $\tilde{\eta}_{L,mach,reg}$ as $\tilde{\eta}_{L,mach,reg}^u$, $\tilde{Y}_{k,mach,bang}$ as $\tilde{Y}_{k,mach,bang}^u$, and $\tilde{Y}_{k,mach,reg}$ as $\tilde{Y}_{k,mach,reg}^u$.\noindent
56
5.1 Asymptotic theory of cross-fitting estimators: preliminary background

To study the asymptotic properties of the above estimators we will now formulate a general and unified notation. Given a random sample $S = \{Z_1, \ldots, Z_n\}$ comprised of $n$ i.i.d. copies of a random vector $Z$ from an unknown law $P$ with density $p$ with respect to an underlying measure, and given two subsamples $S_t$ and $S_v$ of $S$, let $P^n_j$ denote the empirical distribution of sample $S_j$, $j = t, v$. Let $m \left( z, r^t \right)$ be a given function of $z$ and $r^t$ where $r^t \equiv r^t(\cdot) : z \rightarrow r^t(z)$ is some map on the sample space of $Z$, and let $\mu (P) \equiv E_P \left[ m(O, r(P)) \right]$ where $r(P)(\cdot) : z \mapsto r(P)(z)$ is a map that depends on $P$. Define $M(r) \equiv m(Z, r)$.

Consider an estimator $\hat{\mu} \equiv P^n_v[M(\hat{r}^t)] \equiv P^n_v[m(Z, \hat{r}^t)]$ of $\mu (P)$ that depends on $\hat{r}^t \equiv \hat{r}(P^n_t)(\cdot)$, an estimator of $r(P)(\cdot)$ based on data from $S_t$. Consider the decomposition

$$\begin{align*}
P^n_v[M(\hat{r}^t)] - \mu (P) &= P^n_v[M(\hat{r}^t)] - E^n_v[M(\hat{r}^t)] + E^n_v[M(\hat{r}^t)] - \mu (P)
\end{align*}$$

where throughout $E^n_v[\cdot]$ stands for the population expectation operator that regards the data from $S_t$ as fixed, i.e. non-random, e.g. $E^n_v[M(\hat{r}^t)]$ stands for $E_P[M(r)]|_{r=\hat{r}^t} \equiv \int m(z, \hat{r}^t) \, dP(z)$. Note that $E^n_v[M(\hat{r}^t)]$ is random as it depends on the data from the random sub-sample $S_t$.

We refer to $E^n_v[M(\hat{r}^t)] - \mu (P)$ as the drift. We refer to $P^n_v[M(\hat{r})] - E^n_v[M(\hat{r}^t)]$ as the centered term.

Consider now estimation of $\mu (P)$ by sample splitting. Specifically, randomly partition the sample into $U$ equal sized subsamples indexed by $u = 1, \ldots, U$, where $U$ is a small fixed number. For a given $u$, let $S_{v,u}$ denote the set of sample units in the $u^{th}$ partition. Call $S_{v,u}$ the $u^{th}$ validation sample. Let $S_{t,u}$ denote the set comprised by the remaining sample units, call it the $u^{th}$ training sample. As before, we let $P^n_{j,u}$ be the empirical distribution of the data in $S_{j,u}$, $j = v, t$.

The split-specific estimator of $\mu (P)$ is given by

$$\hat{\mu} (\hat{r}^t) = P^n_v[M(\hat{r}^t)]$$

where $\hat{r}^t = \hat{r}(P^n_t)(\cdot)$ is some estimator of $r(P)(\cdot)$ based on data in the training sample and where, by convention, we eliminate the superscript $u$ when we refer to a single split. One of our goals in this section is to study the asymptotic properties of the cross-fitting (CF) estimator $\hat{\mu}^{CF}$ obtained as the average of estimators $\hat{\mu} (\hat{r}^t)$ over
all \( \mathbf{U} \) validation samples, that is,

\[
\hat{\mu}^{cf} = \mathbf{U}^{-1} \sum_{u=1}^{\mathbf{U}} \mathbb{P}_{n,u}^{v} [M (\hat{r}^{t,u})]
\]

Consider now a single split. Henceforth, we suppose that there exists a bounded function \( r^* (P) (\cdot) \), not necessarily equal to \( r (P) (\cdot) \), such that for \( \hat{r} = \hat{r} (\mathbb{P}_n) (\cdot) \) it holds that

\[
\int [m (z, \hat{r}^t) - m (z, r^*)]^2 dP (z) \to 0 \text{ as } n \to \infty
\]

Then, with \( n_v \) denoting the cardinality of \( S_v \),

\[
\sqrt{n_v} \{ \mathbb{P}_n [M (\hat{r}^{t})] - E_v [M (\hat{r}^{t})] \} = \sqrt{n_v} \{ \mathbb{P}_n [M (r^*)] - E_v [M (r^*)] \} + o_P (1)
\]

as \( n \to \infty \) as is well known (see van der Vaart, 1998).

Hence as \( n \to \infty \), we have

\[
\sqrt{n_v} \{ \mathbb{P}_n [M (\hat{r}^{t})] - \mu (P) \} = \sqrt{n_v} \{ \mathbb{P}_n [M (r^*)] - E_v [M (r^*)] \} + \sqrt{n_v} \{ P_v [M (\hat{r}^{t})] - \mu (P) \} + o_P (1),
\]

Thus if \( E_v [M (\hat{r}^{t})] - \mu (P) = o_P (1/\sqrt{n_v}) \), we can conclude that \( \hat{\mu} (\hat{r}^{t}) = \mathbb{P}_n [M (\hat{r}^{t})] \) is an asymptotically linear estimator of \( \mu (P) \), and thus, since \( \mathbb{P}_n [M (r^*)] = \mathbf{U}^{-1} \sum_{u=1}^{\mathbf{U}} \mathbb{P}_{n,u}^{v} [M (r^*)] \), we conclude that \( \hat{\mu}^{cf} \) is an asymptotically linear estimator of \( \mu (P) \) with influence function \( M (r^*) \). That is, as \( n \) goes to \( \infty \)

\[
\sqrt{n} \{ \hat{\mu}^{cf} - \mu (P) \} = \sqrt{n} \mathbb{P}_n [M (r^*)] + o_P (1).
\]

5.2 Analysis of the drifts of the machine learning DR and MR estimators

We will now apply the generic formulation of the preceding section to compare the distribution of machine learning doubly robust and multiply robust estimators of \( \theta (g) \). To do so, we begin by comparing the asymptotic properties of \( \hat{\theta}_{DR,CF,mach} \) and \( \hat{\theta}_{MR,CF,mach} \).
First note that $\theta (g) = P \left[ Q_1 (\tilde{h}, \tilde{\eta}) \right]$, so that in our general formulation of asymptotic theory we identify $r (P)$ with $(\tilde{h}, \tilde{\eta})$ and for any $r^\dagger = (\tilde{h}^\dagger, \tilde{\eta}^\dagger)$ we define

$$m (Z; r^\dagger) = M (r^\dagger)$$

$$= Q_1 (\tilde{h}^\dagger, \tilde{\eta}^\dagger)$$

$$= \frac{\pi^* K}{\pi^!} \psi (\tilde{L}_{K+1}) - \sum_{k=1}^K \left\{ \frac{\pi^* k}{\pi^!} \eta_k^\dagger (\tilde{A}_k, \tilde{L}_k) - \frac{\pi^* (k-1)}{\pi^! (k-1)} y_k \tilde{\eta}_k^\dagger (\tilde{A}_{k-1}, \tilde{L}_k) \right\}$$

The estimator $\tilde{\theta}_{DR,CF,mach}$ is the average of $\mathbb{P}_{n}^v \left\{ Q_1 \left( \tilde{h}^{(t,u)}, \tilde{\eta}_{mach}^{(t,u)} \right) \right\}$ over $u = 1, \ldots, U$, so $\mathbb{P}_{n}^v [M (\tilde{r}^\dagger)]$ is just a split specific

$$\tilde{\theta}_{DR,mach} = \mathbb{P}_{n}^v \left\{ Q_1 \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) \right\}$$

where, recall that by convention we eliminate the superscript $u$ when referring to a generic split.

Likewise, when studying the limit law of $\tilde{\theta}_{MR,CF,mach}$, $\mathbb{P}_{n}^v [M (\tilde{r}^\dagger)]$ is equal to

$$\tilde{\theta}_{MR,mach} = \mathbb{P}_{n}^v \left\{ Q_1 \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) \right\}$$

We are interested in investigating the rates of convergence to 0 of the drifts of $\tilde{\theta}_{DR,CF,mach}$ and $\tilde{\theta}_{MR,CF,mach}$. In view of the discussion of the preceding section, it suffices to study the rates of the drifts of the single split estimators $\tilde{\theta}_{DR,mach}$ and $\tilde{\theta}_{MR,mach}$. Notice that these drifts are $E^v \left[ Q_1 \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) \right] = \theta (g)$ and

$$E^v \left[ Q_1 \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) \right] = \theta (g)$$

which, by Lemma 3, can be expressed as

$$a^p (h^\dagger, \eta^\dagger) = b^p (h^\dagger, \eta^\dagger) = c^p (h^\dagger, \eta^\dagger) \text{ evaluated at } \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) \text{ or } \left( \tilde{h}^\dagger, \tilde{\eta}_{mach}^\dagger \right) .$$

We will exploit these formulae appropriately to make manifest the difference in the orders of the drifts of $\tilde{\theta}_{DR,mach}$ and $\tilde{\theta}_{MR,mach}$.
Using \( c^p (h^t, \eta^t) \) applied to \( (h^t, \eta^t) = \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \) we obtain the following expression for the drift of \( \hat{\theta}_{DR,mach} \)

\[
E^v \left[ Q_1 \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \right] - \theta (g) = \sum_{k=1}^{K} E^v \left[ \frac{\pi^{*k}}{\pi_k} \left\{ \frac{h_k^*}{h_k} - \frac{\hat{h}_k^*}{\hat{h}_k} \right\} \hat{\eta}_{mach}^t - E_{g_k} \left\{ y_{k+1,\hat{\eta}_{k+1,mach}^t} \left( A_k, L_{k+1} \right) \left| A_k, L_k \right\} \right\} \right]
\]

where \( y_{k+1,\hat{\eta}_{k+1,mach}^t} \left( A_K, L_{K+1} \right) \equiv \psi \left( T_{K+1} \right) \). Using the identity for any \( k \in [K] \) and any \( (\hat{h}_1, ..., \hat{h}_K) \),

\[
\frac{\pi^{*k}}{\pi_k} - \frac{\hat{\pi}^{*k}}{\hat{\pi}_k} = \sum_{j=1}^{k} \frac{\pi^{*j-1}}{\pi_j - 1} \left\{ \frac{h_j^*}{h_j} - \frac{\hat{h}_j^*}{\hat{h}_j} \right\} \frac{\pi^{*K}}{\hat{\pi}_j^+}
\]

we arrive at

\[
E^v \left[ Q_1 \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \right] - \theta (g) = \sum_{k=1}^{K} E^v \left[ \frac{\pi^{*K}}{\pi_k} \left\{ \frac{h_k^*}{h_k} - \frac{\hat{h}_k^*}{\hat{h}_k} \right\} \hat{\eta}_{mach}^t - E_{g_k} \left\{ y_{k+1,\hat{\eta}_{k+1,mach}^t} \left( A_k, L_{k+1} \right) \left| A_k, L_k \right\} \right\} \right]
\]

Likewise, using the formula \( \psi^p (h^t, \eta^t) \) applied to \( (h^t, \eta^t) = \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \) we obtain the following expression for the drift of \( \hat{\theta}_{MR,mach} \)

\[
E^v \left[ Q_1 \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \right] - \theta (g) = \sum_{k=1}^{K} E^v \left[ \frac{\pi^{*(k-1)}}{\pi_{(k-1)}} \left\{ \frac{h_k^*}{h_k} - \frac{\hat{h}_k^*}{\hat{h}_k} \right\} \hat{\eta}_{mach}^t - E_{g_k} \left\{ \tilde{Q}_{k+1,mach} \left| A_k, L_k \right\} \right\} \right]
\]

with \( \tilde{Q}_{K+1,mach} \equiv \psi \left( T_{K+1} \right) \).
Note that \( \tilde{\eta}_{k,\text{mach}} \) is equal to the residual
\[
\tilde{E}_{\text{mach}}^t \left\{ y_{k+1,\tilde{\eta}_{k+1,\text{mach}}} (\overline{A}_k, \overline{L}_{k+1}) \right\} - E_gk \left\{ y_{k+1,\tilde{\eta}_{k+1,\text{mach}}} (\overline{A}_k, \overline{L}_{k+1}) \right\}
\]
and \( \tilde{\eta}_{k,\text{mach}} \) is equal to the residual
\[
\tilde{E}_{\text{mach}}^t \left\{ \tilde{Q}_{k+1,\text{mach}} (\overline{A}_k, \overline{L}_{k+1}) \right\} - E_gk \left\{ \tilde{Q}_{k+1,\text{mach}} (\overline{A}_k, \overline{L}_{k+1}) \right\}
\]

where for any \( W = w (\overline{A}_k, \overline{L}_{k+1}) \), \( E_{\text{mach}}^t \{ W \mid \overline{A}_k, \overline{L}_k \} \) is the machine learning estimator of the true expectation \( E_{gk} \{ W \mid \overline{A}_k, \overline{L}_k \} \). Note also that had we used the expression \( b^p \left( h^\dagger, \eta^\dagger \right) \) to represent the drift of \( \tilde{\theta}_{DR,\text{mach}} \), this would have resulted in an expression involving the differences \( \tilde{\eta}_{k,\text{mach}} - E_{gk} \left\{ \tilde{Q}_{k+1,\text{mach}} \mid \overline{A}_k, \overline{L}_k \right\} = \tilde{E}_{\text{mach}}^t \left\{ y_{k+1,\tilde{\eta}_{k+1,\text{mach}}} (\overline{A}_k, \overline{L}_{k+1}) \right\} - E_gk \left\{ y_{k+1,\tilde{\eta}_{k+1,\text{mach}}} (\overline{A}_k, \overline{L}_{k+1}) \right\} \) with \( \tilde{Q}_{k+1} \) defined iteratively for \( k = K - 1, K - 2, \ldots, 0 \), as
\[
\tilde{Q}_{k+1,\text{mach}} = \frac{h_{k+1}^\dagger}{h_k^\dagger} \left[ \tilde{Q}_{k+2,\text{mach}} - \tilde{\eta}_{k+1,\text{mach}} (\overline{A}_{k+1}, \overline{L}_{k+1}) \right] + \tilde{Y}_{k+1,\text{mach}}.
\]

Because these differences are not the residuals from applying the machine learning algorithm to the outcome \( y_{k+1,\tilde{\eta}_{k+1,\text{mach}}} (\overline{A}_k, \overline{L}_{k+1}) \), using the expression \( c^p \left( h^\dagger, \eta^\dagger \right) \) to represent the drift of \( \tilde{\theta}_{DR,\text{mach}} \) would have made the structure of the drift less transparent. Likewise, a similar situation would arise if we use the expression \( c^p \left( h^\dagger, \eta^\dagger \right) \) to represent the drift of \( \tilde{\theta}_{MR,\text{mach}} \).

Although the drift of \( \tilde{\theta}_{DR,\text{mach}} \) has many more terms than the drift of \( \tilde{\theta}_{MR,\text{mach}} \), at this level of generality it does not seem possible to quantitatively compare the size of the drifts of the two estimators when the precise machine learning algorithm being used is not further specified. In the special case that the machine learning algorithm is a linear operator, direct and easily interpretable comparisons become possible. These are discussed in the next subsection. In particular, we will argue that if \( \eta_k^o \) and the true \( h_k \) lie in specific smoothness classes, then we can quantify the rates of convergence of the drifts to zero. Our analysis relies on a specific representation for \( a^p \left( h, \eta \right) \), given in the next subsection, when \( \eta = (\eta_1, \ldots, \eta_K) \) takes two special forms which mimic the forms that the estimators \( \tilde{\eta}_{k+1,\text{mach}} \) and \( \tilde{\eta}_{k+1,\text{mach}} \) take when the machine learning algorithms used are linear operators.
5.2.1 Analysis when the ML algorithms used are linear operators.

We will now argue that if \( \eta_k^a \) and the true \( h_k \) lie in specific smoothness classes, then we can quantify the rates of convergence of the drifts to zero. Our analysis relies on a specific representation for \( a^p(\ h, \eta \) ), given in the next Theorem, when \( \eta = (\eta_1, ..., \eta_K) \) takes two special forms which mimic the forms that the estimators \( \hat{\eta}_{k+1, mach} \) and \( \tilde{\eta}_{k+1, mach} \) take when the machine learning algorithms used are linear operators. To state the Theorem, we must first define a number of objects, which we now do.

Given \( h^\dagger = (h^\dagger_1, ..., h^\dagger_K) \), define for \( 0 \leq j < u \leq K \),

\[
\nabla_{j,u} \equiv \frac{\pi_{j+1}^{u-1}}{\pi_{j+1}^u} \left( \frac{h^*_u}{h^*_j} - \frac{h^*_j}{h^*_u} \right)
\]

Given linear operators \( \Pi^j[\cdot] : L_2(Q_j) \to L_2(P_j), j \in [K] \), where \( Q_j \) and \( P_j \) are the laws of \( (A_j, L_{j+1}) \) and \( (A_j, L_j) \) respectively, we define the following operators

1. for \( j = 1, ..., K - 1 \),

\[
\Pi_{DR}^j[\cdot] = \Pi^j \left\{ E_p \left( \frac{h^*_j}{h^*_j} \cdot \left| A_j, L_{j+1} \right| \right) \right\}
\]

2. for \( 1 \leq j < k \leq K - 1 \),

\[
\Pi_{DR,j,k}^j[\cdot] = \Pi_{DR}^j \circ \cdots \circ \Pi_{DR}^{k-1}[\cdot]
\]

where \( \circ \) denotes the composition operation. Note that \( \Pi_{DR,j,j+1}[\cdot] = \Pi_{DR}^j[\cdot] \).

3. for \( 1 \leq j < u \leq K \),

\[
\Pi_{MR,j,u}^j[\cdot] \equiv \Pi^j \left[ E_p \left( \nabla_{j,u} \cdot \left| A_j, L_{j+1} \right| \right) \right]
\]

Note that

\[
\Pi_{MR,j,j+1}^j[\cdot] = \Pi^j \left[ E_p \left\{ \left( \frac{h^*_j}{h^*_j} - \frac{h^*_j}{h^*_j} \right) \cdot \left| A_j, L_{j+1} \right| \right\} \right]
\]
4. For $1 \leq r_1 < r_2 < \ldots < r_u \leq K$, 
\[ \Pi_{MR,r_1,r_2,\ldots,r_u}[\cdot] \equiv \Pi_{MR,r_1,r_2} \circ \ldots \circ \Pi_{MR,r_{u-1},r_u} \circ \Pi_{MR,r_{u-1},r_u}[\cdot] \]

Next, define the following random variables:

a. for $j = 1,\ldots,K$, define 
\[ \eta_{j,DR} \equiv \eta_{j,DR}(\overline{A}_j, \overline{L}_j) \equiv \Pi^j \left[ y_{j+1,\eta_{j+1}} (\overline{A}_j, \overline{L}_{j+1}) \right] \]

b. for $j = K, K - 1,\ldots,1$, recursively define 
\[ \tilde{\eta}_{j,DR} \equiv \tilde{\eta}_{j,DR}(\overline{A}_j, \overline{L}_j) \equiv \Pi^j \left[ y_{j+1,\tilde{\eta}_{j+1,DR}} (\overline{A}_j, \overline{L}_{j+1}) \right] \]

c. Given $h^\dagger = \left( h_1^\dagger, \ldots, h_K^\dagger \right)$, for $j = K, K - 1,\ldots,1$, recursively define 
\[ \tilde{\eta}_{j,MR} \equiv \tilde{\eta}_{j,MR}(\overline{A}_j, \overline{L}_j) \equiv \Pi^j \left[ Q_{j+1} \left( \overline{h}_{j+1,K}^{\dagger}, \overline{\eta}_{j+1,MR}^{\dagger} \right) \right] \]
where $Q \left( \overline{h}_{K+1,K}^{\dagger}, \overline{\eta}_{K+1,MR}^{\dagger} \right) \equiv \psi(\overline{L}_{K+1})$.

d. Given $h^\dagger = \left( h_1^\dagger, \ldots, h_K^\dagger \right)$, for $j = K, K - 1,\ldots,1$, recursively define 
\[ \eta_{j,MR} \equiv \eta_{j,MR}(\overline{A}_j, \overline{L}_j) \equiv \eta_{j,DR} + \Pi^j \left[ Q_{j+1} \left( \overline{h}_{j+1,K}^{\dagger}, \overline{\eta}_{j+1,MR}^{\dagger} \right) - E_p \left\{ Q_{j+1} \left( \overline{h}_{j+1,K}^{\dagger}, \overline{\eta}_{j+1,MR}^{\dagger} \right) | \overline{A}_j, \overline{L}_{j+1} \right\} \right] . \]

The following Theorem gives special representations for $a^p(h, \eta)$ when $\eta = \tilde{\eta}_{DR}$ and $\eta = \tilde{\eta}_{DR}$.

**Theorem 1.** Let $\tilde{\eta}_{DR} \equiv (\tilde{\eta}_{1,DR},\ldots,\tilde{\eta}_{K,DR})$ and $\tilde{\eta}_{DR} \equiv (\tilde{\eta}_{1,DR},\ldots,\tilde{\eta}_{K,DR})$ where $\tilde{\eta}_{j,DR}$ and $\tilde{\eta}_{j,DR}$, $j \in [K]$ are the random variables defined in (b) and (c) above and $h^\dagger \equiv (h_1^\dagger, \ldots, h_K^\dagger)$, with $h_k^\dagger$ an arbitrary density for the law of $A_k$ given $(\overline{A}_{k-1}, \overline{L}_k)$, $k \in [K]$. The following identities hold.
1. for $k \in [K]$

$$
\tilde{\eta}_{k,DR} - \eta^g_k = \eta_{k,DR} - \eta^g_k + \sum_{j=k+1}^{K} \Pi_{DR,k,j} \left[ \eta_{j,DR} - \eta^g_j \right]
$$  \hspace{1cm} (63)

2.

$$
a^p(h^\dagger, \tilde{\eta}_{DR}) \equiv \sum_{k=1}^{K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\tilde{\eta}_{k,DR} - \eta^g_k) \right\}

= \sum_{k=1}^{K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\eta_{k,DR} - \eta^g_k) \right\}

+ \sum_{1 \leq k < j \leq K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) \Pi_{DR,k,j} \left[ \eta_{j,DR} - \eta^g_j \right] \right\}
$$

3.

$$
\sum_{k=1}^{K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\tilde{\eta}_{k,MR} - \eta^g_k) \right\} \left| L_1 \right.

= \sum_{k=1}^{K} \mathbb{E}_p \left\{ \nabla_{0,k} (\eta_{k,MR} - \eta^g_k) \right\} \left| L_1 \right.

+ \sum_{\emptyset \neq \{r_1, \ldots, r_u\} \subseteq \{K-1\}} \sum_{k=r_u+1}^{K} \mathbb{E}_p \left( \nabla_{0,r_1} \Pi_{MR,r_1,r_2,\ldots,r_u,k} \left[ \eta_{k,MR} - \eta^g_k \right] \right) \left| L_1 \right.
$$  \hspace{1cm} (64)

4.

$$
a^p(h^\dagger, \tilde{\eta}_{MR}) \equiv \sum_{k=1}^{K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\tilde{\eta}_{k,MR} - \eta^g_k) \right\}

= \sum_{k=1}^{K} \mathbb{E}_p \left\{ \frac{\pi^{*k-1}}{\pi^{k-1}} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\eta_{k,MR} - \eta^g_k) \right\}

+ \sum_{\emptyset \neq \{r_1, \ldots, r_u\} \subseteq \{K-1\}} \sum_{k=r_u+1}^{K} \mathbb{E}_p \left( \nabla_{0,r_1} \Pi_{MR,r_1,r_2,\ldots,r_u,k} \left[ \eta_{k,MR} - \eta^g_k \right] \right)
$$

64
Notational remark: in parts (3) and (4) of the Theorem, the summation
\[ \sum_{\emptyset \neq \{r_1, ..., r_u\} \subseteq [K-1]} \]
denote the ordered elements of a subset with cardinality \( u \) with \( r_1 < r_2 < ... < r_u \).

In the special case in which \( K = 2 \), assertions (2) and (4) of the Theorem reduce to

\[ a^p(h^\dagger, \hat{\eta}_{DR}) \equiv E_p \left\{ \frac{\pi^{s_1}}{\pi^{s_1}} \left( \frac{h_2^*}{h_2} - \frac{h_2^*}{h_2^*} \right) \left( \eta_{2,DR} - \eta_{2}^q \right) \right\} + E_p \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1^*}{h_1^*} \right) \left( \eta_{1,DR} - \eta_{1}^q \right) \right\} + E_p \left\{ \frac{h_1^*}{h_1} \Pi^1 \left[ E_p \left\{ \frac{h_2^*}{h_2} \left( \eta_{2,DR} - \eta_{2}^q \right) | A_1, \bar{T}_2 \right\} \right] \right\} \] (65)

and

\[ a^p(h^\dagger, \hat{\eta}_{MR}) \equiv E_p \left\{ \frac{\pi^{s_1}}{\pi^{s_1}} \left( \frac{h_2^*}{h_2} - \frac{h_2^*}{h_2^*} \right) \left( \eta_{2,MR} - \eta_{2}^q \right) \right\} + E_p \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1^*}{h_1^*} \right) \left( \eta_{1,MR} - \eta_{1}^q \right) \right\} + E_p \left\{ \frac{h_1^*}{h_1} \Pi^1 \left[ E_p \left\{ \left( \frac{h_2^*}{h_2} - \frac{h_2^*}{h_2^*} \right) \left( \eta_{2,MR} - \eta_{2}^q \right) | A_1, \bar{T}_2 \right\} \right] \right\} \] (66)

When \( K = 3 \), these formulae are

\[ a^p(h^\dagger, \hat{\eta}_{DR}) \equiv \sum_{k=1}^{3} E_p \left\{ \frac{\pi^{s_k-1}}{\pi^{s_k-1}} \left( \frac{h_k^*}{h_k} - \frac{h_k^*}{h_k^*} \right) \left( \eta_{k,DR} - \eta_{k}^q \right) \right\} + E_p \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1^*}{h_1^*} \right) \Pi^1 \left[ E_p \left\{ \left( \frac{h_2^*}{h_2} - \frac{h_2^*}{h_2^*} \right) \left( \eta_{2,DR} - \eta_{2}^q \right) | A_1, \bar{T}_2 \right\} \right] \right\} \] (67)
and

\[ a_p\left( \hat{L}^t, \tilde{\eta}_{MR} \right) \equiv \sum_{k=1}^{3} E_p \left[ \frac{\pi^{*k-1}}{\pi^{*k-1}} \left( \frac{h^*_k - h^*_k}{h^*_k} \right) (\eta_{k,MR} - \eta_3^0) \right] \]

+ \left[ E_p \left\{ \left( \frac{h^*_3 - h^*_3}{h^*_3} \right) (\eta_{3,MR} - \eta_3^0) \left| A_2, T_3 \right\} \right\} \right] \}ight]

\[ \left[ E_p \left\{ \left( \frac{h^*_3 - h^*_3}{h^*_3} \right) (\eta_{3,MR} - \eta_3^0) \left| A_1, T_2 \right\} \right\} \right]\}

Theorem 1 can be applied to quantify the rates of convergence of the drifts of \( \tilde{\eta}_{DR,mach} \) and \( \tilde{\eta}_{MR,mach} \) when the machine learning algorithm used is a linear operator. Here we will apply it to the special case in which the machine learning algorithms used are series estimators. For any \( k \in [K] \), let \( S_k = s_k (\overline{A}_k, \overline{T}_k) \) be a vector valued function of \((\overline{A}_k; \overline{T}_k)\) of dimension \( m_k = m_k (n) \) which depends on the sample size \( n \). For a given \( s_k (\overline{A}_k, \overline{T}_k) \in L_2 (P_k) \), define the operator \( \Pi_1^n : L_2 (Q_k) \rightarrow L_2 (P_k) \)

\[ \Pi_1^n [f_k] \equiv \tilde{\beta} s_k (\cdot) \]

where \( \tilde{\beta} = \mathbb{P}^{t_n} \left[ f_k (\overline{A}_k, \overline{T}_{k+1}) s_k (\overline{A}_k, \overline{T}_k) \right] \mathbb{P}^{t_n} \left[ s_k (\overline{A}_k, \overline{T}_k) s_k (\overline{A}_k, \overline{T}_k)^T \right]^{-1} \) is the least squares coefficient in the regression of \( f_k (\overline{A}_k, \overline{T}_{k+1}) \) on \( s_k (\overline{A}_k, \overline{T}_k) \) in the training sample \( S_t \).

Notice that \( \tilde{\eta}_{k,DR} \) defined in (b) above coincides with the estimator \( \tilde{\eta}_{k,mach}^{t,u} \) from step (c) of Algorithm 6 when the machine learning algorithm is linear regression on \( S_k \) and the linear operator \( \Pi^{t} \) is \( \Pi_1^n \). Likewise, \( \tilde{\eta}_{k,MR} \) defined in (c) above coincides with the estimator \( \tilde{\eta}_{k,mach}^{t,u} \) of step (d) of Algorithm 6. We can then apply the formula in part (2) of Theorem 1 evaluated at \( \tilde{\eta}_{k,DR} = \tilde{\eta}_{k,mach}^{t,u} \) to compute the drift of \( \tilde{\eta}_{DR,mach} \) and the formula in part (4) of Theorem 1 evaluated at \( \tilde{\eta}_{k,MR} = \tilde{\eta}_{k,mach}^{t,u} \) to compute the drift of \( \tilde{\eta}_{MR,mach} \). We will now do so in the special cases \( K = 2 \) and
$K = 3$. This will illustrate and clarify the relationship between the drifts of $\hat{\theta}_{DR, mach}$ and $\hat{\theta}_{MR, mach}$ without unduly complicating the notation.

Using arguments analogous to those in the sections dealing with parametric nuisance models, it can be shown that the drifts of the estimators $\hat{\theta}_{DR,CF,mach,bang}^{,\theta}$ and $\hat{\theta}_{DR,CF,mach,reg}^{,\theta}$ have the same rate of convergence to 0 as the drift of $\hat{\theta}_{DR,CF,mach}^{,\theta}$, and the drifts of $\hat{\theta}_{MR,CF,mach,bang}^{,\theta}$ and $\hat{\theta}_{MR,CF,mach,reg}^{,\theta}$ have the same rate of convergence as $\hat{\theta}_{MR,CF,mach}^{,\theta}$. Hence, we will restrict our discussion to the analysis of $\hat{\theta}_{DR,CF,mach}^{,\theta}$ and $\hat{\theta}_{MR,CF,mach}^{,\theta}$.

In what follows we let $E_{v^{,}(A_k, L_{k+1})}^{,(\cdot)}$ denote the conditional expectation given $\mathbf{v}^{,}(A_k, L_{k+1})$ operator, regarding the data in the training sample as fixed, i.e. non-random, e.g.

$$E_{v^{,}(A_2, L_3)}^{,(\cdot)} \left\{ \left( \frac{h_3}{h_3^*} - \frac{h_3^*}{h_3} \right) (\tilde{\eta}_{3, MR} - \eta_3^{,g}) \right\} = \int \left( \frac{h_3}{h_3^*} (a_3 | \bar{A}_2, Q_3) - \frac{h_3^*}{h_3} (a_3 | \bar{A}_2, Q_3) \right) (\tilde{\eta}_{3, MR} (a_3, \bar{A}_2, L_3) - \eta_3^{,g} (a_3, \bar{A}_2, L_3)) h (a_3 | \bar{A}_2, L_3) da_3$$

For $K = 2$, applying the formula (65) with $\tilde{\eta}_{k, DR}^{,\theta} = \tilde{\eta}_{k, mach}^{,\theta}$ and $\Pi^k = \Pi^k_{n,k}$, we conclude that the drift $E^{v} \left[ Q_1 \left( \hat{h}^{,\theta}, \hat{\eta}_{mach}^{,\theta} \right) \right] - \theta (g)$ of $\hat{\theta}_{DR, mach}^{,\theta}$ is

$$a^{p} \left( \hat{h}^{,\theta}, \hat{\eta}_{mach}^{,\theta} \right) = E^{v} \left\{ \pi^{*1} \left( \frac{h_2^*}{h_2} - \frac{h_2}{h_2^*} \right) (\eta_{2, DR} - \eta_2^{,g}) \right\} + E^{v} \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1}{h_1^*} \right) (\eta_{1, DR} - \eta_1^{,g}) \right\} + E^{v} \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1}{h_1^*} \right) \Pi^k_{n,1} \left\{ E^{v_{, (A_1, L_2)}} \left\{ \frac{h_2^*}{h_2} (\eta_{2, DR} - \eta_2^{,g}) \right\} \right\} \right\}$$

and applying the formula (66) with $\tilde{\eta}_{k, MR}^{,\theta} = \tilde{\eta}_{k, mach}^{,\theta}$ and $\Pi^k = \Pi^k_{n,k}$, the drift $E^{v} \left[ Q_1 \left( \hat{h}^{,\theta}, \hat{\eta}_{mach}^{,\theta} \right) \right] - \theta (g)$ of $\hat{\theta}_{MR, mach}^{,\theta}$ is

67
\( a^p \left( \frac{\xi}{h}, \frac{\hat{\eta}}{\eta_{mach}} \right) \equiv E^v \left\{ \frac{\pi^{*1}}{\pi} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h_l} \right) (\eta_{2,MR} - \eta^g_2) \right\} \\
+ E^v \left\{ \left( \frac{h^*_1}{h_1} - \frac{h^*_2}{h^*_2} \right) (\eta_{1,MR} - \eta^g_1) \right\} \\
+ E^v \left\{ \left( \frac{h^*_1}{h_1} - \frac{h^*_3}{h^*_3} \right) \Pi_{n,1} \left\{ E^v,(A_1,T_2) \left\{ \left( \frac{h^*_2}{h^*_2} - \frac{h^*_2}{h^*_2} \right) (\eta_{2,MR} - \eta^g_2) \right\} \right\} \right\} \}

For \( K = 3 \), formula (67) applied to \( \hat{\eta}_{k,DR} = \hat{\eta}_{k,mach}^t \) and \( \Pi^k = \Pi_{n,k}^t \) implies that the drift \( E^v \left[ Q \left( \frac{\xi}{h}, \frac{\hat{\eta}}{\eta_{mach}} \right) - \theta \left( g \right) \right] \) of \( \hat{\theta}_{DR,mach} \) is

\[ a^p \left( \frac{\xi}{h}, \frac{\hat{\eta}}{\eta_{mach}} \right) \equiv \sum_{k=1}^{3} E^v \left( \delta^{DR}_k \right) + \sum_{1 \leq k < j \leq 3} E^v \left( \xi^{DR}_{k,j} \right) \quad (69) \]

and formula (68) applied to \( \hat{\eta}_{k,MR} = \hat{\eta}_{k,mach}^t \) and \( \Pi^k = \Pi_{n,k}^t \) implies that the drift \( E^v \left[ Q \left( \frac{\xi}{h}, \frac{\hat{\eta}}{\eta_{mach}} \right) - \theta \left( g \right) \right] \) of \( \hat{\theta}_{MR,mach} \) is

\[ a^p \left( \frac{\xi}{h}, \frac{\hat{\eta}}{\eta_{mach}} \right) = \sum_{k=1}^{3} E^v \left( \delta^{MR}_k \right) + \sum_{1 \leq k < j \leq 3} E^v \left( \xi^{MR}_{k,j} \right) + E^v \left( \xi^{MR}_{1,2,3} \right) \quad (70) \]

where

\[ \delta^{DR}_k \equiv \frac{\pi^{*k-1}}{\pi} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\eta_{k,DR} - \eta^g_k) , \]

\[ \delta^{MR}_k \equiv \frac{\pi^{*k-1}}{\pi} \left( \frac{h^*_k}{h_k} - \frac{h^*_l}{h^*_l} \right) (\eta_{k,MR} - \eta^g_k) , \]

\[ \xi^{DR}_{1,2} \equiv E^v \left\{ \left( \frac{h^*_1}{h_1} - \frac{h^*_2}{h^*_2} \right) \Pi_{n,1} \left\{ E^v,(A_1,T_2) \left\{ \frac{h^*_2}{h^*_2} (\eta_{2,DR} - \eta^g_2) \right\} \right\} \right\} , \]

\[ \xi^{MR}_{1,2} \equiv E^v \left\{ \left( \frac{h^*_1}{h_1} - \frac{h^*_2}{h^*_2} \right) \Pi_{n,1} \left\{ E^v,(A_1,T_2) \left\{ \frac{h^*_2}{h^*_2} (\eta_{2,MR} - \eta^g_2) \right\} \right\} \right\} \]

\[ 68 \]
\[
\xi_{2,3}^{DR} \equiv \frac{\pi^1}{\pi^1} \left( \frac{h_2^* - h_2^*}{h_2^*} \right) \Pi_{n,2} \left\{ E^v, (\bar{\pi}_2, \bar{\tau}_3) \left\{ \frac{h_3^*}{h_3^*} (\eta_{3,DR} - \eta_3^g) \right\} \right\}
\]

\[
\xi_{2,3}^{MR} \equiv E^v \left\{ \frac{\pi^1}{\pi^1} \left( \frac{h_2^*}{h_2^*} - \frac{h_2^*}{h_2^*} \right) \Pi_{n,2} \left\{ E^v, (\bar{\pi}_2, \bar{\tau}_3) \left\{ \left( \frac{h_3^*}{h_3^*} - \frac{h_3^*}{h_3^*} \right) \eta_{3,MR} - \eta_3^g \right\} \right\} \right\}
\]

\[
\xi_{1,3}^{DR} \equiv E^v \left\{ \left( \frac{h_1^*}{h_1^*} - \frac{h_1^*}{h_1^*} \right) \Pi_{n,1} \left\{ E^v, (\bar{\pi}_1, \bar{\tau}_2) \left\{ \frac{h_2^*}{h_2^*} \Pi_{n,2} \left\{ E^v, (\bar{\pi}_2, \bar{\tau}_3) \left\{ \left( \frac{h_3^*}{h_3^*} - \frac{h_3^*}{h_3^*} \right) \eta_{3,DR} - \eta_3^g \right\} \right\} \right\} \right\} \right\}
\]

\[
\xi_{1,3}^{MR} \equiv E^v \left\{ \left( \frac{h_1^*}{h_1^*} - \frac{h_1^*}{h_1^*} \right) \Pi_{n,1} \left\{ E^v, (\bar{\pi}_1, \bar{\tau}_2) \left\{ \left( \frac{h_2^*}{h_2^*} - \frac{h_2^*}{h_2^*} \right) \Pi_{n,2} \left\{ E^v, (\bar{\pi}_2, \bar{\tau}_3) \left\{ \left( \frac{h_3^*}{h_3^*} - \frac{h_3^*}{h_3^*} \right) \eta_{3,MR} - \eta_3^g \right\} \right\} \right\} \right\} \right\}
\]

and

\[
\xi_{1,2,3}^{MR} \equiv E^v \left\{ \left( \frac{h_1^*}{h_1^*} - \frac{h_1^*}{h_1^*} \right) \Pi_{n,1} \left\{ E^v, (\bar{\pi}_1, \bar{\tau}_2) \left\{ \left( \frac{h_2^*}{h_2^*} - \frac{h_2^*}{h_2^*} \right) \Pi_{n,2} \left\{ E^v, (\bar{\pi}_2, \bar{\tau}_3) \left\{ \left( \frac{h_3^*}{h_3^*} - \frac{h_3^*}{h_3^*} \right) \eta_{3,MR} - \eta_3^g \right\} \right\} \right\} \right\} \right\}
\]

We will now compare, under assumptions (i) - (vi) listed below, the rate of convergence of the drifts of the estimators \( \hat{\theta}_{DR,mach} \) and \( \hat{\theta}_{MR,mach} \) when the machine learning algorithms used are series estimators:

For each \( k \in \{1, 2, 3\} \),

i. \( A_k \) is discrete with finite sample space,

ii. \( \bar{T}_k \) is absolutely continuous with respect to Lebesgue measure with support on a compact set in \( \mathbb{R}^{d_k} \)

iii. \( h_k^g (\bar{\pi}_k, \bar{\tau}_k) \) and \( h_k (\bar{\pi}_k, \bar{\tau}_k) \), as functions of \( \bar{\tau}_k \) for each fixed \( \bar{\pi}_k \), lie in Holder balls with exponents \( \nu_{h,k} \) and \( \nu_{h,k} \)

iv. the machine learning algorithm procedure in steps (b.1) and (c.1) of Algorithm 6 is least squares on the covariate vector

\[
S_k^n = vec \left[ A_k \otimes s_k^n (\bar{T}_k)^T \right]
\]

where \( s_k^n (\bar{T}_k) \) is the vector of the first \( m_k^n \) (\( n \) elements of a complete basis having optimal rates of approximation for Holder classes in \( L_r (\mu), 1 \leq r \leq \infty \).
v. the machine learning algorithm procedure in step (a) of Algorithm 6 is linear
logistic regression with covariate vector

\[ S_h^k = \text{vec} \begin{bmatrix} A_k \otimes s_h^k (L_k)^T \end{bmatrix} \]

where \( s_h^k (L_k) \) is the vector of the first \( m_h^k (n) \) elements of a complete basis having optimal rates of approximation for Hölder classes in \( L_r (\mu) \), \( 1 \leq r \leq \infty \).

vi. the Hölder exponents \( \nu_{\eta,k} \) and \( \nu_{h,k} \) are known, \( m_{\eta}^k (n) = n^{1/(1+2\gamma_{\eta,k})} \) and \( m_{h}^k (n) = n^{1/(1+2\gamma_{h,k})} \), \( k = 1, 2, 3 \), where \( \gamma_{\eta,k} \equiv \nu_{\eta,k}/d_k \) and \( \gamma_{h,k} \equiv \nu_{h,k}/d_k \).

Note that in assumption (ii), \( d_k > d_{k'} \) if \( k > k' \). Also, a function \( f(\cdot) \) with compact domain in \( \mathbb{R}^d \) is said to belong to a Hölder ball \( H(\nu, C) \), with Hölder exponent \( \nu > 0 \) and radius \( C > 0 \), if and only if \( f(\cdot) \) is uniformly bounded by \( C \), all partial derivatives of \( f(\cdot) \) up to order \( [\nu] \) exist and are bounded, and all partial derivatives \( \nabla^{[\nu]} \) of order \( [\nu] \) satisfy

\[ \sup_{x,x+\delta x \in [0,1]^d} \left| \nabla^{[\nu]} f(x + \delta x) - \nabla^{[\nu]} f(x) \right| \leq C ||\delta x||^{\nu-[\nu]} . \]

It is well known that under assumptions (i)-(iii) the optimal rates of convergence for \( \eta_{\eta}^k \) and \( h_k \) are \( n^{-\gamma_{\eta,k}/(1+2\gamma_{\eta,k})} \) and \( n^{-\gamma_{h,k}/(1+2\gamma_{h,k})} \) in \( L_r (\mu) \) norms, \( 1 \leq r < \infty \). For estimation of \( \eta_{\eta}^k \), the optimal rate of convergence is obtained by least squares regression of \( y_{\eta_{\eta}^{k+1}} (A_k, L_{k+1}) \) on \( S_h^k \) where \( s_h^k (L_k) \) is the vector of the first \( n^{1/(1+2\gamma_{\eta,k})} \) elements of a complete basis having optimal rates of approximation for Hölder classes in \( L_r (\mu) \), \( 1 \leq r < \infty \). Two examples of such basis are B-splines with sufficient number of derivatives or Daubechies compact wavelets of sufficient order.

Based on these facts and Theorem 1, we conjecture that the following result holds:

**Result 1:** When \( K = 3 \), if assumptions (i)-(vi) hold, the drift

\[ E^v \left[ Q_1 \left( \tilde{\eta}^i, \tilde{\eta}_{\text{mach}}^i \right) \right] - \theta (g) \text{ of the } \tilde{\theta}_{DR, mach} \text{ is} \]
\[
a_p\left(\hat{h}, \hat{\eta}_{\text{mach}}\right) = O_p\left[\max\left\{\left(n - \left(\frac{\gamma_{h,1}}{1 + 2\gamma_{h,1} + 1 + 2\gamma_{h,1}}\right)\right), \max_{k \in \{1,2,3\}} \left(n - \left(\frac{\gamma_{h,k}}{1 + 2\gamma_{h,k} + 1 + 2\gamma_{h,k}}\right)\right)\right\}\right]
\]

and the drift \(E^v\left[Q_1\left(\hat{h}, \hat{\eta}_{\text{mach}}\right)\right] - \theta\left(g\right)\) of \(\hat{\theta}_{DR,\text{mach}}\) is

\[
a_p\left(\hat{h}, \hat{\eta}_{\text{mach}}\right) = O_p\left[\max\left\{\left(n - \left(\frac{\gamma_{h,1}}{1 + 2\gamma_{h,1} + 1 + 2\gamma_{h,1}}\right)\right), \max_{k \in \{1,2,3\}} \left(n - \left(\frac{\gamma_{h,k}}{1 + 2\gamma_{h,k} + 1 + 2\gamma_{h,k}}\right)\right)\right\}\right]
\]

We provide here a sketch of the argument why we believe Result 1 is true and towards the end of this argument we explain why we view it as a conjecture and not as a theorem. Under assumption (iv) \(\eta_{k,DR} = \Pi^\eta_{n,k}y_{k+1,\eta_{k+1}}(\bar{A}_k, \bar{T}_{k+1})\) where \(\Pi^\eta_{n,k}\) is the projection operator \(\Pi^\eta_{n,k}\) but with \(S_k = vec\left[\bar{A}_k \otimes s_k(\bar{T}_k)^T\right]\) instead of \(s_k(\bar{T}_k)\). On the other hand, \(\eta^g_k = E_g\left[y_{k+1,\eta_{k+1}}(\bar{A}_k, \bar{T}_{k+1}) | \bar{A}_k, \bar{T}_k\right]\). Thus, under assumptions (i)-(iv) and (vi), we have \(\eta_{k,DR} - \eta^g_k = O_p\left(n^{-\gamma_{n,k}/(1 + 2\gamma_{n,k})}\right)\). Now, invoking Cauchy-Schwartz repeatedly in the right hand side of formula (69), we obtain the rate of convergence stated for the drift \(a_p\left(\hat{h}, \hat{\eta}_{\text{mach}}\right)\) of \(\hat{\theta}_{DR,\text{mach}}\) in Result 1. Next, for \(k = 1, 2, 3\), when \(\tilde{\eta}_{k,MR} = \tilde{\eta}_{k,\text{mach}}^u\) and \(\Pi_k = \Pi^\eta_{n,k}\), the formula for \(\eta_{k,MR}\) becomes

\[
\eta_{k,MR} \equiv \eta_{k,MR}(\bar{A}_k, \bar{T}_k) = \eta_{k,DR} + \Pi^\eta_{n,k} Q_{k+1}\left(\hat{h}_{k+1, \eta_{k+1,\text{mach}}}^3, \hat{\eta}_{k+1,\text{mach}}^3\right) - E^v\left(\bar{\eta}_{k+1,\text{mach}}^3, \bar{\eta}_{k+1,\text{mach}}^3\right)\]

71
where $Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) = \psi(L_4)$ when $k = 3$. If $\hat{h}_{j+1}, \tilde{\eta}_{j+1, \text{mach}}$ had been fixed functions, i.e. they had not depended on the training sample data, then the difference $Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) - E^v(\mathcal{T}_{k+1}) \left\{ Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) \right\}$ would have been a mean zero random variable and $\Pi_{n,k}^{t,\eta}$ would have been applied to i.i.d. mean zero random variables. Thus, $\Pi_{n,k}^{t,\eta} \left[ Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) - E^v(\mathcal{T}_{k+1}) \left\{ Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) \right\} \right]$ would have been of order $O_p(m_{\eta,k}(n)/n)$, which would then have proved Result 1. However, in the training sample to which $\Pi_{n,k}^{t,\eta}$ is applied to, the random variables $Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) - E^v(\mathcal{T}_{k+1}) \left\{ Q_{k+1}(\mathcal{H}_{k+1}, \mathcal{E}_{k+1}, \text{mach}) \right\}$ are neither mean zero nor i.i.d. because the functions $\hat{h}_{j+1}, \tilde{\eta}_{j+1, \text{mach}}$ are not fixed, but rather they depend on data from that training sample. As a consequence we view Result 1 as a conjecture, although we expect it to be true. As an alternative to Algorithm 6, in the Appendix (section 7.5) we provide two multi-layer cross-fit algorithms that avoid the within training sample dependence described above.

Even if true, Result 1 is of no direct practical application because, in reality, one does not know the Holder exponents $\nu_{\eta,k}$ and $\nu_{h,k}$. However, it is known that the rates of convergence $n^{-\gamma_{\eta,k}/(1+2\gamma_{\eta,k})}$ and $n^{-\gamma_{h,k}/(1+2\gamma_{h,k})}$ for estimation of $\eta_k$ and $h_k$ can be achieved, up to log factors, even if the smoothness of the functions is unknown. For example, such adaption to unknown smoothness can be achieved by choosing the number of basis functions by cross-validation (Dudoit and van der Laan, 2003). This leads to the following conjecture.

**Result 2**: Result 1 holds if we replace assumption (vi) with the following assumption:

\((vi')\) $m_{\eta,k}^\eta(n)$ and $m_{h,k}^h(n)$ are chosen by V-fold cross-validation using empirical $L_2-$loss, $k = 1, 2, 3$.

To proceed with the discussion, we will assume henceforth that Results 1 and 2 hold. The formula (71) for the order of the drift of $\hat{\theta}_{\text{DR,mach}}$ involves the maximum over six second order terms corresponding to the six terms in the right hand side of
(69). On the other hand, formula (72) for the order of the drift of \( \hat{\theta}_{MR,\text{mach}} \) involves the maximum over three second order terms (corresponding to the terms \( E^v(\delta_k^{MR}) \), \( k = 1, 2, 3 \) in (70)), three third order terms and one fourth order term.

Under assumptions (i)-(vi), or assumptions (i)-(v) and (vi'), for each \( k = 1, 2, 3 \), \( E^v(\delta_k^{DR}) \) and \( E^v(\delta_k^{MR}) \) are of the same order, namely \( O_p\left(n^{-\frac{\gamma_{h,k}}{1+2\gamma_{h,k}} + \frac{\gamma_{k,g}}{1+2\gamma_{k,g}}}ight) \). Also, for each \( (i,j) \in \{(1,2),(1,3),(2,3)\} \), \( E^v(\xi_{i,j}^{DR}) \) converges to 0 slower than \( E^v(\xi_{i,j}^{MR}) \) because \( E^v(\xi_{i,j}^{MR}) \) is a third order term that involves the expectation of the product of three differences, two of which agree with the differences in \( E^v(\xi_{i,j}^{DR}) \). By the same reasoning, \( E^v(\xi_{1,2}^{MR}) \) converges to 0 faster than \( E^v(\xi_{1,3}^{MR}) \) and \( E^v(\xi_{2,3}^{MR}) \).

In general, one might expect that the terms \( E^v(\delta_3^{DR}) \) and \( E^v(\delta_3^{MR}) \) would be the dominating terms in the drifts of \( \hat{\theta}_{DR,\text{mach}} \) and \( \hat{\theta}_{MR,\text{mach}} \), i.e. the terms with the slowest rates of convergence, because (1) these terms involve two regressions on the covariates \((\bar{A}_3, \bar{L}_3)\) and (2) these covariates are a superset of the covariates conditioned upon in the regressions involved in all other terms that appear in the right hand sides of (69) and (70). By the same reasoning, one might expect that the second largest term of the drift \( \hat{\theta}_{DR,\text{mach}} \) should be \( E^v(\xi_{2,3}^{DR}) \) with rate of convergence \( O_p\left(n^{-\frac{\gamma_{h,2}+\gamma_{h,3}}{1+2\gamma_{h,2}+\gamma_{h,3}}} \right) \). However, it could happen that at the particular law that generated the data, \( \gamma_{h,2} = \nu_{h,2}/d_2 \) could be less than \( \gamma_{h,3} = \nu_{h,3}/d_3 \) even though \( d_3 \) is greater than \( d_2 \). If \( E^v(\xi_{2,3}^{DR}) \) were, in fact, the dominating term of the drift of \( \hat{\theta}_{DR,\text{mach}} \) then, in view of the comparisons of the orders of the terms of the two drifts made above, the drift of \( \hat{\theta}_{MR,\text{mach}} \) would have a faster rate of convergence to 0 than that of \( \hat{\theta}_{DR,\text{mach}} \). Thus, it could be the case that the drift of \( \hat{\theta}_{MR,\text{mach}} \) is \( o_p\left(n^{-1/2}\right) \) but the drift of \( \hat{\theta}_{DR,\text{mach}} \) is not, in which case, by the analysis of section 5.1, \( \hat{\theta}_{MR,\text{mach}} \) would be an asymptotically linear estimator of \( \theta(g) \) but \( \hat{\theta}_{DR,\text{mach}} \) would not.

In the next subsection we will need to refer to the following additional result which follows from arguments analogous to those used to establish Results 1 and 2.

**Result 3.** Under assumptions (i)-(vi) or, assumptions (i)-(v) and (vi'), \( \tilde{\eta}_{k,\text{mach}} - E_{g_k}\left\{y_{k+1,\tilde{\eta}_{k+1,\text{mach}}(\bar{A}_k, \bar{T}_{k+1})} \mid \bar{A}_k, \bar{T}_k\right\} \), \( \tilde{\eta}_{k,\text{mach}} - E_{g_k}\left\{Q_{k+1,\text{mach}}(\bar{A}_k, \bar{T}_k) \mid \bar{A}_k, \bar{T}_k\right\} \), \( \eta_{k,DR} - \eta_k^g \) and \( \eta_{k,MR} - \eta_k^g \) all converge to 0 at the same rate.
5.2.2 Analysis when the ML algorithms are arbitrary.

In this section we consider estimators \( \hat{\theta}_{MR,mach} \) and \( \hat{\theta}_{DR,mach} \) that use arbitrary machine learning algorithms to estimate the nuisance functions. In order to analyze the rates of convergence to 0 of the drifts of \( \hat{\theta}_{DR,mach} \) and \( \hat{\theta}_{MR,mach} \) we return to the formulae (61) and (62). To facilitate the discussion we write formula (61) for the drift of \( \hat{\theta}_{DR,mach} \) as

\[
E^v \left[ Q_1 \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \right] - \theta \left( g \right) = \sum_{k=1}^{K} E^v \left[ \frac{\pi_{k-1} \hat{\eta}_{mach}^t}{\pi_k} \left( \frac{h_k^*}{h_k} - \frac{h_k^*}{\hat{h}_k} \right) R_{DR,k} \right]
\]

\[
= \rho_{DR}^{MR} + \sum_{1 \leq j < k \leq K} \chi_{j,k}^{DR}
\]

and formula (62) for the drift of \( \hat{\theta}_{MR,mach} \) as

\[
E^v \left[ Q_1 \left( \hat{h}^t, \hat{\eta}_{mach}^t \right) \right] - \theta \left( g \right) = \sum_{k=1}^{K} E^v \left[ \frac{\pi_{k-1} \hat{\eta}_{mach}^t}{\pi_k} \left( \frac{h_k^*}{h_k} - \frac{h_k^*}{\hat{h}_k} \right) R_{MR,k} \right]
\]

\[
= \rho_{MR}^{MR}
\]

where

\[
R_{DR,k} \equiv \hat{\eta}_{k,mach}^t - E_{g_k} \left\{ y_{k+1, \hat{\eta}_{k+1,mach}}^t \left( \bar{A}_k, \bar{L}_{k+1} \right) \left| \bar{A}_k, \bar{L}_k \right. \right\}
\]

and

\[
R_{MR,k} \equiv \hat{\eta}_{k,mach}^t - E_{g_k} \left\{ Q_{k+1,mach} \left| \bar{A}_k, \bar{L}_k \right. \right\}
\]

Suppose it were the case that, as with the linear machine learning algorithm, the rate of convergence to 0 of \( R_{DR,k} \) and \( R_{MR,k} \) were the same. Then \( \rho_{MR}^{MR} \) and \( \rho_{DR}^{DR} \) would converge to 0 at the same rate. Now, one would generally expect that the terms \( \rho_{DR}^{DR} \) and \( \rho_{MR}^{MR} \) would be the dominating terms in the drifts of \( \hat{\theta}_{DR,mach} \) and
\[ \hat{\theta}_{MR,mach}, \text{i.e. the terms with the slowest rate of convergence, because (1) } \rho_{1,K}^{DR} \text{ and } \rho_{1,K}^{MR} \text{ involve two regressions on the covariates } (A_K, L_K) \text{ and (2) these covariates are a superset of the covariates conditioned upon in the regressions involved in all other terms. However, again it could happen that at the particular law that generated the data, one of the } K(K - 1)/2 \text{ terms } \chi_{j,k}^{DR} \text{ in } \chi^{DR} \text{ dominates } \rho^{DR}, \text{i.e. it converges to 0 slower than any of the terms in } \rho^{DR}. \text{ In such case, the drift } \rho^{MR} \text{ of } \hat{\theta}_{MR,mach} \text{ would have a faster rate of convergence to 0 than the drift of } \hat{\theta}_{DR,mach}. \text{ In particular, it could happen that } \hat{\theta}_{MR,mach} \text{ is an asymptotically linear estimator of } \theta(g) \text{ even though } \hat{\theta}_{DR,mach} \text{ is not. The frequency with which the law generating the data has the drift of } \hat{\theta}_{MR,mach} \text{ converging to 0 faster than the drift of } \hat{\theta}_{DR,mach} \text{ may be greater for } K \text{ large because the ratio of the number of terms in } \chi^{DR} \text{ to that in } \rho^{DR} \text{ increases linearly with } K, \text{ providing an increasing number of opportunities for } \chi^{DR} \text{ to dominate the drift of } \hat{\theta}_{DR,mach}. \text{ Of course this discussion must be tempered by the fact that, for non-linear machine learning algorithms, we have no guarantee that the residuals } R_{MR,k} \text{ converge to 0 as fast or faster than the residuals } R_{DR,k}. \]

6 References

Ayyagari, R. (2010), Applications of Influence Functions to Semiparametric Regression Models. Harvard School of Public Health, Department of Biostatistics Doctoral Thesis

Babino, L., Rotnitzky, A. and Robins, J. (2017). Multiple Robust Estimation of Marginal Structural Mean Models. Submitted to Biometrics.

Bang, H., and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Belloni, Alexandre, Chen, Daniel, Chernozhukov, Victor and Hansen, Christian. (2010, 2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain’, arXiv (2010) preprint, Econometrica (2012) 80, 2369-2429

Chernozhukov, V. et al. Double machine learning for treatment and causal parameters. arXiv:1608.00060, 1{37 (2016).

Dudoit, S, and van der Laan, M. (2003) "Asymptotics of cross-validated risk estimation in model selection and performance assessment." UC Berkeley Division
Molina, J., Rotnitzky, A., Sued, M. and Robins, J.M. (2017). Multiple robustness in factorized likelihood models. *Biometrika*. To appear.

Petersen M1, Schwab J1, Gruber S2, Blaser N3, Schomaker M4, van der Laan M1. (2014) Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. *J Causal Inference*. Jun 18;2(2):147-185.

Robins JM. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12), 1393-1512.

Robins, J. M. (1987). Addendum to “a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect”. *Computers & Mathematics with Applications*, 14(9), 923-945.

Robins JM. (1993). Analytic methods for estimating HIV treatment and cofactor effects. *Methodological Issues of AIDS Mental Health Research*. Eds: Ostrow D.G., Kessler R. New York: Plenum Publishing. pp. 213-290

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*. Springer New York.

Robins J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 6-10.

Robins JM. (2002). Commentary on ”Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard.” *Statistics in Medicine*. 21:1663-1680

Robins JM, Li L, Tchetgen E, van der Vaart A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman* 2:335-421.

Robins JM, Rotnitzky A, Zhao LP. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical
7 Appendix

7.1 Proof of Lemma 1

By definition, and the absolute continuity of $g h^*$ with respect to $g h$, we have that for $k \in [K - 1]$,

$$
y_{k+1, \eta_k+1} (\overline{A}_k, \overline{T}_{k+1}) = E_{h_{k+1}} (\eta_{k+1} | \overline{A}_{k+1}, \overline{T}_{k+1}) = E_{h_{k+1}} \left( \frac{h^*_{k+1} \eta_{k+1}}{h_{k+1}} | \overline{A}_{k+1}, \overline{T}_{k+1} \right),
$$

where $h^*_k \equiv h^*_k (A_k | \overline{A}_{k-1}, \overline{L}_k)$, $h_k \equiv h_k (A_k | \overline{A}_{k-1}, \overline{L}_k)$ and $\eta_k \equiv \eta_k (\overline{A}_k, \overline{T}_k)$. Thus, for $k \in [K - 1]$,

$$
\frac{\Delta_k (\eta_k, \eta_{k+1}; g_k)}{\pi^k} = \frac{\pi^* k}{\pi^k} \left\{ \eta_k - E_{g_k} \left[ E_{h_{k+1}} \left( \frac{h^*_{k+1} \eta_{k+1}}{h_{k+1}} | \overline{A}_{k+1}, \overline{T}_{k+1} \right) | \overline{A}_k, \overline{T}_k \right] \right\} \tag{73}
$$

Consequently,

$$
E_{\overline{g}_{k-1}, \overline{h}_k} \left\{ \frac{\Delta_k (\eta_k, \eta_{k+1}; g_k)}{\pi^k} \right\} = E_{\overline{g}_{k-1}, \overline{h}_k} \left( \frac{\pi^* k}{\pi^k} \eta_k \right) - E_{\overline{g}_k, \overline{h}_{k+1}} \left( \frac{\pi^{*k+1}}{\pi^{k+1}} \eta_{k+1} \right).
$$
In addition,

\[
E_{g_{K-1},\pi_K} \left\{ \frac{\Delta K (\eta_K, \eta_{K+1}; g_K)}{\pi^K} \right\} =
\]

\[
= E_{g_{K-1},\pi_K} \left( \frac{\pi^* K}{\pi^K} \eta_K \right) - E_{g_{K-1},\pi_K} \left[ \frac{\pi^* K}{\pi^K} E_{g_K} \left\{ \psi (L_{K+1}) \big| A_K, L_K \right\} \right]
\]

\[
= E_{g_{K-1},\pi_K} \left( \frac{\pi^* K}{\pi^K} \eta_K \right) - E_{g_{K-1},\pi_K} \left\{ \frac{\pi^* K}{\pi^K} \psi (L_{K+1}) \right\}
\]

\[
= E_{g_{K-1},\pi_K} \left( \frac{\pi^* K}{\pi^K} \eta_K \right) - \theta (g).
\]

Consequently,

\[
\sum_{k=1}^{K} E_{g_{k-1},\pi_k} \left\{ \frac{\Delta_k (\eta_k, \eta_{k+1}; g_k)}{\pi^k} \right\} =
\]

\[
= \sum_{k=1}^{K-1} \left\{ E_{g_{k-1},\pi_k} \left( \frac{\pi^k}{\pi^K} \eta_k \right) - E_{g_k,\pi_{k+1}} \left( \frac{\pi^{k+1}}{\pi^K} \eta_{k+1} \right) \right\} + E_{g_{K-1},\pi_K} \left( \frac{\pi^* K}{\pi^K} \eta_K \right) - \theta (g)
\]

\[
= E_{g_0,h_1} \left( \frac{h_1^*}{h_1} \eta_1 \right) - \theta (g)
\]

\[
= E_{g_1} \left\{ y_1, n_1 (L_1) \right\} - \theta (g),
\]

as we wished to show.

7.2 Proof of Lemma 2

The identity (34) coincides with (35) for \( j = 0 \) if \( \overline{A}_{j-1} \) and \( \overline{L}_{j-1} \) are defined as nil when \( j = 1 \). It thus suffices to show (35) for an arbitrary \( j \in \{0, ..., K\} \). We prove it by reverse induction. For \( j = K \) the result holds by definition of \( \eta^0_K (A_K, L_K) \) since \( Q_{K+1} (\pi^*_K, \pi^*_K) = \psi (L_{K+1}) \). Suppose now that (35) holds for a given \( j \in [K] \),
we want to show that it also holds for \( j - 1 \). Now,

\[
E_{g_{j-1}, h_j} \left\{ Q_j \left( \eta_j, \eta_j^\dagger \right) \left| \overline{A}_{j-1}, \overline{L}_{j-1} \right. \right\} - \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) =
\]

\[
= E_{g_{j-1}, h_j} \left\{ \eta_j \left( \eta_j, \eta_j^\dagger \right) \left| \overline{A}_{j-1}, \overline{L}_{j-1} \right. \right\} - \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right)
\]

\[
= E_{g_{j-1}, h_j} \left[ \eta_j^\dagger \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right]
\]

\[
+ E_{g_{j-1}, h_j} \left[ \eta_j \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right] - \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right)
\]

\[
= E_{g_{j-1}, h_j} \left[ \eta_j^\dagger \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right]
\]

\[
+ E_{g_{j-1}, h_j} \left[ \eta_j \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right] - \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right)
\]

\[
= E_{g_{j-1}, h_j} \left[ \eta_j^\dagger \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right]
\]

\[
+ E_{g_{j-1}, h_j} \left[ \eta_j \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right) \right] - \eta_j^g \left( \overline{A}_{j-1}, \overline{L}_{j-1} \right)
\]

where the third equality is by the inductive hypothesis. This concludes the proof.
7.3 Proof of Lemma 3

We prove (1) by reverse induction that for \( k \in \{0, 1, \ldots, K\} \),

\[
\pi^* \{ \eta_k^+ - \eta_k^q \} = \Gamma_k + \sum_{s=k+1}^K E_{gh} \left\{ \frac{1}{\pi_{s-1}} \left( \frac{1}{h_s} - \frac{1}{h_s^+} \right) \Gamma_s \bigg| \bar{A}_K, \bar{L}_k \right\} \quad (74)
\]

where to simplify notation we use the shortcut \( \Gamma_k \equiv \Gamma_k(h_{k+1}^+, \eta_{k+1}^+; g_k) \) and \( \sum_{s=K+1}^K (\cdot) \equiv 0 \).

Applying this equality to \( k = 0 \) with \( \eta_0^+ = E_{gh} \left\{ Q_1 \left( \bar{h}_1^+, \bar{\eta}_1^+ \right) \right\} \quad (75) \)
we obtain that

\[
\pi^0 \left[ E_{gh} \left\{ Q_1 \left( \bar{h}_1^+, \bar{\eta}_1^+ \right) \right\} - \eta_0^q \right] = \Gamma_0 + \sum_{s=1}^K E_{gh} \left\{ \frac{1}{\pi_{s-1}} \left( \frac{1}{h_s} - \frac{1}{h_s^+} \right) \Gamma_s \bigg| \bar{A}_0, \bar{L}_0 \right\}
\]

Recalling that \( \pi^0 \equiv 1, \eta_0^q = \theta(g), (\bar{A}_0, \bar{L}_0) \equiv \text{nill} \), and that with \( \eta_0^+ \) defined as in \( (75) \), \( \Gamma_0 \equiv \eta_0^+ - E_{gh} \left\{ Q_1 \left( \bar{h}_1^+, \bar{\eta}_1^+ \right) \right\} = 0 \), we conclude that

\[
E_{gh} \left\{ Q_1 \left( \bar{h}_1^+, \bar{\eta}_1^+ \right) \right\} - \theta(g) = \sum_{s=1}^K E_{gh} \left\{ \frac{1}{\pi_{s-1}} \left( \frac{1}{h_s} - \frac{1}{h_s^+} \right) \Gamma_s \bigg| \bar{A}_0, \bar{L}_0 \right\} \equiv b^p \left( h^+ \downarrow, \eta^+ \right)
\]

which, invoking Lemma 2, proves that \( b^p (h^+ \downarrow, \eta^+ \downarrow) = a^p (h^+ \downarrow, \eta^+ \downarrow) \).

We now prove identity (74) by induction.

For \( k = K \), (74) holds because by definition

\[
\pi^k \left( \eta_K^+ - \eta_K^q \right) \equiv \pi^k \left[ \eta_K^+ - E_{gK} \left\{ \psi (\bar{T}_{K+1}) \bigg| \bar{A}_K, \bar{L}_K \right\} \right] \\
\equiv \pi^k \left[ \eta_K^+ - E_{gK} \left\{ Q_{K+1} \left( \bar{h}_{K+1}^+, \bar{\eta}_{K+1}^+ \right) \bigg| \bar{A}_K, \bar{L}_K \right\} \right] \\
\equiv \Gamma_K.
\]

Suppose (74) holds for \( k = K, \ldots, j + 1 \). We will show that it holds for \( k = j \).
By Lemma 2 we have

\[ \pi^*j (\eta_j^\dagger - \eta_j^g) = \pi^*j \left[ \eta_j^\dagger - E_{g_j, L_{j+1}} \left\{ Q_{j+1} \left( \frac{\eta_j^K}{h_{j+1}}, \eta_j^g \right) | A_j, T_j \right\} \right] \]

\[ + \pi^*j \left[ E_{g_j, L_{j+1}} \left\{ Q_{j+1} \left( \frac{\eta_j^K}{h_{j+1}}, \eta_j^g \right) | A_j, T_j \right\} - \eta_j^g \right] \]

\[ = \Gamma_j + \sum_{k=j+1}^K E_{g_j, L_{j+1}} \left\{ \frac{\pi^{*(k-1)}}{\pi^{j+1}_j} \left( \frac{h_k^*}{h_k} - \frac{h_k^*}{h_k} \right) \left( \eta_k^\dagger - \eta_k^g \right) \right\} | A_j, T_j \]

Then, invoking the inductive assumption we obtain

\[ \pi^*j (\eta_j^\dagger - \eta_j^g) = \Gamma_j + \sum_{k=j+1}^K E_{g_j, L_{j+1}} \left\{ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \Gamma_k \right\} | A_j, T_j \]

\[ + \sum_{k=j+1}^K \sum_{s=k+1}^K E_{g_j, L_{j+1}} \left\{ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \pi^{s-1}_k \left( \frac{1}{h_s} - \frac{1}{h_s} \right) \Gamma_s \right\} | A_j, T_j \]

Now, rearranging the terms in the double-sum we obtain

\[ \sum_{k=j+1}^K \sum_{s=k+1}^K E_{g_j, L_{j+1}} \left[ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \pi^{s-1}_k \left( \frac{1}{h_s} - \frac{1}{h_s} \right) \Gamma_s \right] | A_j, T_j \]

\[ = \sum_{s=j+2}^K E_{g_j, L_{j+1}} \left[ \left( \frac{1}{h_s} - \frac{1}{h_s} \right) \Gamma_s \sum_{k=j+1}^{s-1} \left\{ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \pi^{s-1}_k \right\} | A_j, T_j \right] \]

and we prove below that

\[ \sum_{k=j+1}^{s-1} \left\{ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \pi^{s-1}_k \right\} = \frac{1}{\pi^{j+1}_j} - \frac{1}{\pi^{j+1}_j} \]

Thus,

\[ \sum_{k=j+1}^{s-1} \left\{ \frac{1}{\pi^{j+1}_j} \left( \frac{1}{h_k} - \frac{1}{h_k} \right) \pi^{s-1}_k \right\} = \frac{1}{\pi^{j+1}_j} - \frac{1}{\pi^{j+1}_j} \] (76)
\[ \pi^s j \left( \eta_j^\dagger - \eta_j^s \right) = \Gamma_j + \sum_{k=j+1}^{K} E_{g_j a_{j+1}} \left\{ \frac{1}{\pi_{j+1}} \left( \frac{1}{h_k} - \frac{1}{h_k^s} \right) \frac{1}{\pi_{k+1}} \right\} \mathcal{A}_j, \mathcal{L}_j \]

\[ + \sum_{s=j+2}^{K} E_{g_j a_{j+1}} \left\{ \frac{1}{\pi_{j+1}} \left( \frac{1}{h_s} - \frac{1}{h_s^s} \right) \frac{1}{\pi_{k+1}} \right\} \mathcal{A}_j, \mathcal{L}_j \]

\[ = \Gamma_j + \sum_{s=j+1}^{K} E_{g_j a_{j+1}} \left\{ \frac{1}{\pi_{j+1}} \left( \frac{1}{h_s} - \frac{1}{h_s^s} \right) \frac{1}{\pi_{s+1}} \right\} \mathcal{A}_j, \mathcal{L}_j \]

as we wish to show.

We now show (76).

\[
\sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{k+1}} \frac{1}{\pi_{k+1}} = \sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{k+1}} \frac{1}{\pi_{k+1}} - \sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{k+1}}
\]

\[ = \frac{1}{\pi_{j+1}} + \sum_{k=j+2}^{s-1} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{k+1}} - \sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}} \frac{1}{\pi_{k+1}} \frac{1}{\pi_{k+1}} - \frac{1}{\pi_{j+1}}
\]

\[ = \frac{1}{\pi_{j+1}} \frac{1}{\pi_{(s-1)}} \]

This concludes the proof that \( b^p (h^t, \eta^t) = a^p (h^t, \eta^t) \).
We now prove that $c^p (h^†, \eta^†) = a^p (h^†, \eta^†)$.

For any $(\mathcal{A}_k, \mathcal{L}_k)$ such that $\pi^{*k} > 0$ define

$$\delta_k (\eta_k, \eta_{k+1}; g_k) \equiv \eta_k (\mathcal{A}_k, \mathcal{L}_k) - E_{g_k} \left\{ y_{k+1, \eta_{k+1}} (\mathcal{A}_k, \mathcal{L}_{k+1}) \mid \mathcal{A}_k, \mathcal{L}_k \right\}$$

Then, conditioning in $(\mathcal{A}_k, \mathcal{L}_k)$, the proof of Lemma 1 can be immediately adapted to show that

$$E_{g_k} \left\{ y_{k+1, \eta_{k+1}^†} (\mathcal{A}_k, \mathcal{L}_{k+1}) \mid \mathcal{A}_k, \mathcal{L}_k \right\} - \eta_k^g (\mathcal{A}_k, \mathcal{L}_k) = \sum_{j=k+1}^{K} E_{g_k} \left\{ \frac{\pi^{*j}_{k+1}}{\pi^{*j}_{k+1}} \delta_j (\eta_j^†, \eta_{j+1}^†; g_j) \mid \mathcal{A}_k, \mathcal{L}_k \right\},$$

from where we deduce that

$$\eta_k^† (\mathcal{A}_k, \mathcal{L}_k) - \eta_k^g (\mathcal{A}_k, \mathcal{L}_k) = \sum_{j=k}^{K} \delta_j (\eta_j^†, \eta_{j+1}; g_j) \equiv \sum_{j=k}^{K} \sum_{k=1}^{K} \frac{\pi^{*j}_{k+1}}{\pi^{*j}_{k+1}} \delta_j (\eta_j^†, \eta_{j+1}; g_j) \left\{ \frac{\pi^{*j}_{k+1}}{\pi^{*j}_{k+1}} \right\}.$$
Now,
\[ \sum_{k=1}^{j} \left\{ \frac{\pi^{*}(k-1)}{\pi^!(k-1)} \left( \frac{h^*_k}{h_k} - \frac{h^*_j}{h^!_k} \right) \right\} = \pi^* \sum_{k=1}^{j} \left\{ \frac{1}{\pi^!(k-1)} \left( \frac{1}{h_k} - \frac{1}{h^!_k} \right) \right\}, \]
so (78) follows from (76) by evaluating in (76) \( j \) at 0 and \( s \) at \( j + 1 \). This concludes the proof.

7.4 Proof of Theorem 1

The proof of Theorem 1 invokes the following Lemma.

**Lemma A.1.**

For any \( j \in [K] \),
\[ E_{g_j} \left\{ Q_j \left( \frac{h^!_{j+K-1}}{h^!_{j}}, \frac{h^*_{j+K}}{h_{j+K}} \right) \bigg| A_{j-1}, L_j \right\} - E_{h_j} \left( \frac{h^*_j}{h^!_j} \eta^g_j \bigg| A_{j-1}, L_j \right) = (79) \]
\[ = \sum_{k=j}^{K} E_{g_j} \left\{ \frac{\pi^{*}(k-1)}{\pi^!(k-1)} \left( \eta^g_k - \eta^g_{j+K} \right) \left( \frac{h^*_k}{h_k} - \frac{h^*_j}{h^!_k} \right) \bigg| A_{j-1}, L_j \right\}. \]

**Proof of Lemma A.1.**

We prove it by reverse induction.
For $j = K$ we have

$$E_{g_K, h_K} \left\{ Q_K \left( \frac{h_K^*}{h_K^\dagger}, \eta_K^\dagger \right) \left| \overline{A}_{K-1}, \overline{L}_K \right\} - E_{h_K} \left( \frac{h_K^*}{h_K^\dagger}, \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right) =$$

$$= E_{g_K, h_K} \left\{ \frac{h_K^*}{h_K^\dagger} \left( \psi \left( \overline{L}_{K+1} \right) - \eta_K^\dagger \right) + y_K, \eta_K^\dagger \left( \overline{A}_{K-1}, \overline{L}_K \right) \left| \overline{A}_{K-1}, \overline{L}_K \right\} - E_{h_K} \left( \frac{h_K^*}{h_K^\dagger}, \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right)$$

$$- E_{h_K} \left( \frac{h_K^*}{h_K^\dagger} \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right)$$

$$= E_{g_K, h_K} \left\{ \frac{h_K^*}{h_K^\dagger} \left( \psi \left( \overline{L}_{K+1} \right) - \eta_K^\dagger \right) \left| \overline{A}_{K-1}, \overline{L}_K \right\} \right\} - E_{h_K} \left( \frac{h_K^*}{h_K^\dagger}, \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right)$$

$$+ E_{h_K} \left( \frac{h_K^*}{h_K^\dagger} \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right) - E_{h_K} \left( \frac{h_K^*}{h_K^\dagger}, \eta_K^g \left| A_{K-1}, \overline{L}_K \right\right)$$

$$= E_{g_K, h_K} \left\{ \frac{h_K^*}{h_K^\dagger} \left( \eta_K^g - \eta_K^\dagger \right) \left| \overline{A}_{K-1}, \overline{L}_K \right\} \right\} + E_{h_K} \left( \frac{h_K^*}{h_K^\dagger} \left( \eta_K^\dagger - \eta_K^g \right) \left| A_{K-1}, \overline{L}_K \right\right)$$

Suppose now that (79) holds for a given $j \in [K]$, we want to show that it also holds

85
for $j - 1$. Now,

$$E_{g, \Delta_j} \{ Q_j \left( \bar{n}^{K}_{j}^*, \bar{n}^{K}_{j} \right) \left| A_{j-1}, \bar{T}_j \right. \} - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \} =$$

$$= E_{g, \Delta_j} \left[ \frac{h^*_j}{h_j} \left\{ Q_{j+1} \left( \bar{n}^{K}_{j+1}, \bar{n}^{K}_{j+1} \right) - \eta^g_j \right\} + y_{j, \eta^g_j} \left( \bar{A}_{j-1}, \bar{T}_j \right) \right] - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= E_{g, \Delta_j} \left[ \frac{h^*_j}{h_j} \left\{ Q_{j+1} \left( \bar{n}^{K}_{j+1}, \bar{n}^{K}_{j+1} \right) \right| \bar{T}_j, \bar{T}_{j+1} \right] \left| \bar{T}_j \right. \} \right]$$

$$- E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} + y_{j, \eta^g_j} \left( \bar{A}_{j-1}, \bar{T}_j \right) - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= E_{g, \Delta_j} \left[ \frac{h^*_j}{h_j} \left[ E_{g, \Delta_j+1} \left\{ Q_{j+1} \left( \bar{n}^{K}_{j+1}, \bar{n}^{K}_{j+1} \right) \right| \bar{T}_j, \bar{T}_{j+1} \right] \left| \bar{T}_j \right. \} \right]$$

$$+ E_{g, \Delta_j} \left\{ \frac{h^*_j}{h_j} \left( \frac{h^*_{j+1}}{h_{j+1}} \eta^g_j + 1 \right) A_{j}, \bar{T}_{j+1} \right) \left| \bar{T}_j \right. \} \right]$$

$$- E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} + E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= E_{g, \Delta_j} \left[ \left\{ 1 \right| \sum_{k=j+1}^K \frac{h^*_j}{h_j} \pi_j^{(k-1)} \left( \eta^g_j - \eta^g_k \right) \left( \frac{h^*_j}{h_j} \right) \left( \frac{h^*_k}{h_k} \right) \right| \bar{T}_j, \bar{T}_{j+1} \right] \right]$$

$$+ E_{g, \Delta_j} \left\{ \frac{h^*_j}{h_j} \left( \frac{h^*_{j+1}}{h_{j+1}} \eta^g_j + 1 \right) A_{j}, \bar{T}_{j+1} \right) \left| \bar{T}_j \right. \} \right]$$

$$- E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} + E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= \sum_{k=j+1}^K E_{g, \Delta_j} \left\{ \frac{\pi_j^{(k-1)}}{\pi_j^{(k-1)}} \left( \eta^g_j - \eta^g_k \right) \left( \frac{h^*_j}{h_j} \right) \left( \frac{h^*_k}{h_k} \right) \right| \bar{A}_{j-1}, \bar{T}_j \} \right] + E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \}$$

$$+ E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= \sum_{k=j+1}^K E_{g, \Delta_k} \left\{ \frac{\pi_j^{(k-1)}}{\pi_j^{(k-1)}} \left( \eta^g_j - \eta^g_k \right) \left( \frac{h^*_j}{h_j} \right) \left( \frac{h^*_k}{h_k} \right) \right| \bar{A}_{j-1}, \bar{T}_j \} \right]$$

$$+ E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) \bar{A}_{j-1}, \bar{T}_j \} - E_{h_j} \left( \frac{h^*_j}{h_j} | \eta^g_j \right) A_{j-1}, \bar{T}_j \}$$

$$= \sum_{k=j}^K E_{g, \Delta_k} \left\{ \frac{\pi_j^{(k-1)}}{\pi_j^{(k-1)}} \left( \eta^g_j - \eta^g_k \right) \left( \frac{h^*_j}{h_j} \right) \left( \frac{h^*_k}{h_k} \right) \right| \bar{A}_{j-1}, \bar{T}_j \} \right]$$
This concludes the proof of the Lemma A.1.

**Proof of Theorem 1:**

We prove part (1) by induction. Part (2) follows immediately.

For \( k = K \), (63) is true because \( \hat{\eta}_{K,DR} = \eta_{K,DR} \) since
\[
y_{K+1,\eta_{K+1,DR}} (A_{K}, \mathcal{T}_{K+1}) \equiv y_{K+1,\hat{\eta}_{K+1,DR}} (A_{j}, \mathcal{T}_{j+1}) \equiv \psi (\mathcal{T}_{K+1}).
\]

Suppose (63) is true for \( k = K, \ldots, j + 1 \). We will show it is true for \( k = j \).

\[
\hat{\eta}_{j,DR} - \eta_{j} = \Pi^j \left[ y_{j+1,\hat{\eta}_{j+1,DR}} (A_{j}, \mathcal{T}_{j+1}) \right] - \eta_{j}^g
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \Pi^j \left[ y_{j+1,\hat{\eta}_{j+1,DR}} (A_{j}, \mathcal{T}_{j+1}) - \eta_{j,DR} \right]
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \Pi^j \left[ y_{j+1,\hat{\eta}_{j+1,DR}} (A_{j}, \mathcal{T}_{j+1}) - y_{j+1,\hat{\eta}_{j+1,DR}} (A_{j}, \mathcal{T}_{j+1}) \right]
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \Pi^j \left[ E_{h_{j+1}} \left\{ \frac{h_{j+1}}{h_{j+1}} \left( \hat{\eta}_{j+1,DR} - \eta_{j+1}^g \right) \left| A_{j}, \mathcal{T}_{j+1} \right. \right\} \right]
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \Pi^j_{DR} \left[ \hat{\eta}_{j+1,DR} - \eta_{j+1}^g \right]
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \Pi^j_{DR} \left[ \eta_{j+1,DR} - \eta_{j+1}^g \right] + \sum_{k=j+2}^{K} \Pi^j_{DR,j,k} \left[ \eta_{k,DR} - \eta_{k}^g \right]
\]
\[
= (\eta_{j,DR} - \eta_{j}^g) + \sum_{k=j+1}^{K} \Pi^j_{DR,j,k} \left[ \eta_{k,DR} - \eta_{k}^g \right]
\]

The third to last equality is by the inductive hypothesis and the second to last is by the assumed linearity of the operator \( \Pi^j \) which induces linearity of the operator \( \Pi^j_{DR} \).

This concludes the proof of part (1).

We now prove part (3) by induction in \( K \). Part (4) follows immediately.

First we show (64) is true when \( K = 1 \).
For $K = 1$, we have

$$E_p \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1^*}{h_1^*} \right) (\tilde{\eta}_{1,MR} - \eta_1^g) \bigg| L_1 \right\} =$$

$$= E_p \left\{ \left( \frac{h_1^*}{h_1} - \frac{h_1^*}{h_1^*} \right) \eta_{1,MR} - \eta_1^g \bigg| L_1 \right\}$$

$$= E_p \left\{ \nabla_{0,1} (\eta_{1,MR} - \eta_1^g) \bigg| L_1 \right\}$$

$$= \sum_{k=1}^K E_p \left\{ \nabla_{0,k} (\eta_{k,MR} - \eta_k^g) \bigg| L_1 \right\}$$

$$+ \sum_{1 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{k=r_u+1}^K E_p \left\{ \nabla_{0,r_1 \Pi_{MR,r_1,r_2,\ldots,r_u},k} [\eta_{k,MR} - \eta_k^g] \bigg| L_1 \right\}$$

where the first equality follows because when $K = 1$, $Q_2 \left( \tilde{h}_2^1, \tilde{\eta}_2^{1} \right) \equiv \psi (L_2)$ so

$$\tilde{\eta}_{1,MR} \equiv \Pi^1 \left[ Q_2 \left( \tilde{h}_2^1, \tilde{\eta}_2^{1} \right) \big| S_1 \right] \equiv \Pi^1 \left[ \psi (L_2) \big| S_1 \right] \equiv \eta_{1,MR}$$

and the third equality is true because

$$\sum_{1 \leq r_1 < r_2 < \ldots < r_u \leq 0} (\cdot) \equiv 0.$$ 

This proves (64) for $K = 1$.

Next, assume (64) is true for $K - 1$, we will show it is true for $K$.

If (64) is true for $K - 1$, then it holds that

$$\sum_{k=2}^K E_p \left\{ \frac{\eta_{k-1}^*}{\eta_{k-1}^*} \left( \frac{h_k^*}{h_k} - \frac{h_k^*}{h_k^*} \right) (\tilde{\eta}_{k,MR} - \eta_k^g) \bigg| A_1, \bar{T}_2 \right\}$$

$$= \sum_{k=2}^K E_p \left\{ \nabla_{1,k} (\tilde{\eta}_{k,MR} - \eta_k^g) \bigg| A_1, \bar{T}_2 \right\}$$

$$= \sum_{k=2}^K E_p \left\{ \nabla_{1,k} (\eta_{k,MR} - \eta_k^g) \bigg| A_1, \bar{T}_2 \right\}$$

$$+ \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{k=r_u+1}^K E_p \left\{ \nabla_{1,r_1 \Pi_{MR,r_1,r_2,\ldots,r_u},k} [\eta_{k,MR} - \eta_k^g] \bigg| A_1, \bar{T}_2 \right\}$$

Note that in the preceding expression we used the inductive hypothesis pretending that our study started at cycle 2 instead of cycle 1, i.e. with $(A_1, \bar{T}_2)$ playing the
role of $L_1$, with each $(L_j, A_j)$ playing the role of $(L_{j-1}, A_{j-1})$, and with $\nabla_{1,k}$ playing the role of $\nabla_{0,k}$.

We also have that

$$\tilde{\eta}_{1,MR} - \eta_1^g = \Pi_1 \left[ Q_2 (\hat{h}_2^K, \hat{\eta}_{2,MR}^K) - E_p \left\{ Q_2 (\hat{h}_2^K, \hat{\eta}_{2,MR}^K) \big| A_1, \mathcal{L}_2 \right\} \right]$$

$$+ \Pi_1 \left[ y_{2, \eta_2^g} (A_1, \mathcal{L}_2) \right] - \eta_1^g$$

$$+ \Pi_1 \left[ E_p \left\{ Q_2 (\hat{h}_2^K, \hat{\eta}_{2,MR}^K) \big| A_1, \mathcal{L}_2 \right\} - y_{2, \eta_2^g} (A_1, \mathcal{L}_2) \right]$$

$$= (\eta_{1,MR} - \eta_1^g)$$

$$+ \Pi_1 \left[ \sum_{r=2}^{K} E_p \left\{ \frac{\pi^*(r-1)}{\pi_2^*(r-1)} \left( \tilde{\eta}_{r,MR} - \eta_1^g \right) \left( \frac{h_r^*}{h_r} - \frac{h_r^*}{h_r} \right) \big| A_1, \mathcal{L}_2 \right\} \right]$$

$$= (\eta_{1,MR} - \eta_1^g) + \Pi_1 \left[ \sum_{r=2}^{K} E_p \left\{ \nabla_{1,r} (\tilde{\eta}_{r,MR} - \eta_1^g) \big| A_1, \mathcal{L}_2 \right\} \right]$$

$$= (\eta_{1,MR} - \eta_1^g) + \Pi_1 \left[ \sum_{k=2}^{K} E_p \left\{ \nabla_{1,k} (\eta_k,MR - \eta_1^g) \big| A_1, \mathcal{L}_2 \right\} \right]$$

$$+ \sum_{2 \leq r_1 < r_2 < ... < r_u \leq K-1} \sum_{k=r_u+1}^{K} \Pi_1 \left[ E_p \left\{ \nabla_{1,1} \Pi_{MR, r_1, r_2, ..., r_u, k} (\eta_k,MR - \eta_1^g) \big| A_1, \mathcal{L}_2 \right\} \right]$$

$$= (\eta_{1,MR} - \eta_1^g) + \sum_{k=2}^{K} \Pi_{MR,1,k} [\eta_{k,MR} - \eta_1^g]$$

$$+ \sum_{2 \leq r_1 < r_2 < ... < r_u \leq K-1} \sum_{k=r_u+1}^{K} \Pi_{MR,1,r_1, r_2, ..., r_u, k} [\eta_{k,MR} - \eta_1^g]$$

where the second equality follows after invoking Lemma A.1. So,
\[
\sum_{k=1}^{K} E_p \left\{ \frac{\pi_{k-1}^{*}}{\pi_{k-1}^{*}} \left( \frac{h_k^*}{h_k^*} - \frac{h_k^*}{h_k^*} \right) \left( \eta_{k,MR} - \eta_k^q \right) \bigg| L_1 \right\} \\
= \sum_{k=2}^{K} E_p \left[ h_k^* \left( \eta_{k,MR} - \eta_k^q \right) \left| L_1 \right\} + E_p \left\{ \left( h_k^* \right)^2 \left( \eta_{k,MR} - \eta_k^q \right) \bigg| L_1 \right\} \right] \\
= E_p \left[ h_k^* \left( \eta_{k,MR} - \eta_k^q \right) \left| L_1 \right\} \right] \\
= \sum_{k=1}^{K} E_p \left\{ \nabla_{0,k} \left( \eta_{k,MR} - \eta_k^q \right) \bigg| L_1 \right\} \\
+ \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1 \ k = r_u+1} E_p \left\{ \nabla_{0,1} \Pi_{MR,1,r_1,r_2,\ldots,r_{u},k} \left[ \eta_{k,MR} - \eta_k^q \right] \bigg| L_1 \right\} \\
+ \sum_{k=2}^{K} E_p \left\{ \nabla_{0,1} \Pi_{MR,1,k} \left[ \eta_{k,MR} - \eta_k^q \right] \bigg| L_1 \right\} \\
= \sum_{k=1}^{K} E_p \left\{ \nabla_{0,k} \left( \eta_{k,MR} - \eta_k^q \right) \bigg| L_1 \right\}
\]
This concludes the proof of Theorem 1.
7.5 Multi-layer cross-fitting MR machine learning algorithms

In this section we describe two algorithms for multiple robust estimation of $\theta$ in which, not only $\theta$ but also the nuisance functions $h_k$ and $\eta_k$ are estimated by cross-fit sample splitting, thus avoiding the within sample dependence problem discussed after Result 1 of section 5.2.1. The first is a two-layer cross-fit algorithm and the second is a multi-layer cross-fit algorithm. The first is simpler to implement, as it involves (exponentially in $K$) fewer estimation steps, but it ignores part of the data for estimation of each $h_k$ and $\eta_k$. These algorithms are new, having been developed in April 2017; in contrast, all other results in the paper are from the period 2012-2014.

In order to describe the algorithms it is convenient to establish the following notation and definitions.

Given a finite set $S \subseteq \mathbb{N}$, a random partition of size $U$ of $S$ is a collection

$$\{S_u \subseteq S : 1 \leq u \leq U, \bigcup_{u=1}^{U} S_u = S, S_u \cap S_{u'} = \emptyset \text{ if } u \neq u'\}$$

where for each $u$, $S_u$ is a random subset of $S$. The random partition of size $U$ is generated from the uniform distribution of size $U$ if all possible partitions of size $U$ are equally likely. Each subset $S_u$ of a random partition is called a random split, or simply a split-sample, of the partition. We call the complement set

$$S_u^c \equiv S \setminus S_u$$

the $u^{th}$-random c-split, or simply, a $u^{th}$-c-split-sample. In the sequel, the word partition stands for a random partition generated from the uniform distribution of a given size.

For $k = 0, ..., K^*$, where $K^*$ is any non-negative integer and positive integers $U_1, ..., U_{K^*}$, define the random c-splits $S_{u[1],...,u[k]}^c$, $k = 1, ..., K^*$ recursively as follows.

1. The set $S_{u[1]}^c$ is the $u_{[1]}^{th}$ c-split-sample of a partition of $\{1, ..., n\}$ of size $U_1$, for $u_{[1]} = 1, ..., U_1$.

2. Given $S_{u[1],...,u[k-1]}^c$, the set $S_{u[1],...,u[k]}^c$ is the $u_{[k]}^{th}$ c-split-sample of a partition of $S_{u[1],...,u[k-1]}^c$ of size $U_k$, for $u_{[k]} = 1, ..., U_k$. That is,

$$S_{u[1],...,u[k]}^c = S_{u[1],...,u[k-1]}^c \setminus S_{u[1],...,u[k]}$$

92
where $S_{u[1],...,u[k]}$ is the $u_{[k]}^{th}$ split-sample of a partition of $S_{u[1],...,u[k-1]}^c$ of size $U_k$.

Given $n$ iid copies $(A_{K,i}, L_{K+1})$, $i = 1, ..., n$, of $(\overline{A}_K, \overline{L}_{K+1})$ we call the subsample comprised by the units with indexes $i$ in $S_{u[1],...,u[k]}$ the $(u[1], ..., u[k])^{th}$ split-sample and the subsample comprised by the units in $S_{u[1],...,u[k]}^c$ the $(u[1], ..., u[k])^{th}$ c-split-sample. In an abuse of notation, split-samples are denoted by the sets of indexes associated with the units in the sample. Thus, for instance $S_{u[0],u[1],...,u[k]}$ denotes a specific subset of $\{i : j = 1, ..., J\}$ as well as the subsample $\{(\overline{A}_{K,i}, \overline{L}_{K+1,i}) : j = 1, ..., J\}$ of the sample $\{(\overline{A}_{K,i}, \overline{L}_{K+1,i}) , i = 1, ..., n\}$.

**MR two-layer cross-fit ALGORITHM with first layer sample split of size $U$**

Compute the split samples $S_{u[1]}$ and $S_{u[1],u[2]}$, $1 \leq u[1] \leq U$ and $1 \leq u[2] \leq K$, corresponding to random splits of sizes $U$ and $K$ respectively.

Let $\tilde{Q}_{K+1, mach} \equiv \psi (\overline{L}_{K+1})$. For $u[1] = 1, 2, ..., U$,

\[
\begin{cases}
\text{for } k = K, K - 1, ..., 1,
\begin{cases}
\text{i) using the units in the } (u[1], u[2] = k)^{th} \text{ split-sample } S_{u[1],u[2]=k} \text{ compute } \\
\quad \hat{h}_k \text{, the output from a preferred machine learning algorithm for estimating } h_k. \text{ For } r \in \{k, k + 1, ..., K\}, \text{ define } \hat{\pi}_r \equiv \prod_{j=k}^{r} \hat{h}_j. \text{ Also, for units in the } (u[1], u[2] = k)^{th} \text{ split-sample } S_{u[1],u[2]=k} \text{ that have } \pi^* > 0 \text{ compute } \tilde{\eta}_k (\cdot, \cdot) \text{, the output of a preferred machine learning algorithm for estimating } E \left( \tilde{Q}_{k+1, mach} \mid \overline{A}_k, \overline{L}_k \right). \\
\text{ii) for each unit in the } (u[1], u[2] = k - 1)^{th} \text{ split-sample } S_{u[1],u[2]=k-1} \text{ that has } \pi^* > 0, \text{ compute } \\
\quad \tilde{Y}_k \equiv y_{k, \tilde{\eta}_k} (\overline{A}_{k-1}, \overline{L}_k) \\
\quad \quad \equiv \int h_k^* (a_k \mid \overline{A}_{k-1}, \overline{L}_k) \tilde{\eta}_k (\overline{A}_{k-1}, a_k, \overline{L}_k) \, d\mu_k (a_k).
\end{cases}
\end{cases}
\]

93
and
\[
\tilde{Q}_k \equiv Q_k \left( \frac{\pi^K}{\hat{h}_k}, \tilde{\eta}_k \right) \\
\equiv \frac{\pi^{*K}}{\tilde{\pi}^{K}_k} \psi \left( I_{K+1} \right) - \\
\left\{ \sum_{j=k}^K \frac{\pi^{*j}_k}{\tilde{\pi}^{j}_k} \tilde{\eta}_{j,mach} \left( \bar{A}_j, \bar{L}_j \right) - \frac{\pi^{*j-1}_k}{\tilde{\pi}^{j-1}_k} y_j \tilde{\eta}_j \left( \bar{A}_j-1, \bar{L}_j \right) \right\} \\
\equiv y_k \tilde{\eta}_{k,mach} \left( \bar{A}_{k-1}, \bar{L}_k \right) + \\
\sum_{j=k}^K \frac{\pi^{*j}_k}{\tilde{\pi}^{j}_k} \left\{ y_{j+1} \tilde{\eta}_{j+1} \left( \bar{A}_j, \bar{L}_{j+1} \right) - \tilde{\eta}_j \right\}
\]

where \( \pi^{*k-1}_k \equiv 1 \) and \( \hat{\pi}^{k-1}_k \equiv 1 \) and the \((u[1], u[2] = 0)^{th}\) split-sample \( S_{u[1], u[2] = 0} \) is defined to be equal to the \( u^{th}\) split-sample \( S_{u[1]} \).

Let \( \tilde{\theta}_{MR, two-layer} \) be the average of \( \tilde{Q}_1 \) based on units in the \( u^{th}\) split sample \( S_{u[1]} \).

Finally, compute
\[
\tilde{\theta}_{MR, two-layer} \equiv \frac{1}{U} \sum_{u[1]=1}^U \tilde{\theta}_{MR, two-layer}^{u[1]}
\]

MR multi-layer cross-fit ALGORITHM with sample splits at each layer of size \( U \)

For \( k = 1, ..., K \), recursively calculate the random split samples \( S_{u[1], u[2]..., u[K]} \), \( (u[1], u[2], ..., u[K]) \in \{1, ..., U\}^K \).

a) For each \( (u[1], ..., u[K]) \in \{1, 2, ..., U\}^K \),
i) using the units in the \((u_{[1]}, \ldots, u_{[K]})^t\) c-split-sample \(S^c_{u_{[1]}, \ldots, u_{[K]}}\), compute 
\(\hat{h}_K(u_{[1]}, \ldots, u_{[K]})\), the output from a preferred machine learning algorithm for estimating \(h_K\). Define \(\hat{\eta}_K(u_{[1]}, \ldots, u_{[K]}) \equiv \hat{h}_K(u_{[1]}, \ldots, u_{[K]})\). Also for any
\(k \in \{1, \ldots, K - 1\}\), compute
\[\hat{h}_K(u_{[1]}, \ldots, u_{[K]}) \equiv \frac{1}{U^{K-k}} \sum_{i_{k+1}, i_{k+2}, \ldots, i_K=1}^U \hat{h}_K(u_{[1]}, \ldots, u_{[k]}=i_{k+1}, \ldots, u_{[K]}=i_K).
\]

ii) using the units in the \((u_{[1]}, \ldots, u_{[K]})^t\) c-split-sample \(S^c_{u_{[1]}, \ldots, u_{[K]}}\) that have \(\pi^* K > 0\), compute \(\tilde{\eta}_K(u_{[1]}, \ldots, u_{[K]})\), the output from a preferred machine learning algorithm for estimating \(E(\psi(\overline{T}_K)) | \overline{A}_K, \overline{T}_K\). Also for any \(k \in \{1, \ldots, K - 1\}\), compute
\[\tilde{\eta}_K(u_{[1]}, \ldots, u_{[K]}) \equiv \frac{1}{U^{K-k}} \sum_{i_{k+1}, i_{k+2}, \ldots, i_K=1}^U \tilde{\eta}_K(u_{[1]}, \ldots, u_{[k]}=i_{k+1}, \ldots, u_{[K]}=i_K).
\]

b) For \(k = K, \ldots, 1\),

\{ 
for each \((u_{[1]}, \ldots, u_{[k]}) \in \{1, 2, \ldots, U\}^k\),

i) if \(k \neq 1\), using the units in the \((u_{[1]}, \ldots, u_{[k]})^t\) split-sample \(S_{u_{[1]}, \ldots, u_{[k]}}\),
compute \(\hat{h}_{k-1}(u_{[1]}, \ldots, u_{[k]})\), the output from a preferred machine learning algorithm for estimating \(h_{k-1}\). Also, if \(k < K\) then for
\(r \in \{k, k + 1, \ldots, K - 1\}\), compute \(\hat{h}_r(u_{[1]}, \ldots, u_{[k]}) = \frac{1}{U^{r-k+1}} \sum_{i_{k+1}, i_{k+2}, \ldots, i_r, i_{r+1}=1}^U \hat{h}_r(u_{[1]}, \ldots, u_{[k]}=i_{k+1}, \ldots, u_{[r]}=i_{r+1})\)
and for
\(k - 1 \leq s \leq r \leq K\), compute \(\hat{\pi}_s(u_{[1]}, \ldots, u_{[k]}), r = \prod_{j=s}^r \hat{h}_j(u_{[1]}, \ldots, u_{[k]}).\)

ii) For each unit in the \((u_{[1]}, \ldots, u_{[k]})^t\) split sample \(S_{u_{[1]}, \ldots, u_{[k]}}\) that has \(\pi^* K > 0\), compute
\[\hat{Y}_k(u_{[1]}, \ldots, u_{[k]}) \equiv y_{k, \hat{\eta}_k(u_{[1]}, \ldots, u_{[k]})} \overline{A}_{k-1}, \overline{T}_k \]
\[= \int h^*_k(a_k, \overline{A}_{k-1}, \overline{T}_k) \tilde{\eta}_k(u_{[1]}, \ldots, u_{[k]}) \overline{A}_{k-1}, a_k, \overline{T}_k \) \(d\mu_k(a_k)\).

95
and

\[ Q_k(u_{[1]}, ..., u_{[k]}) = Q_k \left( \frac{H_k(u_{[1]}, ..., u_{[k]}), \tilde{\eta}_k(u_{[1]}, ..., u_{[k]}), K}{\tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), K} \right) \]

\[ \equiv \frac{\pi_k + K}{\tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), K} \psi \left( \tilde{L}_{K+1} \right) \]

\[ - \sum_{j=k}^{K} \left\{ \frac{\pi_k j}{\tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), j} \tilde{\eta}_j(u_{[1]}, ..., u_{[k]}) (A_j, \tilde{L}_j) - \right\} \]

\[ \left\{ \frac{\pi_k j-1}{\tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), j-1} y_{j-1, \tilde{\eta}_j+1}^{(u_{[1]}, ..., u_{[k]})} (A_{j-1}, \tilde{L}_j) - \tilde{\eta}_{j-1}^{(u_{[1]}, ..., u_{[k]})} \right\} \]

\[ \equiv y_{k, \tilde{\eta}_k}^{(u_{[1]}, ..., u_{[k]})} (A_{k-1}, \tilde{L}_k) \]

\[ + \sum_{j=k}^{K} \frac{\pi_k j}{\tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), j} \left\{ y_{j+1, \tilde{\eta}_j+1}^{(u_{[1]}, ..., u_{[k]})} (A_{j+1}, \tilde{L}_{j+1}) - \tilde{\eta}_{j+1}^{(u_{[1]}, ..., u_{[k]})} \right\} \]

where \( \pi_k^{*k-1} \equiv 1 \) and \( \tilde{\pi}_k(u_{[1]}, ..., u_{[k]}), k-1 \equiv 1 \).

iii) If \( k = 1 \), then using data in the \( u_{[1]} \) split sample \( \mathcal{S}_{u_{[1]}} \), compute

\[ \tilde{\theta}_{MR,\text{multi-layer}}^{u_{[1]}} = p_{u_{[1]}} \left\{ \tilde{Q}_1(u_{[1]}) \right\}, \]

the average of \( \tilde{Q}_1(u_{[1]}) \) in the \( u_{[1]} \) split sample \( \mathcal{S}_{u_{[1]}} \); otherwise using data in the \( (u_{[1]}, ..., u_{[k]}) \) split sample \( \mathcal{S}_{u_{[1]}, u_{[2]}, ..., u_{[k]}} \), compute \( \tilde{\eta}_{k-1}^{(u_{[1]}, ..., u_{[k]})} \), the output of a preferred machine learning algorithm for estimating \( E \left( \tilde{Q}_k(u_{[1]}, ..., u_{[k]}) \mid A_{k-1}, \tilde{L}_{k-1} \right) \). Also, if \( 1 < k < K \), then for \( r \in \{k, k+1, ..., K-1\} \), compute

\[ \tilde{\eta}_r^{(u_{[1]}, ..., u_{[k]})} = \frac{1}{\sum_{i_{k+1}, i_{k+2}, ..., i_r, i_{r+1}}} \tilde{\eta}_r^{(u_{[1]}, ..., u_{[k]}, u_{[k+1]}=i_{k+1}, ..., u_{[r+1]}=i_{r+1})}. \]
Finally, compute

\[ \hat{\theta}_{MR,\text{multi-layer}} = \frac{1}{U} \sum_{u[i]=1}^{U} \hat{\theta}_{MR,\text{multi-layer}}^{u[i]} \]
Figure 1: Illustration of the two-layer cross-fit algorithm for $K = 3$ and $U = 5$.

\[
\hat{\theta}_{\text{MR-two-layer}} = \frac{1}{5} \sum_{t=1}^{5} \theta_{\text{MR-two-layer}}^{U[t]=t}
\]
Figure 2: Illustration of multi-layer cross-fit algorithm for $K = 3$ and $U = 5$. 

\[
\theta_{\text{Mx-multi-layer}} = \frac{1}{S} \sum_{i=1}^{S} \theta_{\text{Mx-multi-layer}}^{i}
\]
Second layer computations for $\eta$ functions

Same splitting and cross-fit strategy for estimating $\eta$ functions, not shown here

Third and fourth layer computations for $\eta$ functions

Same splitting and cross-fit strategy for estimating $\eta$ functions, not shown here

Divide the c-split $S_{ij1}=1$, $a_{ij2}=1$ into 5 second level splits

$S_{ij1}=1$, $a_{ij2}=1$
compute $\eta^c_{i1}$ here

$S_{ij1}=1$, $a_{ij2}=1$
compute $\eta^c_{i2}$ here

$S_{ij1}=1$, $a_{ij2}=1$
compute $\eta^c_{i3}$ here

$S_{ij1}=1$, $a_{ij2}=1$
compute $\eta^c_{i4}$ here

$S_{ij1}=1$, $a_{ij2}=1$
compute $\eta^c_{i5}$ here

$\eta_1^c=\frac{1}{2}\sum_{i=1}^{5} \eta_{i1}^c = \frac{1}{2}\sum_{i=1}^{5} \eta_{i2}^c$

$\eta_2^c=\frac{1}{2}\sum_{i=1}^{5} \eta_{i3}^c = \frac{1}{2}\sum_{i=1}^{5} \eta_{i4}^c$

$\eta_3^c=\frac{1}{2}\sum_{i=1}^{5} \eta_{i5}^c = \frac{1}{2}\sum_{i=1}^{5} \eta_{i6}^c$