University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 317

Targeted Learning of an Optimal Dynamic Treatment, and Statistical Inference for its Mean Outcome

Mark J. van der Laan*

*Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

Targeted Learning of an Optimal Dynamic Treatment, and Statistical Inference for its Mean Outcome

Mark J. van der Laan

Abstract

Suppose we observe n independent and identically distributed observations of a time-dependent random variable consisting of baseline covariates, initial treatment and censoring indicator, intermediate covariates, subsequent treatment and censoring indicator, and a final outcome. For example, this could be data generated by a sequentially randomized controlled trial, where subjects are sequentially randomized to a first line and second line treatment, possibly assigned in response to an intermediate biomarker, and are subject to right-censoring. In this article we consider estimation of an optimal dynamic multiple time-point treatment rule defined as the rule that maximizes the mean outcome under the dynamic treatment, where the candidate rules are restricted to only respond to a user-supplied subset of the baseline and intermediate covariates. This estimation problem is addressed in a statistical model for the data distribution that is nonparametric beyond possible knowledge about the treatment and censoring mechanism, while still providing statistical inference for the mean outcome under the optimal rule. This contrasts from the current literature that relies on parametric assumptions.

For the sake of presentation, we first consider the case that the treatment/censoring is only assigned at a single time-point, and subsequently, we cover the multiple time-point case. We characterize the optimal dynamic treatment as a statistical target parameter in the nonparametric statistical model, and we propose highly data adaptive estimators of this optimal dynamic regimen, utilizing sequential loss-based super-learning of sequentially defined (so called) blip-functions, based on newly proposed loss-functions. We also propose a cross-validation selector (among candidate estimators of the optimal dynamic regimens) based on a cross-validated targeted minimum loss-based estimator of the mean outcome under the

candidate regimen, thereby aiming directly to select the candidate estimator that maximizes the mean outcome. We also establish that the mean of the counterfactual outcome under the optimal dynamic treatment is a pathwise differentiable parameter, and develop a targeted minimum loss-based estimator (TMLE) of this target parameter. We establish asymptotic linearity and statistical inference based on this targeted minimum loss-based estimator under specified conditions. In a sequentially randomized trial the statistical inference essentially only relies upon a second order difference between the estimator of the optimal dynamic treatment and the optimal dynamic treatment to be asymptotically negligible, which may be a problematic condition when the rule is based on multivariate time-dependent covariates. To avoid this condition, we also develop targeted minimum loss based estimators and statistical inference for data adaptive target parameters that are defined in terms of the mean outcome under the $\{\backslash$em estimate$\}$ of the optimal dynamic treatment.

In particular, we develop a novel cross-validated TMLE approach that provides asymptotic inference under minimal conditions, avoiding the need for any empirical process conditions. For the sake of presentation, in the main part of the article we focus on two-time point interventions, but the results are generalized to general multiple time point interventions in the appendix.

1 Introduction

Suppose we observe n independent and identically distributed observations of a time-dependent random variable consisting of baseline covariates, initial treatment and censoring indicator, intermediate covariates, subsequent treatment and censoring indicator, and a final outcome. For example, this could be data generated by a sequentially randomized controlled trial in which one follows up a group of subjects, and treatment assignment at two time-points is sequentially randomized, where the probability of receiving treatment might be determined by a baseline covariate for the first-line treatment, and time-dependent intermediate covariate (such as a biomarker of interest) for the second- line treatment (Robins (1986)). Such trials are often called sequential multiple assignment randomized trials (SMART). A dynamic treatment rule is a rule that deterministically assigns treatment as a function of the available history. If treatment is assigned at two time points, then this dynamic treatment rule consists of two rules, one for each time point (Robins (1986, 2000, 1993, 1997)). The mean outcome under a dynamic treatment is a counterfactual quantity of interest representing what the mean outcome would have been if everybody would have received treatment according to the dynamic treatment rule (Neyman, 1990; Rubin, 1974, 2006; Holland, 1986; Robins, 1987a,b; Pearl, 2000). Dynamic treatments represent pre-specified multiple time-point interventions that at each treatment-decision stage are allowed to respond to the currently available treatment and covariate history. Examples of multiple time-point dynamic treatment regimens are given in Lavori and Dawson (2000, 2008); Murphy (2005); Rosthj et al. (2006); Thall et al. (2000, 2002); Wagner et al. (2001); Petersen et al. (2007); van der Laan and Petersen (2007); Robins et al. (2008) ranging from rules that change the dose of a drug, change or augment the treatment, to making a decision on when to start a new treatment, in response to the history of the subject.

More recently, SMART designs have been implemented in practice: Lavori and Dawson (2000, 2004); Murphy (2005); Thall et al. (2000); Chakraborti et al. (2010); Kasari (2009); Lei et al. (2011); Nahum-Shani et al. (2012a,b); Jones (2010); Lei et al. (2011). For an extensive list of SMARTs, we refer the reader to the website http://methodology.psuedu/ra/adap-treat-strat/projects. For an excellent and recent overview on the literature on dynamic treatments we refer to Chakraborti and Murphy (2013). Researchers have also aimed to learn dynamic treatments from observational studies: Cotton and Heagerty (2011); Orellana et al. (2010a); Robins et al. (2008); Rosthj et al. (2006); van der Laan and Petersen (2007); Petersen et al. (2008, 2007); Moodie et al. (2009). These observational and sequentially randomized studies provide an

opportunity to learn an optimal multiple time-point dynamic treatment defined as the treatment rule that maximizes the mean dynamic-regimen specific counterfactual outcome over a user supplied class of dynamic regimens. The reinforcement learning (i.e., computer science) and statistical literature have made enormous advances in developing statistical methods that aim to learn such optimal rules.

The literature on Q-learning defines the optimal dynamic treatment among *all* dynamic treatments in a sequential manner (Sutton and Sung (1998); Murphy (2003); Robins (2003, 2004); Murphy (2005)): considering a two stage SMART, the optimal treatment rule for the second line treatment is defined as the maximizer of the conditional mean outcome, given the observed past, over the possible second line treatments, and the optimal treatment rule for the first line treatment is defined as the maximizer of the conditional mean counterfactual outcome, given baseline covariates, over the possible values for the initial treatment, under the assumption that the second line treatment is assigned according to the just determined optimal rule for the second line treatment. This characterization of the optimal treatment has its roots in multi-stage decision theory and can be thought of as an example of dynamic programming (Bellman, 1957). This optimal rule can be learned through fitting the likelihood and then just calculating the optimal rule under this fit of the likelihood.. This approach can be implemented with maximum likelihood estimation based on parametric models. Since there is no need to fit the whole likelihood, one can focus on just fitting the sequential regressions, such as sequential linear least squared regression (see e.g., Murphy (2005)), while Ernst et al. (2005) and Ormoneit and Sen (2002) use regression trees and kernel regression estimators, respectively. Moodie et al. (2012) proposes inverse propensity score weighting of the regressions in Q-learning. Q-learning is not limited to particular type of regression models or outcomes: e.g., Goldberg and Kosorok (2012); Zhao et al. (2011) apply Q-learning to the survival outcome setting.

It has been noted (e.g., Robins (2004), Chakraborti and Murphy (2013)) that the estimator of the parameters of one of the regressions (except the first one) when using parametric regression models is a non-smooth function of the estimator of the parameters of the previous regression, and that this results in non-regularity of the estimators of the parameter vector. This raises challenges for obtaining statistical inference, even when assuming that these parametric regression models are correctly specified. Chakraborti and Murphy (2013) discuss various approaches and advances that aim to resolve this delicate issue such as inverting hypothesis testing (Robins (2004)), establishing non-normal limit distributions of the estimators (E. Laber, D. Lizotte, M. Qian, S. Murphy,

submitted), or using the m out of n bootstrap.

(Murphy, 2003; Robins, 2003, 2004) develop so called structural nested mean models tailored for optimal dynamic treatments. These models assume a parametric model for the "blip-function" defined as the additive effect of a blip in current treatment on a counterfactual outcome, conditional on the observed past, in the counterfactual world in which future treatment is assigned optimally. Each blip-function defines the optimal treatment rule for that time point by simply maximizing it over the treatment, so that knowing the blip functions, allows one to calculate the optimal dynamic treatment by starting with maximizing the last blip function and moving backwards in time till the first time point. In the original formulation of structural nested mean models the future treatment in the blip-functions was set equal to some baseline treatment (Robins, 2000), while Murphy (2003) and Robins (2003, 2004) generalized this class of models to structural nested mean models tailored for estimation of optimal dynamic regimens, by defining the future treatment as the optimal treatment. These models are semi-parametric since they only rely on a parametric model of the blip function (at least in a SMART), but they aim to leave the nuisance parameters unspecified. These authors develop estimators for the unknown parameters of the blip-functions using estimating equation methodology. The estimated blip functions now define an estimator of the optimal rule. Statistical inference for the parameters of the blip function proceeds accordingly, but Robins (2004) points out the irregularity of the estimator, resulting in some serious challenges for statistical inference as referenced above.

Structural nested mean models have also been generalized to blip functions that condition on a (counterfactual) subset of the past, thereby allowing the learning of optimal rules that are restricted to only using this subset of the past (Robins (2004) and section 6.5 in van der Laan and Robins (2003)).

An alternative approach, referenced as the direct approach in Chakraborti and Murphy (2013), uses marginal structural models for the dynamic regimen specific mean outcome for a user supplied class of dynamic treatments. If one assumes the marginal structural models are correctly specified, then the parameters of the marginal structural model map into a dynamic treatment that is optimal among the user supplied class of dynamic regimens. In addition, the MSM also provides the complete dose-response curve, i.e. the mean counterfactual outcome for each dynamic treatment in the user-supplied class. This generalization of the original marginal structural models for static interventions to MSMs for dynamic treatments were developed independently by (Orellana et al., 2010a; van der Laan and Petersen, 2007). These articles present inverse probability of treatment and censoring weighted (IPCW)

estimators and double robust augmented IPCW-estimators based on general longitudinal data structures, allowing for right-censoring, time-dependent covariates and survival outcomes, and these articles also include data analysis examples learning the optional rule for when to switch treatment based on CD4-count. Double robust estimating equation based methods that estimate the nuisance parameters with sequential parametric regression models using clever covariates were developed for MSMs for static interventions in Bang and Robins (2005) and and analogue targeted minimum loss-based estimator (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2012) for marginal structural models for a user supplied class of dynamic treatments was developed in Petersen et al. (2013) building on the TMLE for the mean outcome for a single dynamoc treatment developed in van der Laan and Gruber (2012). Additional application papers of interest are (Hernan et al., 2006; Cotton and Heagerty, 2011; Shortreed and Moodie, 2012) which involve fitting MSMs for dynamic treatments defined by treatment-tailoring threshold using IPCW methods.

Each of the above referenced approaches for learning an optimal dynamic treatment rely on parametric assumptions: even the structural nested mean models and the marginal structural models both rely on parametric models for the blip-function and dose-response curve, respectively. As a consequence, even in a SMART, the statistical inference for the optimal dynamic treatment heavily relies on assumptions that are generally believed to be false, and will thus be expected to be biased. Therefore, in this article, we aim to avoid such assumptions and instead define the semi parametric statistical model for the data distribution as nonparametric, beyond the possible knowledge on the treatment mechanism (e.g., known in a RCT) and censoring mechanism. This forces us to define the optimal dynamic treatment and the corresponding mean outcome as parameters defined on this nonparametric model, and to develop data adaptive estimators of the optimal dynamic treatment. In order to not only consider the most ambitious fully optimal rule, we define V-optimal rules as the optimal rule that only use a user-supplied subset V of the available covariates. This allows us to consider sub-optimal rules that are easier to estimate and thereby allow for statistical inference for the counterfactual mean outcome under the sub-optimal rule, i.e., analogue to the generalized structural nested mean models whose blip-functions only condition on a counterfactual subset of the past.

Our estimators of the blip-functions (that also define the V-optimal rule) are based on sequential (analogue to Q-learning) loss-based super-learning (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2012) which involves the application of a

super-learner to fit each of the blip-functions that are defined after having fitted the "previous" blip functions. The super-learner is defined by generating a family of candidate estimators, a risk for each candidate estimator, and selection among all candidate estimators based on a cross-validation based estimator of this risk. Some of these candidate estimators could be based on parametric models of the blip-functions (as in a structural nested mean model), while others are based on available machine learning algorithms. By previously established oracle inequality results on the cross-validation selector established in the above mentioned references, our results guarantee that in SMART the super-learner will be asymptotically equivalent with the estimator selected by the oracle selector (selecting the best) and thereby outperforms any of the parametric model based estimators and any of the other estimators in the family of candidate estimators, under the assumption that non of the parametric models are correctly specified, while it achieves the same rate of convergence as the correctly specified parametric model otherwise. In this manner, our sequential super-learner is at each stage doing an asymptotically optimal job in fitting the blip-function relative to its user supplied class of candidate estimators. Practical findings strongly suggest that this will also result in superior performance in most practical situations relative to sticking to one particular estimation procedure (Polley et al., 2012; van der Laan and Rose, 2012).

Such an estimator of the blip-functions could be substituted in the formula for the optimal dynamic treatment in terms of these blip-function. We also propose a cross-validation selector that selects among candidate estimators of the optimal dynamic treatment based directly on the performance of the candidate rule (instead of the performance in fitting the blip-function). For that purpose, we use cross-validation based on a loss-function whose risk equals the mean outcome under the candidate rule, and we discuss oracle inequalities for this cross-validation selector whose loss-based dissimilarity equals the mean outcome under the candidate rule minus the mean outcome under the optimal rule. We also develop cross-validated targeted minimum loss-based estimator of the same risk in order to improve finite sample performance of the cross-validation selector. In this manner, we target the fine-tuning of the fit of the optimal rule w.r.t a measure of performance that directly measures the performance of the rule in minimizing the mean outcome, the very measure that defines the optimal rule. In particular, this cross-validation selector could be used to select among different candidate estimators of the optimal dynamic treatment indexed by a choice of estimator of the blip-functions.

Since we are not assuming parametric models, we are not able to obtain statistical inference for these optimal rules, although the proposed cross-validated

risks for the blip-function at each stage provides a mean to assess the practical performance of these blip-functions and thereby indirectly the corresponding rules. However, we will show that the mean outcome under the optimal rule is a pathwise differentiable parameter of the data distribution, indicating that it is possible to develop asymptotically linear estimators of this target parameter under conditions. In fact, we obtain the surprising result that the pathwise derivative of this target parameter equals the pathwise derivative of the mean counterfactual outcome under a given dynamic treatment rule set at the optimal rule, treating the latter as known. By a reference to the current literature for double robust and efficient estimation of the mean outcome under a given rule, we then obtain a targeted minimum loss-based estimator for the mean outcome under the optimal rule. Subsequently, we prove asymptotic linearity and efficiency of this TMLE, allowing us to construct confidence intervals for the mean outcome under the optimal dynamic treatment or its contrast w.r.t. a standard treatment. Thus contrary to the irregularity of the estimators of the unknown parameters in the semi parametric structural nested mean model, we can construct regular estimators of the mean outcome under the optimal rule in the nonparametric model.

In a SMART the statistical inference would only rely upon a second order difference between the estimator of the optimal dynamic treatment and the optimal dynamic treatment itself to be asymptotically negligible. This is a reasonable condition if we restrict ourselves to rules only responding to a one dimensional time-dependent covariate, or if we are willing to make smoothness assumptions. To avoid this condition, we also develop targeted minimum loss based estimators and statistical inference for data adaptive target parameters that are defined in terms of the mean outcome under the *estimate* of the optimal dynamic treatment (see van der Laan et al. (2013) for a general approach for statistical inference for data adaptive target parameters). In particular, we develop a novel cross-validated TMLE approach that provides asymptotic inference under minimal conditions.

For the sake of presentation, we focus on two-time point treatments in the main part of the article. In the Appendix we generalize these results to general multiple time point treatments, and develop general (sequential) super-learning based on the efficient cross-validated TMLE of the risk of a candidate estimator. In the appendix we also develop a TMLE of a projection of the blip functions on a parametric working model (with corresponding statistical inference), which can be used as candidate estimators in our super-learners, but also present a result of interest in its own right.

1.1 Organization of article

The first part of this article concerns estimation and statistical inference for the optimal dynamic treatment for a single time-point treatment. This estimation problem in the context of randomized controlled trials, a binary outcome, targeting the optimal rule using all the covariates, was handled through loss-based super learning of the conditional additive effect of treatment given all the baseline covariates in Polley and van der Laan (2009). In (Qian and Murphy, 2011; Zhao et al., 2012) it was shown that the estimation of the optimal dynamic treatment can be reduced to a classification problem. Rubin and van der Laan (2012) identifies an entire family of such reductions to classification for the binary outcome problem, and proposed a more efficient reduction. In this paper, we target the blip-function that indirectly identifies the optimal treatment, and target the optimal rule directly (where the IPCW-loss relates to the classification problem formulation), which is also the approach we follow for multiple time-point interventions.

In Section 2 we define the V-optimal rule for the point-treatment data structure, and present the formal estimation problem to be addressed: i.e. data adaptive estimation of the V-optimal rule, and statistical inference for the mean counterfactual outcome under the V-optimal rule. In Section 3 we present a data adaptive loss-based estimation procedure for the V-optimal rule, using super-learning. The super-learner is based on a cross-validated estimator of a measure of performance of the blip-function. In Section 4 we define a super-learner based on a cross-validated estimator of the mean outcome of the rule implied by the blip-function. In Section 5 we study the mean counterfactual outcome under the V-optimal treatment as a statistical target parameter, establish pathwise differentiability with known canonical gradient/efficient influence curve, and obtain a closed form expression of the expectation of the efficient influence curve at misspecified nuisance parameters. The latter (generalized double robustness results) provides a key ingredient for analyzing the TMLE. In Section 6 we present the TMLE of this target parameter, and in Section 7 we present a formal theorem establishing asymptotic linearity of the TMLE and corresponding confidence intervals based on this TMLE. The implications of this theorem for the analysis of RCTs are discussed.

The second part of this article covers the two-time point treatment case, and thereby in essence the multiple time point treatment case. This part is organized in the same manner: Section 8 defines the estimation problem; Section 9 presents a sequential data adaptive loss-based super learner of the blip-functions for the V-optimal treatment rule, and a super-learner based on a cross-validated estimator of the mean outcome under the candidate rule

implied by the candidate blip function; Section 10 establishes the desired pathwise differentiability of the mean counterfactual outcome under the V-optimal rule and a closed form expression of the expectation of the efficient influence curve under misspecified nuisance parameters; Section 11 presents the TMLE of this target parameter; and Section 12 presents an asymptotic linearity theorem for this TMLE and corresponding statistical inference.

The third part of this article concerns statistical inference for data adaptive target parameters that are defined in terms of the mean outcome under the *estimate* of the optimal dynamic treatment, thereby avoiding the consistency and rate condition for the fitted V-optimal rule as required for asymptotic linearity of the TMLE of the mean outcome under the actual V-optimal rule. Firstly, in Section 13 we present the asymptotic linearity theorem for the TMLE for the mean outcome under the actual fitted dynamic treatment regimen. In Section 14 we present a cross-validated TMLE (CV-TMLE) approach that provides asymptotic inference under minimal conditions for the mean outcome under a dynamic treatment fitted on a training sample, averaged across the different splits in training sample and validation sample. Both results allow us to construct confidence interval that have the correct asymptotic coverage of the random true target parameter, but statistical inference based on the CV-TMLE avoids an empirical process condition that can put a brake on the allowed data adaptivity of the estimator.

Section 15 concludes with a summary, and some remarks, indicating possible directions for future research. The Appendix describes the generalization of the two time-point treatment case to the general case and studies optimal estimation of the risk for a candidate estimator of the V-optimal rule resulting in sequential super-learning based on a CV-TMLE of the risk. In the appendix we also develop the TMLE of the projection of the blip-functions on a user-supplied parametric working model.

2 Formulation of optimal dynamic treatment estimation problem: single time-point treatment

Suppose we observe n independent and identically distributed copies $O_1, \ldots, O_n$ of $O = (W, A, Y) \sim P_0$, where W are baseline-covariates, $A = (A_1, A_2) \in \{0, 1\}^2$ is a subsequently assigned binary treatment A_1 and missing indicator A_2, and Y is a final outcome of interest. Consider a model that makes no assumptions on the marginal distribution $Q_{W,0} = Q_W(P_0)$ of W and the con-

ditional distribution $Q_{Y,0} = Q_Y(P_0)$ of Y, given A, W, but might assume a model on the conditional distribution $g_0 = g(P_0)$ of A, given W. In particular, the data might be generated by a randomized controlled trial in which the outcome is not subject to missingness, in which case g_0 is known. Let's denote the collection of possible probability distributions of O with $\mathcal{M}$, which we refer to as the statistical model for the true data distribution P_0. Let $\bar{Q}_0 \equiv E_{P_0}(Y \mid A, W)$ denote the conditional mean of Y, given A, W.

Let V be a function of W. Define the blip-function

$$\begin{aligned}\bar{Q}_0(V) &\equiv E_{P_0}(E_{P_0}(Y \mid A_1 = 1, A_2 = 1, W) - E_{P_0}(Y \mid A_1 = 0, A_2 = 1, W) \mid V) \\ &= E_{P_0}(\bar{Q}_0(1, 1, W) - \bar{Q}_0(0, 1, W) \mid V).\end{aligned}$$

This parameter of P_0 generates an optimal treatment rule $V \to d_0(V) \in \{0, 1\} \times \{1\}$ for assigning treatment and missing indicator defined as

$$d_0(V) \equiv (I(\bar{Q}_0(V) > 0), 1).$$

Under a causal model, such as the Neyman-Rubin model (Neyman (1990); Rubin (1974, 2006); Holland (1986); Robins (1987a,b)), or the structural causal model (Pearl, 2000), which allows the representation of the observed data as a missing data structure $(W, A, Y = Y(A))$ on the counterfactuals

$$X = (W, Y(0, 0), Y(1, 0), Y(0, 1), Y(1, 1)),$$

and assumes A is independent of $Y(a)$, given W, for $a \in \{0, 1\}^2$ (i.e., randomization assumption), we have

$$\bar{Q}_0(V) = E_{P_0}(Y(1, 1) - Y(0, 1) \mid V),$$

and thus

$$d_0(V) \equiv (I(E_{P_0}(Y(1, 1) - Y(0, 1) \mid V) > 0), 1).$$

In this case, d_0 assigns the *causally* optimal treatment based on the baseline covariates V and assigns "no missingness". It follows that

$$d_0 = \arg\max_{d \in \mathcal{D}} E_{P_0} Y_d$$

is the rule that maximizes the mean outcome over all possible dynamic treatments $\mathcal{D}$ that are only a function of V and that assign $A(2) = 1$.

Beyond estimation of this V-optimal rule d_0, in this article we are also concerned with statistical estimation and inference for $E_0 Y_{d_0}$ which is represented by the following statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}$, defined as

$$\Psi(P_0) = E_{P_0}\left\{I(\bar{Q}_0(V)) > 0)\bar{Q}_0(1, 1, W) + I(\bar{Q}_0(V) \leq 0)\bar{Q}_0(0, 1, W)\right\}.$$

One can write this as $\Psi(P_0) = E_{Q_{W,0}}\bar{Q}_0(d_0(V), W)$. We will also denote this parameter with $\Psi(Q_0)$, where $Q_0 = (\bar{Q}_0, Q_{W,0})$ is the relevant part of the data distribution P_0 this statistical target parameter depends upon. Sometimes, we will also denote it with $\Psi(d_0, \bar{Q}_0, Q_{W,0})$ to emphasize the dependence on the rule d_0, $\bar{Q}_0(A, W)$, and $Q_{W,0}$. The definition of $\Psi(P_0)$ relies on the so called positivity assumption that $g_0(1, 1 \mid W) > 0$ and $g_0(0, 1 \mid W) > 0$ a.e., since otherwise the rule $d_0(V)$ is not defined.

Under the causal model and randomization assumption we have

$$\Psi(P_0) = E_{P_0}Y_{d_0} = \max_{d \in \mathcal{D}} E_{P_0}Y_d,$$

where the maximum is over all rules that are functions of V and assign $A(1) = 1$.

By using that $I(\bar{Q}_0(V) \leq 0) = (1 - I(\bar{Q}_0(V) > 0))$, it follows that we can also represent Ψ as:

$$\begin{aligned} \Psi(Q_0) &= E_{P_0}\bar{Q}_0(0, 1, W) + E_{P_0}I(\bar{Q}_0(V) > 0)\{\bar{Q}_0(1, 1, W) - \bar{Q}_0(0, 1, W)\} \\ &= E_{P_0}\bar{Q}_0(0, 1, W) + E_{P_0}I(\bar{Q}_0(V) > 0)\bar{Q}_0(V). \end{aligned} \quad (1)$$

The estimation problem is defined: we observe n i.i.d. copies of $O = (W, A, Y) \sim P_0 \in \mathcal{M}$ and we wish to estimate the V-optimal rule d_0, and its mean $E_{P_0}Y_{d_0} = \Psi(P_0)$, where $\Psi(P_0) = E_{P_0}\bar{Q}_0(0, 1, W) + E_{P_0}I(\bar{Q}_0(V) > 0)\bar{Q}_0(V)$.

Throughout the paper we will use counterfactual notation to denote parameters since it simplifies notation and helps the presentation, as if we assume a causal model and the randomization assumption, but, we always mean the corresponding statistical parameter of the data distribution whose definition only relies on the positivity assumption.

3 Data adaptive estimation of the V-optimal rule: Targeting the V-adjusted blip-function

We propose to utilize the loss-based super-learning approach to estimate $\bar{Q}_0(V)$, which implies a corresponding estimator of the V-optimal rule $d_0(V) = I(\bar{Q}_0(V) > 0)$. We will first present loss-functions, and subsequently, we present the loss-based super-learning method (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2012).

3.1 Loss functions

We propose the following loss-function directly inspired by (Rubin and van der Laan, 2007):

$$L_{Q_0,g_0}(\bar{Q})(O) = (D_1(Q_0,g_0)(O) - \bar{Q}(V))^2.$$

This loss function is indexed by nuisance parameters (Q_0, g_0) required to evaluate

$$D_1(Q_0,g_0) \equiv I(A(2)=1)\frac{2A(1)-1}{g_0(A \mid W)}(Y-\bar{Q}_0(A,W))+\bar{Q}_0(1,1,W)-\bar{Q}_0(0,1,W).$$

In fact, $D_1(Q_0,g_0)-E_{P_0}(Y(1,1)-Y(0,1))$ is the efficient influence curve for the parameter $E_{P_0}(Y(1,1)-Y(0,1))$, and it has the property that $E_{P_0}(D_1(Q,g) \mid V) = \bar{Q}_0(V)$, if either $Q = Q_0$ or $g = g_0$ (and that $0 < g(1,1 \mid W)$, $0 < g(0,1 \mid W)$). Due to this property it follows that if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then

$$\begin{aligned} &P_0 L_{Q,g}(\bar{Q}) = P_0 D_1(Q,g)^2 + P_0\bar{Q}^2(V) - 2P_0 D_1(Q,g)\bar{Q}(V) \\ &= P_0 D_1(Q,g)^2 + P_0\bar{Q}^2(V) - 2P_0\bar{Q}_0(V)\bar{Q}(V) \\ &= P_0(\bar{Q}-\bar{Q}_0)^2(V) + P_0 D_1(Q,g)^2 - P_0\bar{Q}_0^2(V). \end{aligned}$$

Thus this proves that, if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then the true risk of this loss function $L_{Q,g}(\bar{Q})$ equals $P_0(\bar{Q}-\bar{Q}_0)^2(V)$ up till a constant (not depending on the candidate $\bar{Q}$), so that $\bar{Q} \to P_0 L_{Q,g}(\bar{Q})$ is minimized over $\bar{Q}$ by the true $\bar{Q}_0$. Moreover, it can be shown that $P_n D_1(Q_0,g_0)\bar{Q}$ is an efficient estimator of $P_0\bar{Q}_0\bar{Q}$ demonstrating that the empirical mean of the loss efficiently estimates the true underlying risk, making $L_{Q_0,g_0}(\bar{Q})$ a double robust and efficient loss function for this true underlying squared error risk. That is, $P_n L_{Q_n,g_n}(\bar{Q})$ (and its cross-validated counterpart as used in cross-validation) is a double robust locally efficient estimator of the true underlying risk (up till the irrelevant constant), under regularity conditions.

A special choice of loss-function is obtained by setting $\bar{Q}_0(A,W) = 0$ so that $D_1(Q_0,g_0)$ simplifies to

$$D_1(g_0)(O) = I(A(2)=1)\frac{2A(1)-1}{g_0(A \mid W)}Y.$$

In this case, the loss-function $L_{g_0}(\bar{Q})(O) = (D_1(g_0)(O)-\bar{Q}(V))^2$ only depends on the single nuisance parameter g_0, which would be known in an RCT without missingness. However, even in such an RCT, we would recommend to use a loss-function $L_{Q_n,g_0}(\bar{Q})$ based on an estimator $\bar{Q}_n$ or $E_{P_0}(Y \mid A,W)$, so that

the empirical mean of the loss-function is a more efficient estimator of the underlying risk of $\bar{Q}$.

Suppose that it is known that $Y \in [0,1]$ and $\bar{Q}_0(V) \in (0,1)$. Then it is known that $\bar{Q}_0(V) \in (-1,1)$. More generally, suppose that it is known that $\bar{Q}_0(V) \in (a,b)$. Define $D_1^{a,b}(Q,g) = \frac{D_1(Q,g)-a}{b-a}$, and $\bar{Q}_0^{a,b} = (\bar{Q}_0 - a)/(b-a)$, and define the loss

$$-L_{1,Q,g}(\bar{Q}) = D_1^{a,b}(Q,g) \log \bar{Q}^{a,b} + (1 - D_1^{a,b}(Q,g)) \log(1 - \bar{Q}^{a,b}).$$

By the same argument, it follows that, if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then

$$P_0 L_{1,Q,g}(\bar{Q}) = -P_0 \left\{ \bar{Q}_0^{a,b} \log \bar{Q}^{a,b} + (1 - \bar{Q}_0^{a,b}) \log(1 - \bar{Q}^{a,b}) \right\}$$

showing that the true risk of this loss function is a Kullback-Leibler dissimilarity between $\bar{Q}^{a,b}$ and $\bar{Q}_0^{a,b}$. In addition, $P_n L_{1,Q_0,g_0}(\bar{Q})$ is an efficient estimator of this true underlying risk. Thus, this quasi-log-likelihood loss function satisfies that $\arg\min_{\bar{Q}} P_0 L_{1,Q,g}(\bar{Q}) = \bar{Q}_0$ if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$. Analogue to above, we can define $L_{1,g_0}(\bar{Q})$ as the loss-function that replaces $D_1^{a,b}(Q,g)$ in $L_{1,Q,g}$ by $D_1^{a,b}(g)$ defined above.

We state the validity of these loss functions as a formal result.

Theorem 1 *Assume $0 < g_0(1,1|W)$, $0 < g_0(0,1 \mid W)$, and that there exists a, b with $a < b < \infty$ so that $\bar{Q}_0(V) \in (a,b)$. Define the following two loss-functions for the parameter $\bar{Q}_0$:*

$$\begin{aligned} L_{Q_0,g_0}(\bar{Q})(O) &= (D_1(Q_0,g_0)(O) - \bar{Q}(V))^2 \\ L_{1,Q_0,g_0}(\bar{Q}) &= -\left\{ D_1^{a,b}(Q_0,g_0) \log \bar{Q}^{a,b}(V) + (1 - D_1^{a,b}(Q_0,g_0)) \log(1 - \bar{Q}^{a,b}(V)) \right\} \end{aligned}$$

We have

$$\bar{Q}_0 = \arg\min_{\bar{Q}} P_0 L_{Q,g}(\bar{Q}) \text{ if either } D_1(Q,g) = D_1(Q_0,g) \text{ or } D_1(Q,g) = D_1(Q,g_0),$$

and $0 < g(1,1|W)$, $0 < g(0,1 \mid W)$. The same result is true for $L_{1,Q,g}$.

3.2 Loss-based super-learning

For the sake of presentation, let's consider a randomized controlled trial without missingness so that we can use the loss-functions $L_{g_0}(\bar{Q})$ or $L_{1,g_0}(\bar{Q})$ for $\bar{Q}_0$. We first need to generate a library of candidate estimators for $\bar{Q}_0(V)$. The

loss function $L_{g_0}(\bar{Q})$ teaches us that we can apply any least-squares regression algorithm to regress $D_1(g_0)(O)$ on V, and $L_{1,g_0}(\bar{Q})$ teaches us that we can apply any logistic regression algorithm regressing $D_1^{a,b}(g_0)(O)$ on V. In this manner, we obtain a library of candidate estimators $\hat{\bar{Q}}_j$ of $\bar{Q}_0$, $j = 1, \ldots, J$, where the estimators are viewed as mappings from the empirical distribution P_n of $O_1, \ldots, O_n$ into the parameter space for $\bar{Q}_0$. These can include estimators assuming a parametric regression model, or highly data adaptive machine learning algorithms. This library of J estimators generates a family of candidate estimators $\hat{\bar{Q}}_\alpha = \sum_j \alpha_j \hat{\bar{Q}}_j$ indexed by a weight-vector α. We can now use loss-based cross-validation to select the optimal choice

$$\alpha_n = \arg\min_\alpha E_{B_n} P^1_{n,B_n} L_{g_0}(\hat{\bar{Q}}_\alpha(P^0_{n,B_n})),$$

where $B_n \in \{0,1\}^n$ denotes a random split of the sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, P^0_{n,B_n} and P^1_{n,B_n} denote the empirical distributions of the training and validation sample, respectively, and we used the notation $Pf = \int f(o)dP(0)$. The final estimator of $\bar{Q}_0(V)$ is now defined as

$$\bar{Q}_n = \hat{\bar{Q}}_{\alpha_n}(P_n),$$

which is called the super-learner. This implies a corresponding plug-in estimator $d_n(V) = (I(\bar{Q}_n(V) > 0), 1)$ of the V-optimal rule d_0.

Due to the oracle inequality for the cross-validation selector α_n (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006), if none of the candidate estimators $\hat{\bar{Q}}_\alpha$ converges at the parametric rate $1/\sqrt{n}$ to $\bar{Q}_0$, then we have that $\hat{\bar{Q}}_{\alpha_n}(P_n)$ is asymptotically equivalent (i.e. ratio of loss-based dissimilarities with $\bar{Q}_0$ converges to 1) with the oracle selected estimator $\hat{\bar{Q}}_{\tilde{\alpha}_n}(P_n)$ w.r.t. the loss-based dissimilarity $d(\bar{Q}, \bar{Q}_0) = E_{P_0}\{L_{g_0}(\bar{Q}) - L_{g_0}(\bar{Q}_0)\}$, where the oracle selector is defined as

$$\begin{aligned} \tilde{\alpha}_n &= \arg\min_\alpha E_{B_n} P_0 L_{g_0}(\hat{\bar{Q}}_\alpha(P^0_{n,B_n})) \\ &= \arg\min_\alpha E_{B_n} P_0 \left(\hat{\bar{Q}}_\alpha(P^0_{n,B_n})(V) - \bar{Q}_0(V)\right)^2. \end{aligned}$$

This result only relies on the loss-function $L_{g_0}(\bar{Q})$ to be uniformly bounded in O and $\bar{Q}$, which is arranged by assuming the strong version of the positivity assumption: there exists a $\delta > 0$ so that $\delta < g_0(1,1 \mid W)$, and $\delta < g_0(0,1 \mid W)$, with probability 1. If one of the candidate estimators converges at rate $1/\sqrt{n}$ (e.g., one of candidate estimators is based on a correctly specified parametric model), then the super-learner also converges at rate $1/\sqrt{n}$, but in this case,

it is not asymptotically equivalent with the oracle selector. These results still hold if $J = J(n)$ converges to infinity as fast as a polynomial power in n. The same method can be applied with the quasi-log-likelihood loss function L_{1,g_0}.

We could improve the cross-validated risk estimators, and thereby the cross-validation selector, by using the estimated loss $L_{Q_n,g_0}(\bar{Q})$ or L_{1,Q_n,g_0}, based on an estimator $D_1(Q_n, g_0)$ of $D_1(Q_0, g_0)$. In that case, the cross-validation selector is defined as:

$$\alpha_n = \arg\min_{\alpha} E_{B_n} P^1_{n,B_n} L_{Q_{n,B_n},g_0}(\hat{\bar{Q}}_\alpha(P^0_{n,B_n})),$$

where Q_{n,B_n} denotes the estimator of the nuisance parameters of the loss function based on the training sample P^0_{n,B_n}.

In an observational study, we would use the estimated loss $L_{Q_n,g_n}(\bar{Q})$ or L_{1,Q_n,g_n}. Finite sample oracle inequalities and asymptotic results for the resulting cross-validation selector based on such unified loss functions are presented in van der Laan and Dudoit (2003); van der Laan and Petersen (2012); Diaz and van der Laan (2013): in essence, one still obtains powerful oracle results for the cross-validation selector but the rate of convergence is upper-bounded by the product of the rates at which g_n converges to g_0 and Q_n converges to Q_0. Thus in observational studies in which one has strong knowledge about the treatment assignment mechanism or one knows that there is a single covariate (e.g., the outcome process at baseline) that blocks the effect of the history of the subject on the outcome so that it is sufficient to only adjust for this covariate when fitting the treatment mechanism, the cross-validation selector may still be asymptotically equivalent with the oracle selector above that treats g_0 as known, even if Q_n converges to a misspecified Q.

Further improvement can be obtained by estimating the true squared error risk $P_0 L_{Q_0,g_0}(\bar{Q})$ with a cross-validated TMLE, since a TMLE respects the global constraints of the model. In this case, oracle results have been obtained in (van der Laan and Petersen, 2012; Diaz and van der Laan, 2013). The CV-TMLE of risk is developed for the general multiple time point intervention case in the Appendix.

4 Data adaptive estimation of the V-optimal rule: Using performance of rule as criterion

We can generate a family of candidate estimators of the V-optimal rule by generating a family of candidate estimators of the V-adjusted blip-function with the estimation methodology of the previous section. In this manner,

we obtain candidate estimators $\hat{d}_j : \mathcal{M}_{NP} \to \mathcal{D}$, $j = 1, \ldots, J$ defined by $\hat{d}_j(P_n)(V) = I(\hat{\bar{Q}}_j(P_n)(V) > 0)$ based on an estimator $\hat{\bar{Q}}_j$ of the V-adjusted blip-function $\bar{Q}_0(V)$. We can use a parametric family to combine estimators. For example, one might define $\hat{\bar{Q}}_\alpha = \sum_j \alpha_j \hat{\bar{Q}}_j$ for a vector of weights α, and corresponding $\hat{d}_\alpha = I(\hat{\bar{Q}}_\alpha > 0)$. In this manner, we generated a whole family of candidate estimators $\{\hat{d}_\alpha : \alpha\}$ of the V-optimal rule d_0. It remains to propose a data adaptive selector of α. In the super-learner of the previous section we selected α based on the cross-validated estimate of a squared error risk of $\hat{\bar{Q}}_\alpha$ as an estimator of $\bar{Q}_0$. In this section, we consider an alternative criterion for selection of α: namely, a cross-validated estimate of the data adaptive parameter $E_{B_n} E_{P_0} Y_{\hat{d}_\alpha(P^0_{n,B_n})}$. This cross-validated estimate can be defined as a cross-validated empirical mean of an appropriate loss-function or one can use a cross-validated TMLE (Zheng and van der Laan (2010, 2012); van der Laan and Petersen (2012); Diaz and van der Laan (2013)).

4.1 Loss functions

Consider the double robust loss-function

$$-L_{\bar{Q},g}(d)(O) = \frac{I(A = d(V))}{g(A \mid W)}(Y - \bar{Q}(A, W)) + \bar{Q}(d(V), W),$$

indexed by nuisance parameter $(\bar{Q}(A, W) = E_P(Y \mid A, W), g)$. We note that

$$E_{P_0} L_{\bar{Q},g}(d)(O) = -E_{P_0} Y_d \text{ if either } g = g_0 \text{ or } \bar{Q} = \bar{Q}_0.$$

This proves that this loss function is a valid loss function for the optimal rule d_0:

$$d_0 = \arg\min_{d \in \mathcal{D}} E_{P_0} L_{\bar{Q},g}(d) \text{ if either } g = g_0 \text{ or } \bar{Q} = \bar{Q}_0.$$

The loss-based dissimilarity of this loss-function is given by:

$$E_{P_0} L_{\bar{Q},g}(d) - E_{P_0} L_{\bar{Q},g}(d_0) = E_{P_0} Y_{d_0} - E_{P_0} Y_d \geq 0,$$

if either $g = g_0$ or $\bar{Q} = \bar{Q}_0$. This loss-based dissimilarity provides a dissimilarity of a candidate rule d with the optimal rule d_0, and the cross-validation selector using this loss function is aiming to minimize this loss-based dissimilarity. If $d(V) = I(\bar{Q}(V) > 0)$, then this loss-based dissimilarity can be written as:

$$\begin{aligned} E_{P_0}(Y_d - Y_{d_0}) &= E_{P_0}\{I(\bar{Q}(V) > 0) - I(\bar{Q}_0(V) > 0)\}\bar{Q}_0(V) \\ &= E_{P_0} I(d(V) \neq d_0(V)) \mid \bar{Q}_0(V) \mid . \end{aligned}$$

That is, for each V with $d(V) \neq d_0(V)$ it adds a contribution being equal to the effect size at that V given by $\mid \bar{Q}_0(V) \mid$. This risk dissimilarity also shows that the optimal rule is only uniquely determined on the set $\{V : \bar{Q}_0(V) \neq 0\}$: i.e., if for a particular V treatment has no effect, it does not matter how the rule is defined for the purpose of maximizing EY_d.

In particular, we can consider the IPCW-loss function

$$-L_g(d)(O) = \frac{I(A = d(V))}{g(A \mid W)} Y.$$

For this loss-function we have

$$-E_{P_0} L_{g_0}(d)(O) = E_{P_0} Y_d.$$

The advantage of the double robust loss function relative to the IPCW-loss function is that its empirical mean is a double robust efficient estimator of its risk $E_{P_0} Y_d$.

4.2 Loss-based super-learning

For the sake of presentation, let's consider a randomized controlled trial without missingness so that we can use the loss-function $L_{g_0}(d)$. We can now use loss-based cross-validation to select the optimal choice

$$\alpha_n = \arg\min_{\alpha} E_{B_n} P^1_{n,B_n} L_{g_0}(\hat{d}_\alpha(P^0_{n,B_n})),$$

where $B_n \in \{0,1\}^n$ denotes a random split of the sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, P^0_{n,B_n} and P^1_{n,B_n} denote the empirical distributions of the training and validation sample, respectively, and we used the notation $Pf = \int f(o) dP(0)$. Recall $\hat{d}_\alpha(P_n)(V) = I(\hat{\bar{Q}}(P_n)(V) > 0)$. The final estimator of $\bar{Q}_0(V)$ is now defined as $\bar{Q}_n = \hat{\bar{Q}}_{\alpha_n}(P_n)$, which implies a corresponding plug-in estimator $d_n(V) = I(\bar{Q}_n(V) > 0)$ of the V-optimal rule d_0.

4.3 Oracle inequality for cross-validation selector

Since the loss-based dissimilarity $E_{P_0}(Y_{d_0} - Y_d)$ is a second order difference, which, for example, can be bounded by $\| \bar{Q} - \bar{Q}_0 \|_\infty P(\mid \bar{Q}_0(V) \mid < \| \bar{Q} - \bar{Q}_0 \|_\infty)$, we apply the oracle inequality for so called quadratic loss-functions (see (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006)), so that the remainder term will be $O(1/n)$ instead of $O(1/\sqrt{n})$.

However, this relies on a fundamental property of the loss-function, namely that the variance of the d_0-centered loss $L_{g_0}(d) - L_{g_0}(d_0)$ can be bounded by its expectation. The following lemma provides the resulting result.

Lemma 1 *Suppose* $a \equiv \inf_V \mid \bar{Q}_0(V) \mid > 0$ *where the infimum is taken over a support of* V*, and* $\min_A g_0(A \mid W) > b > 0$ *a.e. for some* $b > 0$*. Consider the discretized cross-validations selector in which we minimize over a set of* $K(n)$ *possible* α*-values, and we will still denote it with* α_n*. In that case, we have the following oracle inequality for the cross-validation selector* α_n *defined above: for each* $\delta > 0$*,*

$$E_0 E_{B_n} P_0\{L_{g_0}(\hat{d}_{\alpha_n}(P^0_{n,B_n})) - L_{g_0}(d_0)\} \leq$$
$$(1+\delta) E_0 \min_\alpha E_{B_n} P_0\{L_{g_0}(\hat{d}_{\alpha_n}(P^0_{n,B_n})) - L_{g_0}(d_0)\} + C(\delta)\frac{\log K(n)}{np},$$

where $C(\delta) < \infty$ *is a universal constant depending on* δ*,* a*,* b*, and* $M = \sup_v \mid \bar{Q}_0(v) \mid$*.*

Proof: Firstly, by assumption the loss-function is uniformly bounded in the sense that $\sup_{d \in \mathcal{D}, O} \mid L_{g_0}(d)(O) \mid < M_1 < \infty$ for some $M_1 < \infty$. In addition, we have that

$$\sup_{d \in \mathcal{D}} \frac{\mathrm{VAR}_{P_0}\{L_{g_0}(d) - L_{g_0}(d_0)\}}{E_{P_0}\{L_{g_0}(d) - L_{g_0}(d_0)\}} < M_2,$$

which is shown as follows:

$$E_0\{L_{g_0}(d) - L_{g_0}(d_0)\}^2 = E_0\{I(A = d(V)) - I(A = d_0(V))\}^2 \frac{Y^2}{g_0^2(A|W)}$$
$$E_0 \mid I(A = d(V)) - I(A = d_0(V)) \mid \frac{Y^2}{g_0^2(A|W)}$$
$$\leq C E_0 I(d(V) \neq d_0(V))$$
$$\leq C_1 E_0 I(d(V) \neq d_0(V)) \mid \bar{Q}_0(V) \mid$$
$$= C_1 E_0\{Y_{d_0} - Y_d\}.$$

The stated oracle inequality is now an application of the general oracle inequality for the cross-validation selector presented in (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006). □

This lemma relies on the strong assumption that $\bar{Q}_0(V)$ is bounded away from zero, since only under that assumption we can bound the variance of the d_0-centered loss by its expectation. Consider now the case that we do not want to assume $\bar{Q}_0(V)$ is bounded away from zero. We can obtain the following bound:

$$E_0\{L_{g_0}(d) - L_{g_0}(d_0)\}^2 \leq \sqrt{E_0 \frac{1}{\mid \bar{Q}_0(v) \mid}} \sqrt{E_0\{L_{g_0}(d) - L_{g_0}(d_0)\}}. \qquad (2)$$

Suppose we now only assume that $E_0 \frac{1}{|\bar{Q}_0(V)|} < \infty$. Consider the proof of the oracle inequality for quadratic loss-functions in van der Laan and Dudoit (2003) or any of the other references (see page 20 of technical report 126, www.bepress.com/ucbbiostat). It relies on dealing with bounding a remainder term $R_{k,n}$, whose tail-probability can be bounded with Bernsteins-inequality by:

$$P(R_{k,n} > s \mid P^0_{n,B_n}, B_n) \leq \exp\left(\frac{-np}{2(1+\delta)^2} \frac{(s+\delta\tilde{H}_k)^2}{\sigma_k^2 + M_1(s+\delta\tilde{H}_k)}\right),$$

where $\tilde{H}_k = P_0\{L_{g_0}(\hat{d}_{\alpha_k}(P^0_{n,B_n})) - L_{g_0}(d_0)\}$ and σ_k^2 is the variance of $\{L_{g_0}(\hat{d}_{\alpha_k}(P^0_{n,B_n})) - L_{g_0}(d_0)\}$. In the proof in the above referenced articles, we could bound σ_k^2 by $\tilde{H}_k$ and thereby establish that the tail-probability is $\exp(-Cns)$, and thereby that the remainder $\max_k R_{k,n}$ has an expectation that is $O((\log K(n))/n)$. In this case, (2) shows that we can only bound σ_k^2 by the square-root of $\tilde{H}_k$. This yields a tail-probability bound for $P(R_{k,n} > s \mid P_{n,B_n}, B_n)$ of the form:

$$\exp\left(-\frac{n(s+\tilde{H}_k)^2}{\sqrt{\tilde{H}_k} + (s+\tilde{H}_k)}\right).$$

From this we learn that if $\tilde{H}_k/n^{2/3} \to 0$, then $R_{k,n} = O_P(\tilde{H}_k^{0.25}/n^{1/2})$, and if $\tilde{H}_k/n^{2/3} \to \infty$, then $R_{k,n} = O_P(1/n)$. As a consequence, we can show that for single split cross-validation, we obtain an oracle inequality of the same form as above, but with the remainder term $O(\log K/n)$ in the case that $\tilde{H} = \min_k \tilde{H}_k$ converges to zero slower than $n^{-2/3}$, while the remainder is $O_P(\tilde{H}^{0.25}/n^{0.5})$ if $\tilde{H}$ converges to zero faster than $n^{-2/3}$. So in the latter case the remainder is larger than the leading term in the oracle inequality. This allows us to draw some conclusions regarding the behavior of the cross-validation selector when $E_0(1/ \mid \bar{Q}_0 \mid (V)) < \infty$. Based on the above, we claim that if the loss-based dissimilarity of the oracle selected estimator converges at a slower rate than $n^{-2/3}$, then we have an oracle inequality as above, and thereby the conclusion that the cross-validation selector is asymptotically equivalent with the oracle selector. On the other hand, if the loss-based dissimilarity of the oracle selected estimator converges at a faster rate than $n^{-2/3}$, then the loss-based dissimilarity of the cross-validation selected estimator might converge to zero at a slower rate than $n^{-2/3}$.

If these theoretical considerations have a practical analogue, then this suggests that in the case that $\bar{Q}_0(V)$ is not uniformly bounded away from zero, this cross-validation selector might be inferior to the cross-validation selector targeting $\bar{Q}_0$ itself as presented in the previous section, *if* we expect rates of

convergence for estimation of $\bar{Q}_0$ faster than $n^{-1/3}$ (e.g., dimension of V is low), while this cross-validation selector is asymptotically optimal otherwise (by being equivalent with the oracle selector). Future simulation studies will have to shed more light on the comparison of these two cross-validation selectors.

4.4 CV-TMLE of risk and corresponding super-learner

Above we used the cross-validated empirical mean $E_{B_n}P^1_{n,B_n}L_{g_0}(\hat{d}_\alpha(P^0_{n,B_n}))$ as an estimator of the data adaptive parameter $E_{B_n}E_{P_0}Y_{\hat{d}_\alpha(P^0_{n,B_n})}$, and thereby as criterion for selecting α. Instead, we can use the cross-validated TMLE of this data-adaptive target parameter, which will be presented in the third part of this paper. Since the cross-validated TMLE is a substitution estimator of this data adaptive parameter and thereby also respects global constraints in the statistical model, this can result in meaningful finite sample improvements relative to using the DR-IPCW or IPCW loss function.

5 The efficient influence curve of the mean outcome under the V-optimal rule: single time-point treatment

The following theorem shows that $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ with $\Psi(P_0) = E_{P_0}Y_{d_0}$ is pathwise differentiable with a specified canonical gradient, also called the efficient influence curve (Bickel et al., 1997; van der Vaart, 1998; van der Laan and Robins, 2003).

Theorem 2 *Assume* $P_0\left(0 < \min(g_0(1,1 \mid W), g_0(0,1 \mid W))\right) = 1$, $P_0(\mid Y \mid < M) = 1$ *for some* $M < \infty$. *The parameter* $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ *is pathwise differentiable with canonical gradient* $D^*(Q_0, g_0)$ *given by*

$$\begin{aligned} D^*(Q_0, g_0) &= D^*(d_0, Q_0, g_0) \\ &\equiv \frac{I(A = d_0(V))}{g_0(A \mid W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(d_0(V), W) - \Psi(Q_0). \end{aligned}$$

That is, $D^*(Q_0, g_0)$ *equals the efficient influence curve* $D^*(d, Q_0, g_0)$ *for the parameter* $\Psi_d(P) \equiv E_P Y_d = E_P\bar{Q}(P)(d(V), W)$ *treating* d *as given, at* $d = d_0$*:* $D^*(Q_0, g_0) = D^*(d_0, Q_0, g_0)$.

We will also denote this efficient influence curve $D^*(Q_0, g_0)$ with $D^*(d_0, \bar{Q}_0, Q_{W,0}, g_0)$ to stress its dependence on each of these components of P_0. The above theorem represents a surprising result at first sight. In general, if our statistical

target parameter is $EY_{d_{10}}$ for some rule d_{10} that depends on P_0, then the dependence of the statistical parameter on the unknown rule, assuming d_{10} is a smooth function of P_0, will generate another component to the efficient influence curve beyond the efficient influence curve that treats d_{10} as known. However, for our very special choice of optimal rule d_0, due to the representation (1), as shown in our proof below, the contribution of the dependence of the rule $d_0 = I(\bar{Q}_0(V) > 0)$ on P_0 to the derivative of our target parameter along paths through P_0 equals zero, so that the pathwise derivative is identical to what it would have been if one treats the rule d_0 as known.

Proof of Theorem 2: Consider the mapping $\Psi(Q) = E_{Q_W} I(\bar{Q}(V) > 0)\bar{Q}(V)$, ignoring the term $E_{Q_W}\bar{Q}(0, 1, W)$ in the definition of EY_{d_0}, since that one does not depend on the V-optimal rule d_0. The pathwise derivative of $\Psi(Q)$ is defined as $\frac{d}{d\epsilon}\Psi(Q(\epsilon))\big|_{\epsilon=0}$ along paths $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$. The derivative w.r.t. ϵ equals the sum of the three contributions $\frac{d}{d\epsilon}E_{Q_W(\epsilon)} I(\bar{Q}(V) > 0)\bar{Q}(V)$, $\frac{d}{d\epsilon}E_{Q_W} I(\bar{Q}(V) > 0)\bar{Q}_\epsilon(V)$ and $\frac{d}{d\epsilon}E_{Q_W} I(\bar{Q}_\epsilon(V) > 0)\bar{Q}(V)$, where all derivatives are at $\epsilon = 0$. The sum of the first two terms equals the pathwise derivative that treats the rule d_0 as known. The latter pathwise derivative has a canonical gradient given by the expression $D^*(d_0, Q_0, g_0) - D^*_{EY(0,1)}(P_0)$, where $D^*(d_0, Q_0, g_0)$ is presented in the theorem and $D^*_{EY(0,1)}(P_0)$ is the efficient influence curve of $E_{Q_W}\bar{Q}(0, 1, W)$. Thus, it remains to show that the third derivative equals zero. Consider a path $Q_{Y,\epsilon}(Y \mid A, W) = (1 + \epsilon S_Y(Y \mid A, W))Q_Y(Y \mid A, W)$ with $E(S_Y \mid A, W) = 0$, S is uniformly bounded, and $Q_{W,\epsilon} = (1 + \epsilon S_W(W))Q_W$ with $ES_W(W) = 0$ and S_W uniformly bounded. Then

$$\begin{aligned}&\bar{Q}_\epsilon(1, 1, W) = \textstyle\int_Y Y(1 + \epsilon S_Y(Y \mid A = (1, 1), W))dQ_Y(Y \mid A = (1, 1), W)\\ &= E(Y \mid A = (1, 1), W) + \epsilon E(Y S_Y(Y \mid A, W) \mid A = (1, 1), W),\end{aligned}$$

and similarly for $A = (0, 1)$. Under the assumption that there exists an $M < \infty$ so that $P(\mid Y \mid < M) = 1$, it follows that the supremum norm $\bar{Q}_\epsilon(W) - \bar{Q}(W)$ is bounded by $C(M)\epsilon$ for some $C(M) < \infty$, where $\bar{Q}(W) = \bar{Q}(1, 1, W) - \bar{Q}(0, 1, W)$. Similarly, we can show that $E_{Q_{W,\epsilon}}(\bar{Q}(W) \mid V) - E_{Q_W}(\bar{Q}(W) \mid V)$ is bounded by $C\epsilon$ for some $C = C(M) < \infty$, uniformly in choices $\bar{Q}$ that are uniformly bounded. As a consequence, it follows that the supremum norm of $\bar{Q}_\epsilon(V) - \bar{Q}(V)$ is bounded by a $C(M)\epsilon$ for some $C(M) < \infty$. We now proceed as follows:

$$\begin{aligned}&\mid EI(\bar{Q}_\epsilon(V) > 0)\bar{Q}(V) - EI(\bar{Q}(V) > 0)\bar{Q} \mid\\ &= \mid E\{I(\bar{Q}_\epsilon(V) > 0) - I(\bar{Q}(V) > 0)\}\bar{Q}(V) \mid\\ &\leq EI(\mid \bar{Q}(V) \mid < C \mid \epsilon \mid) \mid \bar{Q}(V) \mid,\end{aligned}$$

since, if $\bar{Q}(V) > C \mid \epsilon \mid > 0$, then $\bar{Q}_\epsilon(V) > 0$ and thus the difference between the two indicators equals zero, and similarly, if $\bar{Q}(V) < -C \mid \epsilon \mid < 0$, then

also $\bar{Q}_\epsilon(V) < 0$ and thus the difference between the two indicators equals zero again. Now, note that the last term is bounded as follows:

$$EI(\mid \bar{Q}(V) \mid < C \mid \epsilon \mid) \mid \bar{Q}(V) \mid = EI(\mid \bar{Q}(V) \mid < C \mid \epsilon \mid, \mid \bar{Q}(V) \mid > 0) \mid \bar{Q}(V) \mid$$
$$\leq C \mid \epsilon \mid EI(\mid \bar{Q}(V) \mid < C \mid \epsilon \mid, \mid \bar{Q}(V) \mid > 0).$$

Since for any random variable X we have $P(0 <\mid X \mid < \epsilon) \to 0$ as $\epsilon \to 0$, it follows that the last expression converges to zero as $\epsilon \to 0$. Thus, we have shown

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \{EI(\bar{Q}_\epsilon(V) > 0)\bar{Q}(V) - EI(\bar{Q}(V) > 0)\bar{Q}(V)\} = 0.$$

□

We have the following property of the efficient influence curve, which will provide a fundamental ingredient in the analysis of the TMLE presented in the next section.

Theorem 3 *For any Q with $Q_W = Q_{W,0}$ and any g, we have*

$$P_0 D^*(Q, g) = \Psi(Q_0) - \Psi(Q) + R(Q, Q_0, g, g_0), \tag{3}$$

where

$$\begin{aligned} R(Q, Q_0, g, g_0) &= E_{P_0}(\bar{Q} - \bar{Q}_0)(d_Q(V), W) \frac{(g_0 - g)(d_Q(V) \mid W)}{g(d_Q(V) \mid W)} \\ &\quad + E_{P_0} \{d_Q(V) - d_{Q_0}(V)\} \bar{Q}_0(V) \\ &\equiv R_1(Q, Q_0, g, g_0) + R_2(Q, Q_0), \end{aligned}$$

and $d_Q(V) = (I(\bar{Q}(V) > 0), 1)$, but $d_Q(V) - d_{Q_0}(V) \equiv I(\bar{Q}(V) > 0) - I(\bar{Q}_0(V) > 0)$.

If $g(d_Q(V) \mid W) > \delta > 0$ for some $\delta > 0$, then the first term $R_1()$ in $R()$ can be bounded as follows:

$$\begin{aligned} R_1 &\leq \frac{1}{\delta} \sqrt{E_{P_0} \left\{\bar{Q}(d_Q(V), W) - \bar{Q}_0(d_Q(V), W)\right\}^2} \\ &\quad \sqrt{E_{P_0} \{g_0(d_Q(V) \mid W) - g(d_Q(V) \mid W)\}^2}. \end{aligned}$$

The second term R_2 in $R()$ can be bounded as

$$\begin{aligned} R_2 &= E_{P_0} \{d_Q(V) - d_{Q_0}(V)\} \bar{Q}_0(V) \\ &\leq E_{P_0} I(\mid \bar{Q}_0(V) \mid < \mid \bar{Q} - \bar{Q}_0 \mid (V)) \bar{Q}_0(V) \\ &\leq E_{P_0} I(\mid \bar{Q}_0(V) \mid < \mid \bar{Q} - \bar{Q}_0 \mid (V)) \mid \bar{Q} - \bar{Q}_0 \mid (V) \\ &\leq \sqrt{E_{P_0}(\bar{Q} - \bar{Q}_0)^2(V)} \sqrt{E_{P_0} I \mid (\bar{Q}_0(V) \mid < \mid \bar{Q} - \bar{Q}_0 \mid (V))}, \end{aligned}$$

or, by bounding by the supremum norm instead of L^2-norm in the last-inequality, as

$$R_2(Q,Q_0) \;\leq\; \| \bar{Q} - \bar{Q}_0) \|_\infty E_{P_0} I(| \bar{Q}_0(V) |<| \bar{Q} - \bar{Q}_0 \mid (V)).$$

Note that this theorem proves that $R(Q,Q_0,g,g_0)$ is a second order term.
Proof of Theorem 3: Recall that $D^*(Q,g)$ equals the efficient influence curve for the fixed rule $d_Q(V) = (I(\bar{Q}(V) > 0), 1)$ at Q, g. For a fixed rule $d = d_Q(V)$, the expansion of $P_0 D^*(d,Q,g)$ for the efficient influence curve $D^*(d,Q,g)$ of EY_d is easily derived (see e.g., van der Laan (2012)), which yields: (recall $Q_W = Q_{W,0}$)

$$\begin{aligned}
&P_0 D^*(Q,g) = E_{P_0}\bar{Q}_0(0,1,W) + E_{P_0} I(\bar{Q}(V) > 0)\bar{Q}_0(V) \\
&-E_{P_0}\bar{Q}(0,1,W) - E_{P_0} I(\bar{Q}(V) > 0)\bar{Q}(V) \\
&+E_{P_0}\{\bar{Q}(d_Q(V),W) - \bar{Q}_0(d_Q(V),W)\}\tfrac{g_0(d_Q(V)|W) - g(d_Q(V)|W)}{g(d_Q(V)|W)} \\
&= E_{P_0}\bar{Q}_0(0,1,W) + E_{P_0} I(\bar{Q}_0(V) > 0)\bar{Q}_0(V) \\
&-E_{P_0}\bar{Q}(0,1,W) - E_{P_0} I(\bar{Q}(V) > 0)\bar{Q}(V) \\
&+E_{P_0}\{\bar{Q}(d_Q(V),W) - \bar{Q}_0(d_Q(V),W)\}\tfrac{g_0(d_Q(V)|W) - g(d_Q(V)|W)}{g(d_Q(V)|W)} \\
&+E_{P_0}\left\{I(\bar{Q})(V) > 0) - I(\bar{Q}_0(V) > 0)\right\}\bar{Q}_0(V).
\end{aligned}$$

The bounds are obtained as stated in the theorem, using Cauchy-Schwarz inequality. This completes the proof of the theorem. □

Note that, in a randomized controlled trial without missingness, in which case one applies this result at $g = g_0$, we have $R(Q,Q_0,g,g_0) = R_2(Q,Q_0)$.

6 Targeted minimum loss-based estimation of the mean outcome under V-optimal rule: single time-point treatment

Our proposed estimator is to first estimate the optimal rule d_0, giving us an estimated rule $d_n(V) = (I(\bar{Q}_n(V) > 0), 1)$, and subsequently apply the TMLE of EY_d for a fixed rule d at $d = d_n$. This TMLE of the additive causal effect of a single time point intervention has been previously developed: (Scharfstein et al., 1999; van der Laan and Rubin, 2006; van der Laan and Rose, 2012). Since the efficient influence curve satisfies a double robustness property, the TMLE and the DR-IPCW estimator (defined as a solution of the efficient influence curve based estimating equation) are double robust (Robins and Rotnitzky, 1992; van der Laan and Robins, 2003).

In a previous section we described a data adaptive estimator d_n of d_0. We now describe the TMLE for $\Psi_d(P_0) = E_{P_0}Y_d = E_{P_0}\bar{Q}_0(d(W), W)$ at a fixed rule d, and our proposed TMLE is this TMLE applied to $d = d_n$. This TMLE for a fixed dynamic treatment rule has been presented in the literature, but for the sake of being self-contained it will be shortly described here. Firstly, without loss of generality we can assume that $Y \in [0, 1]$. Let $\bar{Q}_n^0$ be an initial estimator of $\bar{Q}_0(A, W) = E_{P_0}(Y \mid A, W)$, and let g_n be an estimator of g_0. Since we only need to estimate $\bar{Q}_0^d(W) = \bar{Q}_0(d(V), W)$, this initial estimator $\bar{Q}_n^0$ could be based on the loss function

$$-L(\bar{Q}) = I(A = d(V)) \left\{ Y \log \bar{Q}(A, W) + (1 - Y) \log(1 - \bar{Q}(A, W)) \right\},$$

so that it only measures the performance of $\bar{Q}_n^0$ in estimating the function $\bar{Q}_0^d$. Note that this is indeed a valid loss function for $\bar{Q}_0^d = \bar{Q}_0(d(V), W)$ since $\bar{Q}_0^d = \arg\min_{\bar{Q}^d} E_{P_0} L(\bar{Q})$. In a randomized controlled trial, we can set $g_n = g_0$.

Consider the submodel

$$\text{Logit}\bar{Q}_n^0(\epsilon) = \text{Logit}\bar{Q}_n^0 + \epsilon H(g_n),$$

where $H(g_n)(A, W) = I(A_2 = 1, A_1 = d(V))/g_n(A \mid W)$. Let $\epsilon_n = \arg\min_\epsilon P_n L(\bar{Q}_n^0(\epsilon))$, which can be obtained with univariate logistic regression of Y on $H(g_n)$ using $\text{Logit}\bar{Q}_n^0$ as off-set, and only using the observations with $A_i = d(W_i)$. This defines now an update $\bar{Q}_n^* = \bar{Q}_n^0(\epsilon_n)$. The TMLE of $E_{P_0}Y_d$ is defined as the resulting plug-in estimator

$$\Psi_d(Q_{W,n}, \bar{Q}_n^*) = \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n^*(d(V_i), W_i).$$

Thus, our TMLE of $\Psi(Q_0) = \Psi_{d_0}(Q_{W,0}, \bar{Q}_0)$ is given by

$$\psi_n^* = \Psi_{d_n}(Q_{W,n}, \bar{Q}_n^*) = E_{Q_{W,n}} \bar{Q}_n^*(d_n(W), W)).$$

This TMLE $(d_n, Q_{W,n}, \bar{Q}_n^*)$ solves the efficient influence curve estimating equation:

$$P_n D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n) = 0.$$

7 Asymptotic efficiency and linearity of the TMLE of the mean counterfactual outcome under V-optimal rule: single time-point treatment

We now wish to analyze the TMLE $\psi_n^* = \Psi(d_n, Q_{W,n}, \bar{Q}_n^*)$ of $\psi_0 = \Psi(d_0, Q_{W,0}, \bar{Q}_0) = \Psi(Q_0)$. By Theorem 3, we have

$$-P_0 D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n) = \psi_0 - \Psi(d_n, Q_{W,n}, \bar{Q}_n^*) + R(Q_n, Q_0, g_n, g_0).$$

Combining this with $P_n D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n) = 0$ yields

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n) + R(Q_n, Q_0, g_n, g_0).$$

This provides a basis for proving the desired asymptotic efficiency of the TMLE. That is, if $D_n^* \equiv D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n)$ falls in a P_0-Donsker class with probability tending to 1 (van der Vaart and Wellner (1996)), $P_0\{D_n^* - D^*(Q_0, g_0)\}^2$ converges to zero in probability, and $R(Q_n, Q_0, g_n, g_0) = o_P(n^{-1/2})$, then it follows that

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(Q_0, g_0) + o_P(1/\sqrt{n}).$$

Thus, under these conditions, we have shown that the TMLE is asymptotically linear with influence curve the efficiency influence curve $D^*(Q_0, g_0) = D^*(d_0, Q_0, g_0)$, thereby establishing that the TMLE is asymptotically efficient.

In our theorem below we generalize this result by allowing that $\bar{Q}_n^*(A, W)$ is misspecified, even though the rule d_n and g_n are assumed to be consistent for d_0 and g_0.

Theorem 4 *Assume $Y \in [0, 1]$, $P_0(0 < \min(g_0(1, 1 \mid W), g_0(0, 1 \mid W))) = 1$, $D_n^* \equiv D^*(d_n, \bar{Q}_n^*, Q_{W,n}, g_n)$ falls in a P_0-Donsker class with probability tending to 1, $P_0\{D_n^* - D^*(d_0, \bar{Q}, Q_{W,0}, g_0)\}^2$ converges to zero in probability, and*

$$R_2(\bar{Q}_n, \bar{Q}_0) = E_{P_0}\left\{I(\bar{Q}_n(V) > 0) - I(\bar{Q}_0(V) > 0)\right\} \bar{Q}_0(V) = o_P(1/\sqrt{n}).$$

We refer to Theorem 3 for a second order representation of $R_2(\bar{Q}_n, \bar{Q}_0)$. Then,

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(d_0, \bar{Q}, Q_{W,0}, g_0) + R_1(Q_n, Q_0, g_n, g_0) + o_P(n^{-1/2}).$$

If $g_n = g_0$ (i.e., RCT), then $R_1(Q_n, Q_0, g_n, g_0) = 0$, so that ψ_n^* is asymptotically linear with influence curve $D^*(d_0, \bar{Q}, Q_{W,0}, g_0)$.

For general g_n, we also assume that

$$E_{P_0}\{\bar{Q}(d_0(V),W)-\bar{Q}_0(d_0(V),W)\}\frac{g_n(d_0(V)\mid W)-g_0(d_0(V)\mid W)}{g_0(d_0(V)\mid W)}$$
$$=(P_n-P_0)D_g(P_0)+o_P(1/\sqrt{n}),$$

for some function $D_g(P_0)(O)\in L_0^2(P_0)$, and (using notation $\bar{Q}^d(W)=\bar{Q}(d(V),W)$, $g^d(W)=g(d(V)\mid W)$, $\parallel f\parallel=\sqrt{P_0f^2}$)

$$\begin{array}{l}\parallel(\bar{Q}-\bar{Q}_0)^{d_n}-(\bar{Q}-\bar{Q}_0)^{d_0}\parallel\parallel g_n^{d_n}-g_0^{d_n}\parallel=o_P(1/\sqrt{n})\\ \parallel g_n^{d_n}-g_0^{d_n}\parallel^2=o_P(1/\sqrt{n})\\ \parallel(g_n-g_0)^{d_0}\parallel\parallel g_0^{d_n}-g_0^{d_0}\parallel=o_P(1/\sqrt{n})\\ \parallel(g_n-g_0)^{d_n}-(g_n-g_0)^{d_0}\parallel=o_P(1/\sqrt{n})\\ \parallel(\bar{Q}_n-\bar{Q})^{d_n}\parallel\parallel(g_n-g_0)^{d_n}\parallel=o_P(1/\sqrt{n}).\end{array}$$

Then,

$$\psi_n^*-\psi_0=(P_n-P_0)\{D^*(d_0,\bar{Q},Q_{W,0},g_0)+D_g(P_0)\}+o_P(1/\sqrt{n}),\qquad(4)$$

so that ψ_n^* is asymptotically linear with influence curve $D^*(d_0,\bar{Q},Q_{W,0},g_0)+D_g(P_0)$.

If g_n is an MLE of g_0 according to a correctly specified model $\mathcal{G}$ for g_0 with tangent space $T_g(P_0)$ at P_0, then it follows that

$$D_g(P_0)=-\Pi(D^*(d_0,\bar{Q},Q_{W,0},g_0)\mid T_g(P_0)),$$

where $\Pi(\cdot\mid T_g(P_0))$ denotes the projection operator onto $T_g(P_0)\subset L_0^2(P_0)$ in the Hilbert space $L_0^2(P_0)$.

Proof of Theorem 4: The first part of the Theorem has already been proven above. We have

$$\begin{array}{l}R_1(Q_n,Q_0,g_n,g_0)=E_{P_0}\{\bar{Q}_n(d_n(V),W)-\bar{Q}_0(d_n(V),W)\}\frac{g_n(d_n(V)|W)-g_0(d_n(V)|W)}{g_n(d_n(V)|W)}\\ =E_{P_0}\{\bar{Q}_n(d_n(V),W)-\bar{Q}(d_n(V),W)\}\frac{g_n(d_n(V)|W)-g_0(d_n(V)|W)}{g_n(d_n(V)|W)}\\ +E_{P_0}\{\bar{Q}(d_n(V),W)-\bar{Q}_0(d_n(V),W)\}\frac{g_n(d_n(V)|W)-g_0(d_n(V)|W)}{g_n(d_n(V)|W)}\\ =E_{P_0}\{\bar{Q}_n(d_n(V),W)-\bar{Q}(d_n(V),W)\}\frac{g_n(d_n(V)|W)-g_0(d_n(V)|W)}{g_n(d_n(V)|W)}\\ +E_{P_0}\{\bar{Q}(d_0(V),W)-\bar{Q}_0(d_0(V),W)\}\frac{g_n(d_0(V)|W)-g_0(d_0(V)|W)}{g_0(d_0(V)|W)}+R_{1b,n},\end{array}$$

where we will denote the first term on right-hand side with $R_{1a,n}$. Note that $R_{1b,n}$ can be decomposed in a sum of terms where, by using Cauchy-Schwarz

inequality, these terms can be bounded by

$$\begin{array}{l} \parallel (\bar{Q} - \bar{Q}_0)^{d_n} - (\bar{Q} - \bar{Q}_0)^{d_0} \parallel \parallel g_n^{d_n} - g_0^{d_n} \parallel \\ \parallel g_n^{d_n} - g_0^{d_n} \parallel^2 \\ \parallel (g_n - g_0)^{d_0} \parallel \parallel g_0^{d_n} - g_0^{d_0} \parallel \\ \parallel (g_n - g_0)^{d_n} - (g_n - g_0)^{d_0} \parallel . \end{array}$$

The first term $R_{1a,n}$ can be bounded by

$$\parallel (\bar{Q}_n - \bar{Q})^{d_n} \parallel \parallel (g_n - g_0)^{d_n} \parallel .$$

This completes the proof of (8). The last statement is a corollary of Theorem 2.3 in van der Laan and Robins (2003). □

7.1 Asymptotic linearity of TMLE in RCT:

Suppose the data is generated by a randomized controlled trial without missingness so that g_0 is known. In addition, assume that we have a univariate score V available, and we want to use the data of the RCT to learn the V-optimal rule d_0 and provide statistical inference for $E_{P_0}Y_{d_0}$. Since V is 1-dimensional, using kernel smoothers or sieve-based estimation to generate a library of candidate estimators for the super-learner based on loss function (e.g.) $L_{g_0}(\bar{Q})$ will generate an estimator $\bar{Q}_n$ of $\bar{Q}_0(V)$ that converges at a rate $n^{-2/5}$ under a minor smoothness assumption on $\bar{Q}_0$, and higher rates of convergence would be obtained under additional smoothness assumptions. As a consequence, in this case $R_2(Q_n, Q_0) = O_P(n^{-4/5})$ or better. As a consequence, all conditions of Theorem 4 hold, and it follows that the proposed TMLE is asymptotically linear with influence curve $D^*(d_0, \bar{Q}, Q_{W,0}, g_0)$, where $\bar{Q}(A, W)$ is the possibly misspecified limit of $\bar{Q}_n^*(A, W)$ in the TMLE. To conclude, randomized controlled trials allow us to learn V-optimal rules at adaptive optimal rates of convergence, and also allow valid asymptotic statistical inference for $E_{P_0}Y_{d_0}$ for univariate V, and, for multivariate V under additional smoothness assumptions on $\bar{Q}_0(V)$.

7.2 Statistical inference

Suppose one is concerned with statistical inference for the target parameter $\psi_{1,0} \equiv E_{P_0}Y_{d_0} - E_{P_0}Y_0$, where Y_0 is the counterfactual corresponding with static intervention $A = (0, 1)$. Above we developed the TMLE for $E_{P_0}Y_{d_0}$, and we could use a separate TMLE for EY_0, or we could use a TMLE of $\bar{Q}_0(A, W)$ targeting directly $E_{P_0}\{Y_{d_n} - Y_0\}$ by using the clever covariate $H(g_n) = I(A_2 =$

1)$\{I(A_1 = d(V)) - I(A_1 = 0)\}/g_n(A \mid W)$. This results in a TMLE $\psi^*_{1,n}$ of $\psi_{1,0}$. By (a trivial generalization of) Theorem 4, if $g_n = g_0$, then this TMLE of $E_{P_0}Y_{d_0} - E_{P_0}Y_0$ is asymptotically linear with influence curve

$$IC(P_0) = \{D^*(d_0, \bar{Q}, Q_{W,0}, g_0) - D_{EY_0}(P_0),$$

where $D_{EY_0}(P_0) = I(A_2 = 1)I(A_1 = 0)/g_0(A \mid W)(Y - \bar{Q}(A, W)) + \bar{Q}(0, 1, W) - E_{P_0}Y_0$. In addition, if g_n is an MLE of g_0 according to a model, then the above influence curve $IC(P_0)$ is a conservative influence curve. Let IC_n be an estimator of this influence curve $IC(P_0)$ obtained by plugging in the available estimates of its unknown components. The asymptotic variance of the TMLE $\psi^*_{1,n}$ of $\psi_{1,0} = E_{P_0}Y_{d_0} - E_{P_0}Y_0$ can now be estimated with

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} IC_n^2(O_i).$$

An asymptotic 0.95-confidence interval for $\psi_{1,0}$ is given by $\psi^*_{1,n} \pm 1.96\sigma_n/\sqrt{n}$. In particular, we can test a null-hypothesis $H_0 : \psi_{1,0} = 0$ to determine if there is statistically significant evidence that an optimal treatment rule outperforms the current standard treatment $A = 0$.

8 Formulation of optimal dynamic treatment estimation problem: two time point treatment

For the sake of presentation, we first consider the case of a two time-point treatment. In the Appendix we present the general K time-point case. Suppose we observe n i.i.d. copies $O_1, \ldots, O_n$ of $O = (L(0), A(0), L(1), A(1), Y = L(2)) \sim P_0$, where $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ is a binary treatment and $A_2(j)$ is a missing or right-censoring indicator at "time" j, $j = 0, 1$. For a time-dependent process $X()$, we will use the notation $\bar{X}(t) = (X(s) : s \leq t)$. Let $\mathcal{M}$ be a statistical model that makes no assumptions on the marginal distribution $Q_{0,L(0)}$ of $L(0)$, and the conditional distributions $Q_{0,L(j)}$ of $L(j)$, given $\bar{A}(j-1), \bar{L}(j-1)$, $j = 0, 1$, but might make assumptions on the conditional distributions $g_{0,A(j)}$ of $A(j)$, given $\bar{A}(j-1), \bar{L}(j)$, $j = 0, 1$. We will refer to g_0 as the intervention mechanism, which can be factorized in a treatment mechanism g_{01} and censoring mechanism g_{02} as follows:

$$g_0(O) = \prod_{j=1}^{2} g_{0,1}(A_1(j) \mid \bar{A}(j-1), \bar{L}(j)) g_{0,2}(A_2(j) \mid A_1(j), \bar{A}(j-1), \bar{L}(j)).$$

In particular, the data might have been generated by a sequential multiple assignment randomized trial (SMART) in which case g_{01} is known.

Let $(A(0), V(1))$ be a function of $(L(0), A(0), L(1))$, and let $V(0)$ be a function of $L(0)$. Let $V = \bar{V} = (V(0), V(1))$. Consider dynamic treatment rules $V(0) \rightarrow d_{A(0)}(V(0)) \in \{0,1\} \times \{1\}$ and $(A(0), V(1)) \rightarrow d_{A(1)}(A(0), V(1)) \in \{0,1\} \times \{1\}$ for assigning treatment $A(0)$ and $A(1)$, respectively, where the rule for $A(0)$ is only a function of $V(0)$, and the rule for $A(1)$ is only a function of $(A(0), V(1))$. Note that these rules are restricted to set the censoring indicators $A_2(j) = 1$, $j = 0, 1$. Let $\mathcal{D}$ be the set of all such rules. We assume that $V(0)$ is a function of $V(1)$ (i.e., observing $V(1)$ includes observing $V(0)$), but in the theorem below we indicate an alternative assumption. For any rule $d \in \mathcal{D}$, let

$$\Psi_d(P) \equiv E_{P_d} Y_d,$$

where Y_d is a random variable with probability density

$$\begin{aligned} &P_d(L(0), A(0), L(1), A(1), Y) \\ &= I(A = d(V)) Q_{L(0)}(L(0)) Q_{L(1)}(L(1) \mid L(0), A(0)) Q_Y(Y \mid \bar{L}(1), \bar{A}(1)), \end{aligned}$$

with respect to some dominating measure μ. This probability distribution P_d is the G-computation formula (Robins (1987b,b, 1997, 1999); Gill and Robins (2001); Yu and van der Laan (2003)) for the counterfactual O_d representing the probability distribution O would have had, if contrary to the fact, A would have been assigned according to the dynamic intervention $d = (d_{A(0)}, d_{A(1)})$. Thus,

$$E_{P_d} Y_d = \int_y y P_d(y) d\mu(y),$$

where

$$P_d(y) = \sum_{l(0), l(1)} P_d(l(0), d_{A(0)}(v(0)), l(1), d_{A(1)}(v(1)), y)$$

is the marginal density of Y_d under the joint distribution P_d. We are concerned with estimation of the V-optimal rule defined as

$$d_0 = \arg \max_{d \in \mathcal{D}} E_{P_{0,d}} Y_d.$$

We are also concerned with statistical inference for the statistical target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ defined by

$$\Psi(P_0) = E_{P_{0,d_0}} Y_{d_0} = \Psi_{d_0}(P_0).$$

This defines the statistical estimation problem addressed in the current (second) part of this article.

If we assume a structural equation model stating that

$$\begin{aligned} L(0) &= f_{L(0)}(U_{L(0)}) \\ A(0) &= f_{A(0)}(L(0), U_{A(0)}) \\ L(1) &= f_{L(1)}(L(0), A(0), U_{L(1)}) \\ A(1) &= f_{A(1)}(\bar{L}(1), A(0), U_{A(1)}) \\ Y &= f_Y(\bar{L}(1), \bar{A}(1), U_Y), \end{aligned}$$

we can define counterfactuals Y_d defined by the modified system in which the equations for $A(0), A(1)$ are replaced by $A(0) = d_{A(0)}(V(0))$ and $A(1) = d_{A(1)}(A(0), V(1))$. One can now define the causally optimal rule as $d_0^* = \arg\max_{d \in \mathcal{D}} E_{P_0} Y_d$. If we assume a sequential randomization assumption stating that $A(0)$ is independent of $U_{L(1)}, U_Y$, given $L(0)$, and $A(1)$ is independent of U_Y, given $\bar{L}(1), A(0)$, then we have that $E_0 Y_d = E_{P_{0,d}} Y_d$ for all rules d, and thereby that the statistical rule d_0 equals this causally optimal rule d_0^*, and thus that $E_0 Y_{d_0^*} = \Psi(P_0)$. Similarly, we have such an identifiability result/G-computation formula under the Neyman-Rubin causal model (Robins (1987a)).

In the remainder of the article, if for a static or dynamic intervention d, we use notation L_d (or Y_d, O_d) we mean the random variable with probability distribution P_d, so that all our quantities are statistical parameters. For example, the quantity $E_{P_0}(Y_{a(0)a(1)} \mid V_{a(0)}(1))$ defined in the next theorem denotes the conditional expectation of $Y_{a(0)a(1)}$, given $V_{a(0)}(1)$, under the probability distribution $P_{0,a(0)a(1)}$ (i.e., G-computation formula presented above for the static intervention $(a(0), a(1))$). In addition, if we write down these parameters, we will automatically assume the positivity assumption required for the G-computation formula to be well defined. For that it will suffice to assume

$$\begin{aligned} P_0\left(0 < \min_{\delta \in \{0,1\}} g_{0,A(0)}(\delta, 1 | L(0))\right) &= 1 \\ P_0\left(0 < \min_{\delta \in \{0,1\}} g_{0,A(1)}(\delta, 1 \mid \bar{L}(1), A(0))\right) &= 1. \end{aligned} \tag{5}$$

The next theorem presents an explicit form of the V-optimal individualized treatment rule d_0 as a function of P_0.

Theorem 5 *We assumed $V(0)$ is a function of $V(1)$. The V-optimal rule d_0*

can be represented as the following explicit parameter of P_0:

$$\begin{aligned}
\bar{Q}_{20}(a(0), v(1)) &= \\
&E_{P_0}(Y_{a(0),A(1)=(1,1)} \mid V_{a(0)}(1) = v(1)) - E_{P_0}(Y_{a(0),A(1)=(0,1)} \mid V_{a(0)}(1) = v(1)) \\
d_{0,A(1)}(A(0), V(1)) &= (I(\bar{Q}_{20}(A(0), V(1)) > 0), 1) \\
\bar{Q}_{10}(v(0)) &= E_{P_0}(Y_{A(0)=(1,1),d_{0,A(1)}} \mid V(0)) - E_{P_0}(Y_{A(0)=(0,1),d_{0,A(1)}} \mid V(0)) \\
d_{0,A(0)}(V(0)) &= (I(\bar{Q}_{10}(V(0)) > 0), 1),
\end{aligned}$$

where $a(0) \in \{0,1\} \times \{1\}$. If $V(1)$ does not include $V(0)$, but, for all $(a(0), a(1)) \in \{\{0,1\} \times \{1\}\}^2$,

$$E(Y_{a(0),a(1)} \mid V(0), V_{a(0)}(1)) = E(Y_{a(0),a(1)} \mid V_{a(0)}(1)), \tag{6}$$

then the above expression for the V-optimal rule d_0 is still true.

Proof: Let $V_d = (V(0), V_d(1))$. For a rule in $\mathcal{D}$, we have

$$\begin{aligned}
&E_{P_d} Y_d = E_{P_d} E_{P_d}(Y_d \mid V_d) \\
&= E_{V_d}\left(E(Y_{a(0),a(1)} \mid V_{a(0)}) I(a(1) = d_{A(1)}(a(0), V_{a(0)}(1)))\right) I(a(0) = d_{A(0)}(V(0))).
\end{aligned}$$

For each value of $a(0)$, $V_{a(0)} = (V(0), V_{a(0)}(1))$ and $d_{A(0)}(V(0))$, the inner conditional expectation is maximized over $d_{A(1)}(a(0), V_{a(0)}(1))$ by $d_{0,A(1)}$ as presented in the theorem, where we used that $V(1)$ includes $V(0)$. This proves that $d_{0,A(1)}$ is indeed the optimal rule for assignment of $A(1)$. Suppose now that $V(1)$ does not include $V(0)$, but the stated assumption holds. Then the optimal rule $d_{0,A(1)}$ that is restricted to be a function of $(V(0), V(1), A(0))$ is given by $I(\bar{Q}_{20}(A(0), V(0), V(1)) > 0)$, where

$$\begin{aligned}
&\bar{Q}_{20}(a(0), v(0), v(1)) = \\
&E_{P_0}(Y_{a(0),A(1)=(1,1)} - Y_{a(0),A(1)=(0,1)} \mid V_{a(0)}(1) = v(1), V(0) = v(0)).
\end{aligned}$$

However, by assumption, the latter function only depends on $(a(0), v(0), v(1))$ through $(a(0), v(1))$, and equals $\bar{Q}_{20}(a(0), v(1))$. Thus, we now still have that $d_{0,A(1)}(V) = (I(\bar{Q}_{20}(A(0), V(1)) > 0), 1)$, and, in fact, it is now also an optimal rule among the larger class of rules that are allowed to use $V(0)$ as well.

Given we found $d_{0,A(1)}$, it remains to determine the rule $d_{0,A(0)}$ that maximizes

$$\begin{aligned}
&E_{V_d}\left(E_P(Y_{a(0),d_{0,A(1)}} \mid V_{a(0)})\right) I(a(0) = d_{A(0)}(V(0))) \\
&= E_{V(0)} E(Y_{a(0),d_{0,A(1)}} \mid V(0)) I(a(0) = d_{A(0)}(V(0))),
\end{aligned}$$

where we used the iterative conditional expectation rule, taking the conditional expectation of $V_{a(0)}$, given $V(0)$. This last expression is maximized over $d_{A(0)}$ by $d_{0,A(0)}$ as presented in the theorem. This completes the proof. □

9 Data adaptive estimation of the V-optimal rule: two time-point treatment

We need to construct a data adaptive estimator of $\bar{Q}_{20}(a(0), v(1)) = E_{P_0}(Y_{a(0)11} - Y_{a(0)01} \mid V_{a(0)}(1) = v(1))$ and, given a resulting estimator $d_{n,A(1)}$ of $d_{0,A(1)}$, we subsequently need to construct a data adaptive estimator of $\bar{Q}_{10,d}(v(0)) = E_{P_0}(Y_{11d_{A(1)}} - Y_{01d_{A(1)}} \mid V(0) = v(0))$ for a given $d_{A(1)} = d_{n,A(1)}$. For that purpose we propose to use sequential loss-based super-learning defined by the application of two subsequent super-learners. Each super-learner relies on the specification of a library of candidate estimators of $\bar{Q}^d_{j0}$, a specification of loss functions $L_j(\bar{Q}^d_j)$ for $\bar{Q}^d_{j0}$, and cross-validation based on this loss function to select among weighted combinations of the candidate estimators, $j = 1, 2$: here $\bar{Q}^d_2 = \bar{Q}_2$ does not depend on d. Our loss functions will be indexed by nuisance parameters that, in general, need to be estimated, but the loss-function can be selected to be known in a sequential RCT in which g_0 is known. We first focus on the specification of valid loss-functions that can be used to both generate candidate estimators and to use the cross-validation in the loss-based super-learner. In the Appendix we develop sequential super-learning based on a cross-validated TMLE of the risk-function, while in loss-based super-learning the risk is estimated with a cross-validated empirical mean (which can be unstable, thereby motivating the CV-TMLE of risk).

9.1 Loss-functions

Define

$$L^F_{D_1(Q_0,g_0)}(\bar{Q}_2)(O) \equiv h(a(0), V(1)) \left\{ D_1(Q_0, g_0)(O) - \bar{Q}_2(A(0), V(1)) \right\}^2,$$

where

$$D_1(Q_0, g_{0,A(1)}) = I(A_2(1) = 1) \frac{2A_1(1) - 1}{g_{0,A(1)}(O)} (Y - E_{P_0}(Y \mid \bar{L}(1), \bar{A}(1)))$$
$$+ E_{P_0}(Y \mid \bar{L}(1), A(0), A(1) = (1, 1)) - E_{P_0}(Y \mid \bar{L}(1), A(0), A(1) = (0, 1)).$$

We have that $\bar{Q}_{20} = \arg\min_{\bar{Q}_2} P_{0,a(0)} L^F_{D_1(Q,g)}(\bar{Q}_2)$ if either $D_1(Q, g) = D_1(Q_0, g)$ or $D_1(Q, g) = D_1(Q, g_0)$, so that L^F is a valid loss function under sampling from the static-intervention specific G-computation distribution $P_{0,a(0)}$. Our proposed double robust loss function is obtained by applying the DR-IPCW

mapping (van der Laan and Dudoit, 2003) to this loss function:

$$
\begin{aligned}
&L_{2,D_1(Q_0,g_0),Q_0,g_0}(\bar{Q}_2)(O) \\
&= \sum_{a(0)} h(a(0),V(1)) \frac{I(A(0)=a(0))}{g_{0,A(0)}(O)} L^F_{D_1(Q_0,g_0)}(\bar{Q}_2)(O) \\
&- \sum_{a(0)} h(a(0),V(1)) \frac{I(A(0)=a(0)}{g_{0,A(0)}(O)} E_{Q_0}\left(L^F_{D_1(Q_0,g_0)}(\bar{Q}_2) \mid A(0),L(0)\right) \\
&+ \sum_{a(0)} h(a(0),V(1)) E_{Q_0}\left(L^F_{D_1(Q_0,g_0)}(\bar{Q}_2) \mid A(0)=a(0),L(0)\right),
\end{aligned}
$$

where $a(0)$ sums over the two values in $\in \{0,1\} \times \{1\}$. This loss function is indexed by nuisance parameters g_0, the stated conditional expectation under Q_0, given $A(0), L(0)$, and the nuisance parameters required to evaluate $D_1(Q_0, g_0)$. In addition, this loss function is indexed by a weight function $h()$, but each such choice defines a valid loss function. We have $\bar{Q}_{20} = \arg\min_{\bar{Q}_2} P_0 L_{2,D_1(Q,g),Q,g}(\bar{Q}_2)$ if one of the following four scenarios applies:

$$
\begin{aligned}
L_{2,D_1(Q,g),Q,g} &= L_{2,D_1(Q_0,g),Q_0,g} \\
L_{2,D_1(Q,g),Q,g} &= L_{2,D_1(Q,g_0),Q_0,g} \\
L_{2,D_1(Q,g),Q,g} &= L_{2,D_1(Q_0,g),Q,g_0} \\
L_{2,D_1(Q,g),Q,g} &= L_{2,D_1(Q,g_0),Q,g_0}.
\end{aligned}
$$

Under any of these 4 scenarios we have

$$
P_0\{L_{2,D_1(Q,g),Q,g}(\bar{Q}_2) - L_{2,D_1(Q,g),Q,g}(\bar{Q}_{20})\} = \sum_{a(0)} P_0 h(\bar{Q}_2 - \bar{Q}_{20})^2(a(0), V_{a(0)}(1)),
$$

demonstrating that $L_{2,D_1(Q_0,g_0),Q_0,g_0}(\bar{Q}_2)$ is indeed a valid double robust loss function for $\bar{Q}_{20}$. A special choice is obtained by setting the nuisance parameter $Q_0 = 0$ so that we obtain a simple IPCW-loss function:

$$
L_{2,g_0}(\bar{Q}_2)(O) = \sum_{a(0)} h(a(0),V(1)) \frac{I(A(0)=a(0)}{g_{0,A(0)}(O)} \left(D_1(g_0)(O) - \bar{Q}_2(A(0),V(1)\right)^2,
$$

where

$$
D_1(g_0)(O) = I(A_2(1)=1) \frac{2A_1(1)-1}{g_{0,A(1)}(O)} Y.
$$

In this case, the loss-function $L_{2,g_0}(\bar{Q}_2)(O)$ only depends on the single nuisance parameter g_0, which would be known in an RCT without missingness. However, even in an RCT, we would recommend to use a loss-function $L_{Q_n,g_0}(\bar{Q}_2)$

based on an estimator Q_n, so that the empirical mean of the loss-function is a more efficient estimator of the true risk.

For a given $d_{A(1)}$, define

$$L^F_{1,d,D_1(d,Q_0,g_0)}(\bar{Q}^d_1)(O) = h(V(0))(D_1(d,Q_0,g_0)(O) - \bar{Q}^d_1(V(0)))^2,$$

where

$$D_1(d,Q_0,g_0) = I(A_2(0)=1)\frac{2A_1(0)-1}{g_{0,A(0)}(O)}(Y - E_{P_0}(Y_d \mid L(0),A(0)))$$
$$+E_{P_0}(Y_d \mid L(0),A(0)=(1,1)) - E_{P_0}(Y_d \mid L(0),A(0)=(0,1)).$$

We have $\bar{Q}^d_{10} = \arg\min_{\bar{Q}^d_1} P_{0,d_{A(1)}} L^F_{1,D_1(d,Q,g)}(\bar{Q}^d_1)$ if either $D_1(d,Q,g) = D_1(d,Q_0,g)$ or $D_1(d,Q,g,) = D_1(d,Q,g_0)$, so that L^F_1 is a valid double robust loss function under sampling from the post-intervention distribution $P_{0,d_{A(1)}}$ corresponding with the dynamic intervention $d_{A(1)}$. Our proposed loss function is obtained by applying the DR-IPCW mapping to this loss function:

$$L_{1,d,D_1(d,Q_0,g_0),Q_0,g_0}(\bar{Q}^d_1)(O) = \frac{I(A(1)=d_{A(1)}(V(1)))}{g_{0,A(1)}(O)} L^F_{1,d,D_1(d,Q_0,g_0)}(\bar{Q}^d_1)$$
$$-\frac{I(A(1)=d_{A(1)}(V(1)))}{g_{0,A(1)}(O)} E_{Q_0}\left(L^F_{1,D_1(d,Q_0,g_0),Q_0,g_0}(\bar{Q}^d_1) \mid \bar{A}(1),\bar{L}(1)\right)$$
$$+E_{Q_0}\left(L^F_{1,d,D_1(d,Q_0,g_0),Q_0,g_0}(\bar{Q}^d_1) \mid A(0),A(1)=d_{A(1)}(V(1)),\bar{L}(1)\right)$$

This loss function satisfies the same double robustness as presented above. In particular, if we denote this loss function with $L_{1,d,Q_0,g_0}(\bar{Q}^d_1)$, then, if either $L_{1,d,Q,g} = L_{1,d,Q_0,g}$ or $L_{1,d,Q,g} = L_{1,d,Q,g_0}$, we have

$$P_0\{L_{1,d,Q,g}(\bar{Q}^d_1) - L_{1,d,Q,g}(\bar{Q}^d_{10})\} = P_0 h(V(0))(\bar{Q}^d_1 - \bar{Q}^d_{10})^2(V(0)),$$

demonstrating that $L_{1,d,Q_0,g_0}(\bar{Q}_1)$ is indeed a valid double robust loss function for $\bar{Q}^d_{10}$ whose loss-based dissimilarity equals a squared error dissimilarity. Again, a special choice is obtained by setting the nuisance parameter $Q_0 = 0$ so that we obtain an IPCW-loss function:

$$L_{1,d,g_0}(\bar{Q}^d_1)(O) = \frac{I(A(1)=d_{A(1)}(V(1)))}{g_{0,A(1)}(O)} h(V(0))(D_1(g_0)(O) - \bar{Q}^d_1(V(0)))^2,$$

where

$$D_1(g_0)(O) = I(A_2(0)=1)\frac{2A_1(0)-1}{g_{0,A(0)}(O)}Y.$$

We state the double robust property of these loss functions in the following theorem, even though the actual robustness is even better and stated above showing that one only needs to correctly specify one of the two nuisance parameters of $D_1()$ and one of the two nuisance parameters of the DR-IPCW mapping applied to L^F.

Theorem 6 *If either $Q = Q_0$ or $g = g_0$ (and the positivity assumption at g and g_0), then*

$$\begin{aligned} P_0\{L_{2,Q,g}(\bar{Q}_2) - L_{2,Q,g}(\bar{Q}_{20})\} &= P_0 \sum_{a(0)} h(\bar{Q}_2 - \bar{Q}_{20})^2(a(0), V_{a(0)}(1)) \\ P_0\{L_{1,d,Q,g}(\bar{Q}_1^d) - L_{1,d_{A(1)},Q,g}(\bar{Q}_{10}^d)\} &= P_0 h(V(0))(\bar{Q}_1^d - \bar{Q}_{10}^d)^2(V(0)), \end{aligned}$$

where $a(0) \in \{0,1\} \times \{1\}$. As a consequence, we have

$$\begin{aligned} \bar{Q}_{20} &= \arg\min_{\bar{Q}_2} P_0 L_{2,Q,g}(\bar{Q}_2) \\ \bar{Q}_{10}^d &= \arg\min_{\bar{Q}_1^d} P_0 L_{1,d,Q,g}(\bar{Q}_1^d) \\ &\quad \textit{if either } g = g_0 \textit{ or } Q = Q_0. \end{aligned}$$

Suppose that it is known that $Y \in [0,1]$ so that $\bar{Q}_{20} \in (-1,1)$ and $\bar{Q}_{10}^d \in (-1,1)$. More generally, suppose that it is known that $\bar{Q}_{20}, \bar{Q}_{10}^d \in (a,b)$ for some $a < b$. We define $D_1^{a,b}(Q,g) = \frac{D_1(Q,g)-a}{b-a} \in (0,1)$, and use the above loss-functions but with the squared error $(D_1 - \bar{Q}_{20})^2$ replaced by the quasi-log-likelihood loss

$$L^F_{2,D_1(Q,g)}(\bar{Q}_2) = -h\left\{D_1^{a,b}(Q,g)\log\bar{Q}_2 + (1 - D_1^{a,b}(Q,g))\log(1-\bar{Q}_2)\right\}.$$

Similarly, we define $D_1^{a,b}(d,Q,g) = \frac{D_1^{a,b}(d,Q,g)-a}{b-a}$ and replace $(D_1 - \bar{Q}_1^d)^2$ by

$$L^F_{1,d,D_1(d,Q,g)}(\bar{Q}_1^d) = -h\left\{D_1^{a,b}(d,Q,g)\log\bar{Q}_1^d + (1 - D_1^{a,b}(d,Q,g))\log(1-\bar{Q}_1^d)\right\}.$$

Let's denote the resulting loss functions with $L_{2,Q,g}$ and $L_{1,d,Q,g}$ again. By the same argument, these loss functions satisfy: if either $Q = Q_0$ or $g = g_0$, then

$$\begin{aligned} &P_0\{L_{2,Q,g}(\bar{Q}_2) - L_{2,Q,g}(\bar{Q}_{20})\} \\ &= -\textstyle\sum_{a(0)} P_0 h\{\bar{Q}_{20}\log\bar{Q}_2 + (1-\bar{Q}_{20})\log(1-\bar{Q}_2)\}(a(0), V_{a(0)}(1)) \\ &P_0\{L_{1,d,Q,g}(\bar{Q}_1^d) - L_{1,d,Q,g}(\bar{Q}_{10}^d)\} \\ &= -P_0 h\{\bar{Q}_{10}^d(V(0))\log\bar{Q}_1^d(V(0)) + (1-\bar{Q}_{10}^d(V(0)))\log(1-\bar{Q}_1^d)\}(V(0)), \end{aligned}$$

again, demonstrating that these are valid double robust loss functions whose loss-based dissimilarity now equals a Kullback-Leibner dissimilarity.

9.2 Loss-based sequential super-learning

For the sake of presentation, let's consider a sequentially randomized controlled trial without missingness. In that case, we can use the loss-functions $L_{2,g_0}(\bar{Q}_2)$ and $L_{1,d,g_0}(\bar{Q}_1^d)$ for $\bar{Q}_{20}$ and $\bar{Q}_{10}^d$, respectively.

We first need to construct a super-learner of $\bar{Q}_{20}$. This requires generating a library of candidate estimators of $\bar{Q}_{20}$. The IPCW-loss function $L_{2,g_0}(\bar{Q}_2)$ teaches us that we can apply any least-squares or logistic regression algorithm to regress $D_1(g_0)(O)$ on $A(0), V(1)$ using weights $h(A(0), V(1))/g_{0,A(0)}(O)$. In this manner, we obtain a library of candidate estimators $\hat{\bar{Q}}_{2,j}$ of $\bar{Q}_{20}$, $j = 1, \ldots, J$.

This generates a family of candidate estimators $\hat{\bar{Q}}_{2,\alpha} = \sum_j \alpha_j \hat{\bar{Q}}_{2j}$ obtained by taking linear combinations of these estimators using a weight-vector α. We can now use loss-based cross-validation to select the optimal choice

$$\alpha_n = \arg\min_{\alpha} E_{B_n} P^1_{n,B_n} L_{2,g_0}(\hat{\bar{Q}}_{2,\alpha}(P^0_{n,B_n})).$$

It can be decided to restrict α to be a vector of positive numbers and sum up till 1. The final super-learner of $\bar{Q}_{20}$ is now defined as $\bar{Q}_{2n} = \hat{\bar{Q}}_{2,\alpha_n}(P_n)$. This estimator $\bar{Q}_{2n}$ implies an estimator $d_{n,A(1)}(A(0), V(1)) = (I(\bar{Q}_{2n}(A(0), V(1)) > 0), 1)$ of $d_{0,A(1)}$.

Given this estimator $d_{A(1)} = d_{n,A(1)}$, we now need to construct a super-learner of $\bar{Q}_{10}^d$. The loss-function $L_{1,d,g_0}(\bar{Q}_1^d)$ teaches us that we can estimate $\bar{Q}_{10}^d$ by applying any regression algorithm to regress $D_1(g_0)(O)$ onto $V(0)$ using weights $I(A(1) = d_{A(1)}(A(0), V(1))/g_{0,A(1)}(O)$. In this manner, we obtain a library of candidate estimators $\hat{\bar{Q}}^d_{1,j}$ of $\bar{Q}^d_{10}$, $j = 1, \ldots, J$. This generates a family of candidate estimators $\hat{\bar{Q}}^d_{1,\alpha} = \sum_j \alpha_j \hat{\bar{Q}}^d_{1j}$. We can use loss-based cross-validation to select the optimal choice

$$\alpha_n = \arg\min_{\alpha} E_{B_n} P^1_{n,B_n} L_{1,d,g_0}(\hat{\bar{Q}}^d_{1,\alpha}(P^0_{n,B_n})).$$

The final super-learner of $\bar{Q}^d_{10}$ is defined as $\bar{Q}^d_{1n} = \hat{\bar{Q}}_{1,\alpha_n}(P_n)$. The above description of a particular super-learner $\bar{Q}^d_{1n}$ is applied to $d_{1,A(1)} = d_{n,A(1)}$. The resulting estimator $\bar{Q}_{1n} = \bar{Q}^{d_n}_{1n}$ implies an estimator $d_{n,A(0)}(V(0)) = (I(\bar{Q}_{1n}(V(0)) > 0), 1)$ of $d_{0,A(0)}$.

Thus, the above sequential loss-based super-learning approach based on the two loss-functions for $\bar{Q}_{20}$ and $\bar{Q}^d_{10}$ provides us with a data adaptive estimator d_n of the V-optimal rule d_0, fully utilizing the available machine learning literature.

The cross-validation selector for $\bar{Q}_{20}$ satisfies the previously discussed oracle inequality and corresponding asymptotic equivalence with the oracle selector under stated conditions (i.e., uniformly bounded loss function and the size of the library can grow polynomial in sample size). This shows that the super-learner is optimal in the sense that it asymptotically outperforms any candidate estimator by simply including it in the library. Of course, this relied on g_0 being known.

Regarding the cross-validation selector for $\bar{Q}_{10}$, we now have to note that $\bar{Q}_{20}$ (i.e., $d_{0,A(1)}$) is another nuisance parameter of the loss-function for $\bar{Q}_1^0$, and, as a consequence, the rate of convergence at which $d_{n,A(1)}$ converges to $d_{0,A(1)}$ will provide an upper-bound on the rate of convergence of the estimator $\bar{Q}_{1n}$ as an estimator of $\bar{Q}_{10}$.

As discussed previously, oracle results for the super-learner can still be obtained when g_0 is estimated, when we use the DR-IPCW loss function using estimators Q_n, g_n, or if we estimate the desired full-data risk with CV-TMLE as carried out in the Appendix. The advantage of using double robust loss functions is that the second order terms in the finite sample oracle inequality are now expressed in terms of product of the approximation errors of the two nuisance parameters, and the further advantage of the CV-TMLE is that it is a substitution estimator respecting global bounds thereby enhancing the finite sample robustness of the risk-estimator.

The performance of the estimators $\bar{Q}_{1n}^d$ and $\bar{Q}_{2n}$ of $\bar{Q}_{10}^d$ and $\bar{Q}_{20}$, respectively, can be assessed with cross-validation, analogue to the use of cross-validation to assess the performance of a super-learner in the regression context.

9.3 Cross-validation based on performance of rule

As in our point-treatment section, given a collection of candidate estimators $\hat{d}_\alpha(P_n)$ of d_0, we can also select α with a minimizer of a cross-validated estimator of $\alpha \rightarrow E_{B_n} E_{P_0} Y_{\hat{d}_\alpha(P^0_{n,B_n})}$. For example, we could use the cross-validated empirical mean $E_{B_n} P^1_{n,B_n} L_{g_0}(\hat{d}_\alpha(P^0_{n,B_n})$ of the IPCW-loss $L_{g_0}(d) = I(\bar{A} = d(V))/g_0(O)Y$ or the DR-IPCW L_{g_0,Q_0} loss defined as the efficient influence curve of EY_d (minus the EY_d-constant so that it has expectation equal to EY_d). Of course, when the nuisance parameters of the loss are unknown, then they are replaced by estimators based on the training samples: e.g., $E_{B_n} P^1_{n,B_n} L_{\hat{g}(P^0_{n,B_n})}(\hat{d}_\alpha(P^0_{n,B_n})$. Alternatively, we can estimate this data adaptive target parameter $E_{B_n} E_{P_0} Y_{\hat{d}_\alpha(P^0_{n,B_n})}$ with the CV-TMLE, as we present in part III of this article (analogue to Zheng and van der Laan (2010, 2012);

van der Laan and Petersen (2012); Diaz and van der Laan (2013)).

10 The efficient influence curve of the mean outcome under V-optimal rule: two time-point treatment

In the next theorem we present a representation of $\Psi(P_0) = E_0 Y_{d_0}$ that explicitly shows how $\Psi(P_0)$ depends on d_0, which will allow us to establish the pathwise differentiability with known efficient influence curve.

Theorem 7 *Recall the definitions of $\bar{Q}_{20}$ and $\bar{Q}_{10}$ in Theorem 5. We can represent $\Psi(P_0) = E_{P_{d_0}} Y_{d_0}$ as follows:*

$$\Psi(P_0) = EY_{0101} + E_{V_{a(0)=(0,1)}} d_{0,A(1)}(a(0) = (0,1), V_{a(0)=(0,1)}) \bar{Q}_{20}(0, 1, V_{a(0)=(0,1)}) \\ + E_{V(0)} d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)).$$

Proof: We have

$$\begin{aligned}
&\Psi(P_0) = E_{V(0)} E(Y_{01,d_{0,A(1)}} \mid V(0)) + d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)) \\
&= E_{V_{a(0)=(0,1)}} E(Y_{01,d_{0,A(1)}} \mid V_{a(0)=(0,1)}) + E_{V(0)} d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)) \\
&= E_{V_{a(0)=(0,1)}} E(Y_{0101} \mid V_{a(0)=(0,1)}) + I(\bar{Q}_{20}(a(0) = (0,1), V_{a(0)=(0,1)}) > 0) \bar{Q}_{20}(0, V_{a(0)=(0,1)}) \\
&+ E_{V(0)} d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)) \\
&= E_{V_{a(0)=(0,1)}} E(Y_{0101} \mid V_{a(0)=(0,1)}) + d_{0,A(1)}(a(0) = (0,1), V_{a(0)=(0,1)}) \bar{Q}_{20}(0, V_{a(0)=(0,1)}) \\
&+ E_{V(0)} d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)) \\
&= EY_{0101} + E_{V_{a(0)=(0,1)}} d_{0,A(1)}(a(0) = (0,1), V_{a(0)=(0,1)}) \bar{Q}_{20}(0, V_{a(0)=(0,1)}) \\
&+ E_{V(0)} d_{0,A(0)}(V(0)) \bar{Q}_{10}(V(0)).
\end{aligned}$$

This completes the proof of the theorem. □

The following theorem presents the efficient influence curve of Ψ.

Theorem 8 *Assume that $P_0(\mid Y \mid < M) = 1$ for some $M < \infty$. The parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is pathwise differentiable with canonical gradient given by*

$$D^*(P_0) = \sum_{k=0}^{2} D_k^*(P_0),$$

where

$$\begin{aligned}
D_0^*(P_0) &= E_{P_0}(Y_{d_0} \mid L(0), A(0) = d_{0,A(0)}(V(0))) - E_{P_0}Y_{d_0} \\
D_1^*(P_0) &= \frac{I(A(0) = d_{0,A(0)}(V(0)))}{g_{0,A(0)}(O)} \\
&\quad \times \left(E_{P_0}(Y_{d_0} \mid \bar{A}(1) = d_0(V), \bar{L}(1)) - E_{P_0}(Y_{d_0} \mid L(0), A(0) = d_{0,A(0)}(V(0)))\right) \\
D_2^*(P_0) &= \frac{I(\bar{A}(1) = d_0(V))}{\prod_{j=0}^1 g_{0,A(j)}(O)}(Y - E_{P_0}(Y_{d_0} \mid \bar{A}(1) = d_0(V), \bar{L}(1)))
\end{aligned}$$

That is, $D^*(P_0)$ equals the efficient influence curve $D_0^*(d, P_0)$ for the parameter $\Psi_d(P) \equiv E_P Y_d$ treating d as given, at the V-optimal rule $d = d_0$: $D^*(P_0) = D_0^*(d_0, P_0)$.

Proof: The expression for $\Psi(P_0)$ presented in the previous theorem, and using the same proof as applied for the point treatment case demonstrates that the dependence of $\Psi(P_0)$ on the rule d_0 is such that the pathwise derivative of Ψ w.r.t. d_0 equals zero. As a consequence, the pathwise derivative is identical to the pathwise derivative of $\Psi_d : \mathcal{M} \to \mathbb{R}$ with $\Psi(P) = E_P Y_d$ for a fixed d at $d = d_0$. The latter pathwise derivative has a known efficient influence curve (Bang and Robins, 2005; van der Laan and Gruber, 2012) and is given by the expression stated in the theorem. □

We have the following property of the efficient influence curve, which will provide a fundamental ingredient in the analysis of the TMLE presented in the next section.

Theorem 9 *Let d_Q be the V-optimal rule corresponding with Q. For any Q, g, we have*

$$P_0 D^*(Q, g) = \psi_0 - \Psi(Q) + R_{1d_Q}(Q, Q_0, g, g_0) + R_2(Q, Q_0)$$

where

$$R_{1d}(Q, Q_0, g, g_0) = P_0 D^*(d, Q, g) - (\Psi_d(Q_0) - \Psi_d(Q)),$$

$\Psi_d(P) = E_P Y_d$ is the statistical target parameter that treats d as known, and $D^*(d, Q_0, g_0)$ is the efficient influence curve of this parameter Ψ_d at P_0. In addition,

$$\begin{aligned}
&R_2(Q, Q_0) = \Psi_{d_Q}(Q_0) - \Psi_{d_0}(Q_0) \\
&= E_{Q_0}(d_{Q,A(1)} - d_{0,A(1)})(a(0) = (0,1), V_{a(0)=(0,1)})\bar{Q}_{20}(0, 1, V_{a(0)=(0,1)}) \\
&\qquad + E_{Q_0}(d_{Q,A(0)} - d_{0,A(0)})(V(0))\bar{Q}_{10}(V(0)) \\
&\equiv R_{2A(1)}(Q, Q_0) + R_{2A(0)}(Q, Q_0).
\end{aligned}$$

The term $R_{2A(1)}$ can be bounded as

$$\begin{aligned} R_{2A(1)} &= E_{Q_0}\left\{I(\bar{Q}_2 > 0) - I(\bar{Q}_{20} > 0)\right\}\bar{Q}_{20}(0,1,V_{a_0=(0,1)}) \\ &\leq E_{Q_0} I(\mid \bar{Q}_{20}(0,1,V_{01}) \mid< \mid \bar{Q}_2 - \bar{Q}_{20} \mid (0,1,V_{01}))\bar{Q}_{20}(0,1,V_{01}) \\ &\leq \sqrt{E_{Q_0}(\bar{Q}_2 - \bar{Q}_{20})^2(0,1,V_{01})}\sqrt{E_{P_0} I \mid (\bar{Q}_{20} \mid< \mid \bar{Q}_2 - \bar{Q}_{20} \mid (0,1,V_{01}))}, \end{aligned}$$

or, by bounding by the supremum norm instead of L^2-norm in the last-inequality, as

$$R_{2A(1)} \leq \parallel (\bar{Q}_2 - \bar{Q}_{20})(0,1,\cdot) \parallel_\infty E_{P_0} I \mid (\bar{Q}_{20}(0,1,V_{01}) \mid< \mid \bar{Q}_2 - \bar{Q}_{20} \mid (0,1,V_{01})).$$

Similarly, the term $R_{2A(0)}$ can be bounded as

$$\begin{aligned} R_{2A(0)} &= E_{Q_0}\left\{I(\bar{Q}_1(V(0)) > 0) - I(\bar{Q}_{10}(V(0)) > 0)\right\}\bar{Q}_{10}(V(0)) \\ &\leq E_{Q_0} I(\mid \bar{Q}_{10}(V(0)) \mid< \mid \bar{Q}_1 - \bar{Q}_{10} \mid (V(0)))\bar{Q}_{10}(V(0)) \\ &\leq \sqrt{E_{Q_0}(\bar{Q}_1 - \bar{Q}_{10})^2(V(0))}\sqrt{E_{P_0} I \mid (\bar{Q}_{10}(V(0)) \mid< \mid \bar{Q}_1 - \bar{Q}_{10} \mid (V(0)))}, \end{aligned}$$

or, by bounding by the supremum norm instead of L^2-norm in the last-inequality, as

$$R_{2A(0)} \leq \parallel \bar{Q}_1 - \bar{Q}_{10}) \parallel_\infty E_{P_0} I \mid (\bar{Q}_{10}(V(0)) \mid< \mid \bar{Q}_1 - \bar{Q}_{10} \mid (V(0))).$$

From the study of the statistical target parameter Ψ_d, we know that $P_0 D^*(d,Q,g) = \Psi_d(Q_0) - \Psi_d(Q) + R_{1d}(Q,Q_0,g,g_0)$, where R_1 is a closed form second order term involving integrals of differences $Q - Q_0$ times differences $g - g_0$ (van der Laan and Gruber (2012)), and the remainder $R_1()$ in the Theorem is just $R_{1,d_Q}(Q,Q_0,g,g_0)$.

Proof: By definition of $R_{1d}(Q,Q_0,g,g_0)$ we have

$$\begin{aligned} &P_0 D^*(Q,g) = P_0 D^*(d_Q,Q,g) = \Psi_{d_Q}(Q_0) - \Psi_{d_Q}(Q) + R_{1d_Q}(Q,Q_0,g,g_0) \\ &= \Psi_{d_0}(Q_0) - \Psi_{d_Q}(Q) + \{\Psi_{d_Q}(Q_0) - \Psi_{d_0}(Q_0)\} + R_{1d_Q}(Q,Q_0,g,g_0) \\ &= \psi_0 - \Psi(Q) + R_2(Q,Q_0) + R_{1d_Q}(Q,Q_0,g,g_0). \end{aligned}$$

The bounding of $R_2(Q,Q_0)$ proceeds as stated in the theorem. □

11 Targeted minimum loss-based estimation of the mean outcome under V-optimal rule: two time-point treatment

Our proposed TMLE is to first estimate the optimal rule d_0, giving us an estimated rule $d_n(V) = (d_{n,A(0)}(V(0)), d_{n,A(1)}(V(1))$, and subsequently apply

the TMLE of EY_d for a fixed rule d at $d = d_n$ as presented in van der Laan and Gruber (2012).This TMLE is an analogue of the double robust estimating equation method presented in Bang and Robins (2005): see also Petersen et al. (2013) for a generalization of the TMLE to marginal structural models for dynamic treatments.

In a previous section we described a data adaptive estimator d_n of d_0. So it remains to describe the TMLE for $\Psi_d(P_0) = E_{P_0}Y_d$ at a fixed rule d, and our proposed TMLE is this TMLE applied to $d = d_n$.

This TMLE for a fixed dynamic treatment rule has been presented in the literature, but for the sake of being self-contained it will be shortly described here. Firstly, without loss of generality we can assume that $Y \in [0, 1]$. Let $\bar{Q}^d_{2n}$ be an initial estimator of $E_{P_0}(Y \mid \bar{A}(1) = d(\bar{L}(1)), \bar{L}(1))$. Consider the submodel $\text{Logit}\bar{Q}^d_{2n}(\epsilon) = \text{Logit}\bar{Q}^d_{2n} + \epsilon H_2(g_n)$, where

$$H_2(g_n) = \frac{I(\bar{A}(1) = d(\bar{L}(1))}{\prod_{j=0}^{1} g_{n,A(j)}(O)}.$$

Let ϵ_n be the estimator of ϵ obtained by fitting ϵ with univariate logistic regression of Y on $H_2(g_n)$ using $\text{Logit}\bar{Q}_{2n}$ as off-set. This defines a targeted estimator $\bar{Q}^{d*}_{2n} = \bar{Q}^d_{2n}(\epsilon_n)$.

Regress $\bar{Q}^{d*}_{2n}$ on $L(0), A(0) = d_{A(0)}(L(0))$ which defines an initial estimator $\bar{Q}^d_{1n}$ of $\bar{Q}^d_{10} = E_{P_0}(Y_d \mid L(0)) = E_{P_0}(\bar{Q}_{20} \mid L(0), A(0) = d_{A(0)}(L(0)))$. Consider the submodel $\text{Logit}\bar{Q}^d_{1n}(\epsilon) = \text{Logit}\bar{Q}^d_{1n} + \epsilon H_1(g_n)$, where

$$H_1(g_n) = \frac{I(A(0) = d_{A(0)}(L(0)))}{g_{n,A(0)}(O)}.$$

Let ϵ_n be the estimator of ϵ obtained by fitting ϵ with univariate logistic regression of $\bar{Q}^d_{1n}$ on $H_1(g_n)$ using $\text{Logit}\bar{Q}^d_{1n}$ as off-set. This defines a targeted estimator $\bar{Q}^{d*}_{1n} = \bar{Q}^d_{1n}(\epsilon_n)$ of $\bar{Q}^d_{10}$. Let $Q_{L(0),n}$ be the empirical distribution of $L_i(0)$, and let $Q^{d*}_n = (Q_{L(0),n}, \bar{Q}^{d*}_{1n}, \bar{Q}^{d*}_{2n})$. The TMLE of $\psi_0 = E_{P_0}Y_d = \Psi_d(Q_{L(0),0}, \bar{Q}^d_{10}, \bar{Q}^d_{20}) = E_{Q_{L(0),0}}\bar{Q}^d_{10}(L(0))$ is defined by the plug-in estimator

$$\psi^*_n = \Psi(Q^*_n) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}^{d*}_{1n}(L_i(0)).$$

Thus, our TMLE of $\Psi(Q_0) = \Psi_{d_0}(Q_{W,0}, \bar{Q}_0)$ is given by

$$\psi^*_n = \Psi_{d_n}(Q_{L(0),n}, \bar{Q}^{d_n*}_{1n}, \bar{Q}^{d_n*}_{2n}) = E_{Q_{L(0),n}}\bar{Q}^{d_n*}_{1n}(L(0)).$$

Recall that $D^*(d, Q^d, g)$ is the efficient influence curve for the target parameter EY_d treating d as fixed, and that we showed that $D^*(d_0, Q^{d_0}_0, g_0)$ is

the efficient influence curve of the target parameter EY_{d_0} where d_0 is the V-optimal rule. The TMLE $(d_n, Q_n^* = Q_n^{d_n*})$ described above solves the efficient influence curve estimating equation:

$$P_n D^*(d_n, Q_n^{d_n*}, g_n) = 0.$$

12 Asymptotic efficiency of the TMLE of the mean outcome under V-optimal rule: two time-point treatment

We now wish to analyze the TMLE $\psi_n^* = \Psi(d_n, Q_n^{d_n*})$ of $\psi_0 = \Psi(d_0, Q_0^{d_0}) = \Psi(Q_0)$. By Theorem 9, we have

$$-P_0 D^*(d_n, Q_n^{d_n*}, g_n) = \psi_0 - \Psi(d_n, Q_n^{d_n*}) + R(Q_n, Q_0, g_n, g_0).$$

Combining this with $P_n D^*(d_n, Q_n^{d_n*}, g_n) = 0$ yields

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(d_n, Q_n^{d_n*}, g_n) + R(Q_n, Q_0, g_n, g_0).$$

This provides a basis for proving the desired asymptotic efficiency of the TMLE. That is, if $D_n^* \equiv D^*(d_n, Q_n^*, g_n)$ falls in a P_0-Donsker class with probability tending to 1, $P_0\{D_n^* - D^*(d_0, Q_0, g_0)\}^2$ converges to zero in probability, and $R(Q_n, Q_0, g_n, g_0) = o_P(n^{-1/2})$, then it follows that

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(d_0, Q_0, g_0) + o_P(1/\sqrt{n}).$$

Thus, under these conditions, we have shown that the TMLE is asymptotically linear with influence curve the efficiency influence curve, thereby establishing that the TMLE is asymptotically efficient.

In our theorem below we generalize this result by allowing that $\bar{Q}_n^*$ is misspecified, even though the rule d_n and g_n are assumed to be consistent for d_0 and g_0.

Theorem 10 *Assume* $Y \in [0,1]$, $g_0(a(0), a(1), \bar{L}(1)) > 0$ *for all* $(a(0), a(1)) \in \{\{0,1\} \times \{1\}\}^2$. $D_n^* \equiv D^*(d_n, Q_n^*, g_n)$ *falls in a* P_0*-Donsker class with probability tending to 1,* $P_0\{D_n^* - D^*(d_0, Q, , g_0)\}^2$ *converges to zero in probability,* $Q_{L(0)} = Q_{L(0),0}$, *and*

$$R_2(\bar{Q}_n, \bar{Q}_0) = o_P(1/\sqrt{n}),$$

where $R_2()$ *is defined and bounded in Theorem 9. Then,*

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(d_0, \bar{Q}, Q_{W,0}, g_0) + R_{1d_n}(Q_n, Q_0, g_n, g_0) + o_P(n^{-1/2}),$$

where $R_{1d} = P_0 D^*(d, Q_n, g_n) - (\Psi_d(Q_0) - \Psi_d(Q_n))$*, as defined in Theorem 9. If* $g_n = g_0$ *(i.e., RCT), then* $R_{1d_n}(Q_n, Q_0, g_n, g_0) = 0$*, so that* ψ_n^* *is asymptotically linear with influence curve* $D^*(d_0, Q, g_0)$*.*

For general g_n*, we also assume the following second order term condition:*

$$R_{1d_n}(Q_n, Q_0, g_n, g_0) - R_{1d_n}(Q, Q_0, g_n, g_0) = o_P(1/\sqrt{n}).$$

In addition, we assume the following asymptotic linearity condition on a smooth functional of g_n*:*

$$R_{1d_n}(Q, Q_0, g_n, g_0) = (P_n - P_0) D_g(P_0) + o_P(1/\sqrt{n}),$$

for some function $D_g(P_0)(O) \in L_0^2(P_0)$*.*

Then,

$$\psi_n^* - \psi_0 = (P_n - P_0)\{D^*(d_0, Q, g_0) + D_g(P_0)\} + o_P(1/\sqrt{n}). \tag{7}$$

If g_n *is an MLE of* g_0 *according to a correctly specified model* $\mathcal{G}$ *for* g_0 *with tangent space* $T_g(P_0)$ *at* P_0*, then it follows that*

$$D_g(P_0) = -\Pi(D^*(d_0, Q, , g_0) \mid T_g(P_0)),$$

where $\Pi(\cdot \mid T_g(P_0))$ *denotes the projection operator onto* $T_g(P_0) \subset L_0^2(P_0)$ *in the Hilbert space* $L_0^2(P_0)$*.*

The proof of this theorem is a straightforward consequence of the template presented before the theorem.

12.1 Asymptotic linearity of TMLE in SMART

Suppose the data is generated by a sequentially randomized controlled trial and there is no missingness so that g_0 is known. In addition, assume that $V(0)$ and $V(1)$ are both univariate scores, and assume condition (12) so that the optimal rule $d_{0,A(1)}$ based on $(A(0), V(0), V(1))$ is the same as the optimal rule $d_{0,A(1)}$ based on $A(0), V(1)$: e.g., $V(1)$ is the same score as $V(0)$ but measured at the next time-point, so that it is reasonable to assume that an effect of $V(0)$ on Y will be fully blocked by $V(1)$. Suppose we want to use the data of the RCT to learn the V-optimal rule d_0 and provide statistical inference for $E_{P_0} Y_{d_0}$. Since both $V(0)$ and $V(1)$ are 1-dimensional, using kernel smoothers or sieve-based estimation to generate a library of candidate estimators for the sequential loss-based super-learner of the blip-functions $(\bar{Q}_{10}, \bar{Q}_{20})$ described in previous section, we can obtain an estimator $\bar{Q}_n = (\bar{Q}_{1n}, \bar{Q}_{2n})$

of $\bar{Q}_0 = (\bar{Q}_{10}, \bar{Q}_{20})$ that converges at a rate such as $n^{-2/5}$ under the assumption that $\bar{Q}_{10}, \bar{Q}_{20}$ are continuously differentiable with a uniformly bounded derivative, or at a better rate under additional smoothness assumptions. As a consequence, in this case $R_2(Q_n, Q_0) = O_P(n^{-4/5})$ at minimal. As a consequence, all conditions of Theorem 10 hold, and it follows that the proposed TMLE is asymptotically linear with influence curve $D^*(d_0, Q, g_0)$, where $\bar{Q}$ is the possibly misspecified limit of $\bar{Q}_n^{d_n*}$ in the TMLE. To conclude, sequentially randomized controlled trials allow us to learn V-optimal rules at adaptive optimal rates of convergence, and allow valid asymptotic statistical inference for $E_{P_0}Y_{d_0}$. If $V(j)$ is higher dimensional, then one will have to rely on enough smoothness assumptions on the blip-functions in order to guarantee that our super-learner $\bar{Q}_n$ (and thus d_n) is still such that $R_2(Q_n, Q_0) = o_P(1/\sqrt{n})$.

If there is actual missingness or right-censoring, then $g_0 = g_{01}g_{02}$ factors in a treatment mechanism g_{01} and censoring mechanism g_{02}, where g_{01} is known, but g_{02} is typically not known. Having a lot of knowledge about how censoring depends on the observed past might make it possible to obtain a good estimator of g_{02}. In that case, the above conclusions still apply, but one now estimates the nuisance parameters of the loss-function (e.g., one uses a double robust loss-function in which g_{02} is replaced by an estimator).

12.2 Statistical inference

Suppose one is concerned with statistical inference for the target parameter $\psi_{1,0} \equiv E_{P_0}Y_{d_0} - E_{P_0}Y_0$, where Y_0 represents the counterfactual outcome $Y_{(0,1),(0,1)}$ for the static intervention that sets both treatments equal to 0, and, as always, sets censoring to "no censoring". Above we developed the TMLE for $E_{P_0}Y_{d_0}$, and we could use a separate TMLE for EY_0, or by modifying the TMLE described for $E_{P_0}Y_d$ by using clever covariates that are the difference of the clever covariates one would use for EY_d and EY_0, we can use a TMLE directly targeting $\psi_{1,0}$. This results in a TMLE $\psi_{1,n}^*$ of $\psi_{1,0}$. By a slight generalization of Theorem 4, if $g_n = g_0$ is known, this TMLE of $\psi_{1,0} = E_{P_0}Y_{d_0} - E_{P_0}Y_0$ is asymptotically linear with influence curve

$$IC(P_0) = \{D^*(d_0, Q, g_0) - D^*(d = 0, Q, g_0),$$

where $D^*(d = 0, Q_0, g_0)$ is the efficient influence curve of EY_0 (i.e, $d = 0$ represents the static intervention $A = ((0,1),(0,1))$). If g_n is an MLE according to a model with tangent space $T_g(P_0)$, then the TMLE is asymptotically linear with influence curve

$$IC(P_0) - \Pi(IC(P_0) \mid T_g(P_0)),$$

so that one could still use $IC(P_0)$ as a conservative influence curve. Let IC_n be an estimator of this influence curve $IC(P_0)$ obtained by plugging in the available estimates of its unknown components. The asymptotic variance of the TMLE $\psi^*_{1,n}$ of $\psi_{1,0}$ can now be (conservatively) estimated with

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} IC_n^2(O_i).$$

An asymptotic 0.95-confidence interval for $\psi_{1,0}$ is given by $\psi^*_{1,n} \pm 1.96\sigma_n/\sqrt{n}$. In particular, we can test a null-hypothesis $H_0 : \psi_{1,0} = 0$ to determine if there is statistically significant evidence that an optimal treatment rule outperforms the current standard treatment $A = 0$.

13 Statistical inference for mean outcome under data adaptively determined dynamic treatment

Let $\hat{d} : \mathcal{M}_{NP} \rightarrow \mathcal{D}$ be an estimator that maps an empirical distribution into an individualized treatment rule. Let $d_n = \hat{d}(P_n)$ be the estimated rule. Up till now we have been concerned with statistical inference for $E_{P_0}Y_{d_0}$, where d_0 is the unknown V-optimal rule while d_n is a best estimator of this rule. As a consequence, statistical inference for $E_{P_0}Y_{d_0}$ based on the TMLE relied on consistency of d_n to d_0, but also relied on a rate of convergence at which d_n needs to converge to d_0: i.e., $R_2(Q_n, Q_0) = o_P(1/\sqrt{n})$. In this section we present statistical inference for the data adaptive target parameter $\psi_{0n} = \Psi_{d_n}(P_0) = E_{P_0}Y_d|_{d=d_n}$. That is, we construct an estimator $\psi^*_n = \hat{\Psi}^*(P_n)$ of $\Psi_{\hat{d}(P_n)}(P_0)$ and a confidence interval $\psi^*_n \pm 1.96\sigma_n/\sqrt{n}$ so that

$$P_0\left(\Psi_{\hat{d}(P_n)}(P_0) \in \hat{\Psi}_{\hat{d}(P_n)}(P_n) \pm 1.96\hat{\sigma}(P_n)/\sqrt{n}\right) \rightarrow 0.95, \text{ as } n \rightarrow \infty.$$

Note that in this definition of the confidence interval the target parameter is itself also random variable through the data P_n.

Statistical inference will be based on the same TMLE of $\Psi_d(P_0)$ at $d = d_n$, and our variance estimator will also be the same, but since the target is not $\Psi_{d_0}(P_0)$ but $\Psi_{d_n}(P_0)$, there will be no need for any consistency or rate condition on d_n. As a consequence, this approach is particularly appropriate in cases where V is high dimensional so that it is not reasonable to expect that d_n converges to d_0 at the required rate. In addition, even when statistical inference

for $E_{P_0}Y_{d_0}$ is feasible, one might be interested in statistical inference for the mean outcome under the actual concretely available rule d_n instead of under the unknown rule d_0.

As previously shown, we have $P_0D^*(d_n, Q_n^*, g_n) = \psi_{0n} - \psi_n^* + R_{d_n}(Q_n^*, Q_0, g_n, g_0)$ and $P_nD^*(d_n, Q_n^*, g_n) = 0$, which yields

$$\psi_n^* - \psi_{0n} = (P_n - P_0)D^*(d_n, Q_n^*, g_n) + R_{d_n}(Q_n^*, Q_0, g_n, g_0).$$

Analogue to Theorem 10 we now have the following theorem.

Theorem 11 *Assume $Y \in [0,1]$. Let $\hat{d}(P_n) \in \mathcal{D}$ with probability tending to 1, and assume the positivity assumption $\inf_{d\in\mathcal{D}} P_0(g_0(A = d(L), L) > 0) = 1$. Let $\psi_{0n} = \Psi_{d_n}(P_0) = E_{P_0}Y_d|_{d=d_n}$ be the data adaptive target parameter of interest. Consider the TMLE (Q_n^*, g_n) of $\Psi_{d_n}(Q_0)$ treating d_n as fixed, and $\psi_n^* = \Psi_{d_n}(Q_n^*)$ is the TMLE of ψ_{0n}. Let $R_{1d}(Q, Q_0, g, g_0) = P_0D^*(d, Q, g) - \{\Psi_d(Q_0) - \Psi_d(Q)\}$, as defined as in Theorem 9.*

Assume $D_n^* \equiv D^*(d_n, Q_n^*, g_n)$ falls in a P_0-Donsker class with probability tending to 1, $P_0\{D_n^* - D^*(d_0, Q, g_0)\}^2$ converges to zero in probability for some $d_0 \in \mathcal{D}$, $Q_{L(0)} = Q_{L(0),0}$.

Then,

$$\psi_n^* - \psi_0 = (P_n - P_0)D^*(d_0, \bar{Q}, Q_{W,0}, g_0) + R_{1d_n}(Q_n, Q_0, g_n, g_0) + o_P(n^{-1/2}).$$

If $g_n = g_0$ (i.e., RCT), then $R_{1d_n}(Q_n, Q_0, g_n, g_0) = 0$, so that ψ_n^* is asymptotically linear with influence curve $D^*(d_0, Q, g_0)$.

For general g_n, we also assume that the following second order term condition:

$$R_{1d_n}(Q_n, Q_0, g_n, g_0) - R_{1d_n}(Q, Q_0, g_n, g_0) = o_P(1/\sqrt{n}).$$

In addition, we assume the following asymptotic linearity condition (for a smooth functional of g_n):

$$R_{1d_n}(Q, Q_0, g_n, g_0) = (P_n - P_0)D_g(P_0) + o_P(1/\sqrt{n}),$$

for some function $D_g(P_0) \in L_0^2(P_0)$.

Then,

$$\psi_n^* - \psi_0 = (P_n - P_0)\{D^*(d_0, Q, g_0) + D_g(P_0)\} + o_P(1/\sqrt{n}). \tag{8}$$

If g_n is an MLE of g_0 according to a correctly specified model $\mathcal{G}$ for g_0 with tangent space $T_g(P_0)$ at P_0, then it follows that

$$D_g(P_0) = -\Pi(D^*(d_0, Q, , g_0) \mid T_g(P_0)),$$

where $\Pi(\cdot \mid T_g(P_0))$ denotes the projection operator onto $T_g(P_0) \subset L_0^2(P_0)$ in the Hilbert space $L_0^2(P_0)$.

14 Statistical Inference for average of sample-split specific mean counterfactual outcomes under data adaptively determined dynamic treatments.

Let $\mathcal{D}$ be an index set for a collection of individualized treatment rules, and for each $d \in \mathcal{D}$, we have a statistical target parameters $\Psi_d : \mathcal{M} \to \mathbb{R}$, defined by $\Psi_d(P) = E_P Y_d$. Let $\hat{d} : \mathcal{M}_{NP} \to \mathcal{D}$ be an estimator that maps an empirical distribution into an individualized treatment rule, and thereby a choice of target parameter. Consider a cross-validation sample split random vector $B_n \in \{0,1\}^n$, and for a split B_n, let P^0_{n,B_n} be the empirical distribution of the training sample $\{i : B_n(i) = 0\}$ and P^0_{n,B_n} is the empirical distribution of the validation sample $\{i : B_n(i) = 1\}$. In this section, we are concerned with presenting a method that provides an estimator and statistical inference for the data-adaptive target parameter

$$\psi_{0n} = E_{B_n} \Psi_{\hat{d}(P^0_{n,B_n})}(P_0).$$

Let J be the number of possible values of B_n. Thus for each of the J training samples one applies the estimator $\hat{d}$, giving a target parameter value $\Psi_{\hat{d}(P^0_{n,B_n})}(P_0)$, and our target parameter ψ_{0n} is defined as the average across these J target parameters. Below we present a cross-validated TMLE ψ^*_n of this data adaptive target parameter ψ_{0n}. We will be able to establish statistical inference for this parameter ψ_{0n}, not only without relying on a consistency or rate condition on the estimated rule as achieved in the previous section, but also removing the reliance on the empirical process condition (i.e., Donsker class condition) that was needed in any of the previous theorems. That means that in a sequentially randomized controlled trial, we obtain valid asymptotic statistical inference without any conditions, even when d_n is a highly data adaptive estimator of a V-optimal rule for a possibly high dimensional V.

The next subsection defines the general cross-validated TMLE for data adaptive target parameters. Subsequently, we present an asymptotic linearity theorem allowing us to construct asymptotic 0.95-confidence intervals. Finally, we present the cross-validated TMLE for the two time-point treatment case in detail.

14.1 Cross-validated TMLE

For each target parameter Ψ_d, let $D_d^*(P_0)$ be its efficient influence curve at P_0. Assume that $\Psi_d(P_0) = \Psi_d(Q_0^d)$ only depends on P_0 through a parameter Q_0^d, and assume that $D_d^*(P_0) = D_d^*(Q_0^d, g_0^d)$ depends on P_0 through Q_0^d and a nuisance parameter g_0^d. Define a second order term $R^d(Q^d, Q_0^d, g^d, g_0^d)$ as follows:

$$P_0 D_d^*(Q^d, g^d) = \Psi_d(P_0) - \Psi_d(Q^d) + R^d(Q^d, Q_0^d, g^d, g_0^d).$$

Let $\hat{Q}^d$, $\hat{g}^d$ be initial estimators of Q_0^d and g_0^d, respectively. Let $L^d(Q^d)$ be a valid loss function for Q_0^d so that $Q_0^d = \arg\min_{Q^d} P_0 L^d(Q^d)$, and let $\{Q^d(\epsilon) : \epsilon\}$ be a submodel through Q at $\epsilon = 0$ with a univariate or multivariate parameter ϵ so that the linear span of the generalized score includes the efficient influence curve at (Q^d, g^d):

$$D_d^*(Q^d, g^d) \in \langle \frac{d}{d\epsilon} L^d(Q^d(\epsilon))\Big|_{\epsilon=0} \rangle,$$

where $\langle f \rangle = \{\sum_j \beta_j f_j : \beta\}$ denotes the linear space spanned by the components of f. Let $\hat{Q}^d(\epsilon)$ be this submodel through $\hat{Q}^d$, using $\hat{g}^d$. For the single time point treatment case, we define

$$\epsilon_n = \arg\min_{\epsilon} E_{B_n} P_{n,B_n}^1 \, L^d(\hat{Q}^d(P_{n,B_n}^0)(\epsilon))\Big|_{d=\hat{d}(P_{n,B_n}^0)},$$

but for the multiple time-point treatment case, we use the sequential TMLE algorithm of the TMLE for $EY_{\hat{d}(P_{n,B_n}^0)}$ but where the ϵ's are determined based on the cross-validated empirical risks averaging over the training samples. In a later subsection, we demonstrate this in detail.

For notational convenience, we use the notation $\hat{Q}(P_{n,B_n}^0) = \hat{Q}^{\hat{d}(P_{n,B_n}^0)}(P_{n,B_n}^0)$, and similarly, we define $\hat{g}(P_{n,B_n}^0) = \hat{g}^{\hat{d}(P_{n,B_n}^0)}(P_{n,B_n}^0)$. For each split B_n, we define the corresponding updates $\hat{Q}^*(P_{n,B_n}^0, \epsilon_n) \equiv \hat{Q}(P_{n,B_n}^0)(\epsilon_n)$. The key assumption about ϵ_n and a corresponding update $\hat{Q}^*(P_{n,B_n}^0, \epsilon_n)$ is that it solves the cross-validated empirical mean of the efficient influence curve:

$$E_{B_n} P_{n,B_n}^1 D_{\hat{d}(P_{n,B_n}^0)}^*(\hat{Q}^*(P_{n,B_n}^0, \epsilon_n), \hat{g}(P_{n,B_n}^0)) = o_P(1/\sqrt{n}). \quad (9)$$

As shown below, the sequential TMLE updating algorithm for the multiple time point intervention case indeed satisfies this equation with $o_P(1/\sqrt{n})$ replaced by 0.

The proposed estimator of ψ_{0n} is given by

$$\psi_n^* \equiv E_{B_n} \Psi_{\hat{d}(P_{n,B_n}^0)}(\hat{Q}^*(P_{n,B_{n^0}}, \epsilon_n)).$$

In the current literature we have referred to this estimator as the cross-validated TMLE (Zheng and van der Laan (2010, 2012); van der Laan and Petersen (2012); Diaz and van der Laan (2013)). The only twist relative to the original CV-TMLE is that we change our target on each training sample into the training sample specific target parameter implied by the fitted rule on the training sample, while in the original CV-TMLE formulation, the target would still be $\Psi_{d_0}(P_0)$. With this minor twist, the (same) CV-TMLE is now used to target the average of training sample specific target parameters averaged across the J training samples. This utilization of CV-TMLE was already used to estimate the average (across training samples) of the true risk of an estimator based on a training sample in (van der Laan and Petersen, 2012; Diaz and van der Laan, 2013), so that this represents a generalization of that application of CV-TMLE to general data adaptive target parameters as proposed in van der Laan et al. (2013).

14.2 Statistical inference based on the CV-TMLE

Let's now proceed with the analysis of this CV-TMLE ψ_n^* of ψ_{0n}. A key identity is given by:

$$E_{B_n} P_0 D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n},\epsilon_n),\hat{g}(P^0_{n,B_n})) = E_{B_n}\Psi_{\hat{d}(P^0_{n,B_n})}(P_0) - \psi_n^*$$
$$+E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}),Q_0,\hat{g}(P^0_{n,B_n}),g_0).$$

This proves

$$\psi_n^* - \psi_{0n} = E_{B_n}(P^1_{n,B_n} - P_0) D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n},\epsilon_n),\hat{g}(P^0_{n,B_n}))$$
$$+o_P(1/\sqrt{n}) + E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}),Q_0,\hat{g}(P^0_{n,B_n}),g_0).$$

Regarding the empirical process term we have the following lemma.

Lemma 2 *Assume that the supremum norm of* $D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n},\epsilon_n),\hat{g}(P^0_{n,B_n}))$ *is bounded by some* $M < \infty$ *with probability tending to 1, and that*

$$P_0\{D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n},\epsilon_n),\hat{g}(P^0_{n,B_n})) - D^*_{d_0}(Q^{d_0},g^{d_0})\}^2 \to 0 \text{ in probability.}$$

Then,

$$E_{B_n}(P^1_{n,B_n} - P_0) D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n},\epsilon_n),\hat{g}(P^0_{n,B_n})) = (P_n - P_0) D^*_{d_0}(Q^{d_0},g^{d_0})$$
$$+o_P(1/\sqrt{n}).$$

Thus, under this very mild consistency condition, we have

$$\psi_n^* - \psi_{0n} = (P_n - P_0)D_{d_0}^*(Q^{d_0}, g^{d_0}) + o_P(1/\sqrt{n}) \\ + E_{B_n} R_{\hat{d}(P_{n,B_n}^0)}(\hat{Q}^*(P_{n,B_n}^0), Q_0, \hat{g}(P_{n,B_n}^0), g_0).$$

Suppose now that $Q^{d_0} = Q_0^{d_0}$ and $g^{d_0} = g_0^{d_0}$, and

$$E_{B_n} R_{\hat{d}(P_{n,B_n}^0)}(\hat{Q}^*(P_{n,B_n}^0), Q_0, \hat{g}(P_{n,B_n}^0), g_0) = o_P(1/\sqrt{n}).$$

Then, it follows that

$$\psi_n^* - \psi_{0n} = (P_n - P_0)D_{d_0}^*(Q_0^{d_0}, g_0^{d_0}) + o_P(1/\sqrt{n}).$$

In general, we only assume that $g^{d_0} = g_0^{d_0}$, and

$$E_{B_n} R_{\hat{d}(P_{n,B_n}^0)}(\hat{Q}^*(P_{n,B_n}^0), Q_0, \hat{g}(P_{n,B_n}^0), g_0) - E_{B_n} R_{\hat{d}(P_{n,B_n}^0)}(Q, Q_0, \hat{g}(P_{n,B_n}^0), g_0) \\ = o_P(1/\sqrt{n}).$$

In many applications, due to linearity of $(Q - Q_0) \to R_d(Q, Q_0, g, g_0)$, this difference is represented by an integral involving the product of a difference $\hat{Q}^*(P_{n,B_n}^0) - Q$ and a difference $\hat{g}(P_{n,B_n}^0) - g_0$. In that case, this assumption correspond with a second order term being $o_P(1/\sqrt{n})$, where the second order term might be bounded by an L^2-norm of a difference $\hat{Q}^*(P_{n,B_n}^0) - Q$ times an L^2-norm of a difference $\hat{g}(P_{n,B_n}^0) - g_0$. In addition, we assume the following asymptotic linearity condition on $\hat{g}$:

$$E_{B_n} R_{\hat{d}(P_{n,B_n}^0)}(Q, Q_0, \hat{g}(P_{n,B_n}^0), g_0) = (P_n - P_0)D_g(P_0) + o_P(1/\sqrt{n}).$$

Then, we can conclude:

$$\psi_n^* - \psi_{0n} = (P_n - P_0)\{D_{d_0}^*(Q, g_0) + D_g(P_0)\} + o_P(1/\sqrt{n}).$$

This proves the following theorem.

Theorem 12 *Let $\mathcal{D}$ be an index set for a collection of individualized treatment rules, and for each $d \in \mathcal{D}$, we have a statistical target parameters $\Psi_d : \mathcal{M} \to \mathbb{R}$, defined by $\Psi_d(P) = E_P Y_d$. Assume $\inf_{d \in \mathcal{D}} P_0(g_0(A = d(L), L) > 0) = 1$Let $\hat{d} : \mathcal{M}_{NP} \to \mathcal{D}$ be an estimator that maps an empirical distribution into an individualized treatment rule, and thereby a choice of target parameter. Consider a sample split random vector $B_n \in \{0,1\}^n$, and for a split B_n, let P_{n,B_n}^0 be the empirical distribution of training sample $\{i : B_n(i) = 0\}$ and*

P^0_{n,B_n} be the empirical distribution of the validation sample $\{i : B_n(i) = 1\}$. The data-adaptive target parameter is defined as follows:

$$\psi_{0n} = E_{B_n}\Psi_{\hat{d}(P^0_{n,B_n})}(P_0).$$

For each target parameter Ψ_d, let $D^*_d(P_0)$ be its efficient influence curve at P_0. Assume that $\Psi_d(P_0) = \Psi_d(Q^d_0)$ only depends on P_0 through a parameter Q^d_0, and assume that $D^*_d(P_0) = D^*_d(Q^d_0, g^d_0)$ depends on P_0 through Q^d_0 and a nuisance parameter g^d_0. Define a second order term $R^d()$ as follows:

$$P_0 D^*_d(Q^d, g^d) = \Psi_d(P_0) - \Psi_d(Q^d) + R^d(Q^d, Q^d_0, g^d, g^d_0).$$

Let $(Q^d, O) \rightarrow L^d(Q^d)(O)$ be a valid loss function for Q^d_0 so that $Q^d_0 = \arg\min_{Q^d} P_0 L^d(Q^d)$, and let $\{Q^d(\epsilon) : \epsilon\}$ be a submodel through Q at $\epsilon = 0$ with a univariate or multivariate parameter ϵ so that the linear span of the generalized score includes the efficient influence curve:

$$D^*_d(Q^d, g^d) \in \langle \frac{d}{d\epsilon} L^d(Q^d(\epsilon))\Big|_{\epsilon=0} \rangle.$$

Let $\hat{Q}^d(\epsilon)$ be this submodel through $\hat{Q}^d$, using $\hat{g}^d$. For notational convenience, we use the notation $\hat{Q}(P^0_{n,B_n}) = \hat{Q}^{\hat{d}(P^0_{n,B_n})}(P^0_{n,B_n})$, and similarly, we define $\hat{g}(P^0_{n,B_n}) = \hat{g}^{\hat{d}(P^0_{n,B_n})}(P^0_{n,B_n})$. For each split B_n, we define the corresponding updates $\hat{Q}^*(P^0_{n,B_n}, \epsilon_n) \equiv \hat{Q}(P^0_{n,B_n})(\epsilon_n)$. Let ϵ_n be computed so that it solves/satisfies the following equation:

$$E_{B_n} P^1_{n,B_n} D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), \hat{g}(P^0_{n,B_n})) = o_P(1/\sqrt{n}). \tag{10}$$

The proposed estimator of ψ_{0n} is given by

$$\psi^*_n \equiv E_{B_n}\Psi_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P_{n,B^0_n}, \epsilon_n)).$$

Assume that the supremum norm of $D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), \hat{g}(P^0_{n,B_n}))$ is bounded by some $M < \infty$ with probability tending to 1, and that

$P_0\{D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*_{d_0}(Q^{d_0}, g^{d_0})\}^2 \rightarrow 0$ *in probability.Then,*

$$\begin{aligned}\psi^*_n - \psi_{0n} &= (P_n - P_0)D^*_{d_0}(Q^{d_0}, g^{d_0}) + o_P(1/\sqrt{n}) \\ &+ E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), Q_0, \hat{g}(P^0_{n,B_n}), g_0).\end{aligned}$$

In general, we will assume $g = g_0$, and

$$E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), Q_0, \hat{g}(P^0_{n,B_n}), g_0) - E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(Q, Q_0, \hat{g}(P^0_{n,B_n}), g_0) = o_P(1/\sqrt{n}),$$

and the following asymptotic linearity condition on $\hat{g}$:

$$E_{B_n} R_{\hat{d}(P^0_{n,B_n})}(Q, Q_0, \hat{g}(P^0_{n,B_n}), g_0) = (P_n - P_0) D_g(P_0) + o_P(1/\sqrt{n}).$$

Then,

$$\psi^*_n - \psi_{0n} = (P_n - P_0)\{D^*_{d_0}(Q, g_0) + D_g(P_0)\} + o_P(1/\sqrt{n}).$$

Suppose g_0 is known and $\hat{g}(P_n) = g_0$. Consider the estimator

$$\sigma^2_n = E_{B_n} P^1_{n,B_n} \left\{ D^*_{\hat{d}(P^0_{n,B_n})}(\hat{Q}^*(P^0_{n,B_n}, \epsilon_n), \hat{g}(P^0_{n,B_n})) \right\}^2$$

of the asymptotic variance $\sigma^2_0 = P_0\{D^*_{d_0}(Q, g_0)\}^2$ of the CV-TMLE ψ^*_n. An asymptotic 0.95-confidence interval for ψ_{0n} is given by $\psi^*_n \pm 1.95\sigma_n/\sqrt{n}$. This same variance estimator and confidence interval can be used for the case that g_0 is not known and $\hat{g}(P_n)$ is an MLE of g_0 according to some model. In that case, the theorem tells us that it is an asymptotically conservative confidence interval.

14.3 CV-TMLE of the mean outcome under data adaptive V-optimal rule: two time-point treatment

Let $\hat{d}$ be the data adaptive estimator of the V-optimal rule d_0, as presented in a previous section. Firstly, without loss of generality we can assume that $Y \in [0, 1]$. Let's denote the realizations of B_n with $j = 1, \ldots, J$. Let $\bar{Q}_{2nj}$ be an initial estimator of $\bar{Q}^{d_{nj}}_{20} = E_{P_0}(Y \mid \bar{A}(1) = d_{nj}(\bar{L}(1)), \bar{L}(1))$ based on the training sample P^0_{nj}, and similarly let d_{nj} and g_{nj} represent the estimated rule and estimated intervention mechanism based on this training sample P^0_{nj}, $j = 1, \ldots, J$. Consider the submodel $\text{Logit}\hat{\bar{Q}}_{2n,j}(\epsilon) = \text{Logit}\hat{\bar{Q}}_{2nj} + \epsilon H_2(g_{nj})$, where

$$H_2(g_{nj}) = \frac{I(\bar{A}(1) = d_{nj}(\bar{L}(1)))}{\prod^1_{l=0} g_{nj,A(l)}(O)}.$$

Let

$$\epsilon_{2n} = \arg\min_{\epsilon} \frac{1}{J} \sum_{j=1}^{J} P^1_{nj} L_{2,nj}(\bar{Q}_{2n,j}(\epsilon)),$$

where

$$-L_{2nj}(\bar{Q}_2) = I(\bar{A}(1) = d_{nj}(\bar{L}(1))) \left\{ Y \log \bar{Q}_2(\bar{L}(1)) + (1-Y) \log(1 - \bar{Q}_2(\bar{L}(1)) \right\}.$$

This estimator of ϵ can be obtained by fitting ϵ with univariate logistic regression of Y on $H_2(g_{nj})$ using $\text{Logit}\bar{Q}_{2nv}$ as off-set, but where observations Y_i in validation sample are coupled to a corresponding offset $\bar{Q}_{2nj}(\bar{L}_i(1))$ and covariate $H_2(g_{nj})(\bar{L}_i)$ based on the corresponding training sample. This defines a targeted estimator $\hat{\bar{Q}}_2^*(P_{nj}^0, \epsilon_{2n}) = \bar{Q}_{2nj}(\epsilon_{2n})$ for each $j = 1, \ldots, J$. We will denote this targeted estimator with $\bar{Q}_{2nj}^*$, and note that it only depends on P_n through the training sample P_{nj}^0 and ϵ_{2n}.

Regress $\bar{Q}_{2nj}$ on $L(0), A(0) = d_{nj,A(0)}(L(0))$ which defines an initial estimator $\bar{Q}_{1nj}$ of $\bar{Q}_{10}^{d_{nj}} = E_{P_0}(Y_{d_{nj}} \mid L(0)) = E_{P_0}(\bar{Q}_{20}^{d_{nj}} \mid L(0), A(0) = d_{nj,A(0)}(L(0)))$. Consider the submodel $\text{Logit}\bar{Q}_{1nj}(\epsilon) = \text{Logit}\bar{Q}_{1nj} + \epsilon H_1(g_{nj})$, where

$$H_1(g_{nj}) = \frac{I(A(0) = d_{nj,A(0)}(L(0)))}{g_{nj,A(0)}(O)}.$$

Let

$$\epsilon_{1n} = \arg\min_{\epsilon} \frac{1}{J} \sum_{j=1}^{J} P_{nj}^1 L_{1,nj}(\bar{Q}_{1nj}(\epsilon)),$$

where

$$\begin{aligned} &-L_{1,nj}(\bar{Q}_1) \\ &= I(A(0) = d_{nj,A(0)}(V(0)) \left\{ \bar{Q}_{2nj} \log \bar{Q}_1(L(0)) + (1 - \bar{Q}_{2nj}) \log(1 - \bar{Q}_1(L(0)) \right\}. \end{aligned}$$

This defines a targeted estimator $\hat{\bar{Q}}_1^*(P_{nj}^0, \epsilon_{1n}) = \bar{Q}_{1nj}(\epsilon_{1n})$ of $\bar{Q}_{10}^{d_{nj}}$, $j = 1, \ldots, J$. We will denote this targeted estimator with $\bar{Q}_{1nj}^*$ and note that it only depends on P_n through the training sample P_{nj}^0 and ϵ_{1n}.

Let $Q_{L(0),nj}$ be the empirical distribution of $L_i(0)$ for the training sample P_{nj}^0. This defines an estimator $\psi_{nj}^* = Q_{L(0),nj}\bar{Q}_{1nj}^* = P_{nj}^1 \bar{Q}_{1nj}^*$ of $\psi_{d_{nj}0} = \Psi_{d_{nj}}(P_0)$ for each $j = 1, \ldots, J$. The cross-validated TMLE is now defined as $\psi_n^* = \frac{1}{J} \sum_{j=1}^{J} \psi_{nj}^*$.

This CV-TMLE solves the cross-validated efficient influence curve equation:

$$0 = \frac{1}{J} \sum_{j=1}^{J} P_{nj}^1 D_{d_{nj}}^*(Q_{nj}^*, g_{nj}),$$

where $Q_{nj}^* = (Q_{1nj}^*, Q_{2nj}^*, Q_{L(0),nj})$ only depends on P_n through P_{nj}^0 and $\epsilon_n = (\epsilon_{1n}, \epsilon_{2n})$. Our general asymptotic linearity Theorem 12 can thus be immediately applied to this CV-TMLE.

Recall that in the second part of the article, we suggested using a CV-TMLE to estimate the risk $E_{B_n}E_{P_0}Y_{\hat{d}(P^0_{n,B_n})}$ for a candidate estimator $\hat{d}$, and to define a cross-validation selector accordingly, resulting in a particular type of super-learner. Thus the above description of CV-TMLE defines this desired estimator, and could thus also be used to define this super-learner.

15 Concluding remarks

This article investigated nonparametric estimation of a V-optimal dynamic treatment, statistical inference for the mean outcome under the V-optimal rule, and statistical inference for the (data adaptive target parameter defined as the) mean outcome under a data adaptively determined V-optimal rule (treating the latter as given). We proposed sequential loss-based super learning with novel choices of loss-functions to construct such a nonparametric estimator of the V-optimal rule. When applied in sequentially randomized controlled trials, at each stage, this method is guaranteed to asymptotically outperform any competitor (w.r.t. loss-based dissimilarity) by simply including it in the library of candidate estimators. In this sequential loss-based super-learner the cross-validation is used to optimize the performance in fitting the V-adjusted blip-function itself. We also proposed a cross-validation selector (and corresponding super-learner) that aims to optimize the performance of the fitted rule itself in maximizing the mean outcome. The latter seems to be more targeted towards our goal, but theoretical results regarding the cross-validation selector tell a more complex story, suggesting that only when V is higher dimensional, the latter can be expected to be superior. We plan to carry out simulation studies to shed light on this important issue.

We proved a surprising/useful result stating that the mean outcome under the V-optimal rule is represented by a statistical parameter whose pathwise derivative is identical to what it would have been if the unknown rule would be treated as known. As a consequence, the efficient influence curve is immediately known, and any of the efficient estimators for the mean outcome under a given rule can be applied at the estimated rule. In particular, we demonstrate a TMLE, and present the asymptotic linearity theorem. However, the dependence of the statistical target parameter on the unknown rule affects the second order terms of the TMLE, and, as a consequence, the asymptotic linearity of the TMLE requires that a second order difference between the estimated rule and the V-optimal rule converges to zero at a rate faster than $1/\sqrt{n}$. We show that this can be expected to hold for rules that are only a function of one continuous score (such as a biomarker), but when V is higher

dimensional, only strong smoothness assumptions will guarantee this, so that, even in an RCT, we cannot be guaranteed valid statistical inference for such V-optimal rules.

Therefore, we proceeded to pursue statistical inference for so called data adaptive target parameters. Specifically, we presented statistical inference for the mean outcome under the dynamic treatment regimen we fitted based on the data. We now show that statistical inference for this data adaptive target parameter does not rely on the second order term condition anymore, and only requires that the data adaptively fitted rule converges to some fixed rule. However, even in a sequentially randomized controlled trial, the asymptotic linearity theorem still relies on a Donsker class condition that limits the data adaptivity of the estimator of the rule. So, even though the assumptions are much weaker, they can still cause havoc when V is high dimensional in finite samples, and possibly, even asymptotically.

Therefore, we proceeded with the average of sample split specific target parameters, as in general proposed in (van der Laan et al., 2013), where we show that statistical inference can now avoid the empirical process condition. Specifically, our data adaptive target parameter is now defined as an average across J sample splits in training and validation sample of the mean outcome under the dynamic treatment fitted on the training sample. We present a cross-validated TMLE of this data adaptive target parameter, and we established an asymptotic linearity theorem that does neither require a consistency or rate condition on the estimated rule, nor does it require the empirical process condition. As a consequence, in a sequential RCT, this method provides valid asymptotic statistical inference without any conditions, beyond the requirement that the estimated rule converges to some fixed rule. In future work we hope to address the practical performance of these methods in a simulation study and apply it to actual data sets of interest, generated by observational as well as randomized controlled trials.

In the current article we defined the treatment as binary at each time point. Consider now a treatment that has k possible values. In that case, we can define a vector of binary indicators, ordered in a user-supplied manner, that identify the treatment. We can now just apply the results for the multiple time-point treatment case in the Appendix, since this represents a special case in which at some time-point there are no inter mediate time-dependent covariates between two subsequent binary treatments. As a consequence, our results also apply to this case.

It might also be of interest to propose working models for the mean outcome $E_{P_0}(Y_{d_0} \mid S)$ under the optimal rule, conditional on some baseline covariates $S \subset W$. This is now a function of S, but we would define the target parameter

of interest as a projection of this true underlying function on the working model. It would now be of interest to develop TMLE for this finite dimensional pathwise differentiable parameter, and we presume that similar results as we found here might appear. Such parameters provide information about how the mean outcome under the optimal rule are affected by certain baseline characteristics.

Acknowledgement

This research was supported by an NIH grant R01 AI074345-06.

References

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–72, 2005.

R. Bellman. *Dynamic Programming*. Univ. Press, Princeton, NJ, 1957.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

B. Chakraborti and S.A. Murphy. Dynamic treatment regimens. *Annu. Rev. Stat. Appl.*, 1:1–18, 2013.

B. Chakraborti, S.A. Murphy, and V. Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.*, 19:317–43, 2010.

C. Cotton and P. Heagerty. A data augmentation method for estimating the causal effect of adherence to treatment regimens targeting control of an intermediate measure. *Stat. Biosc*, 3:28–44, 2011.

I. Diaz and M.J. van der Laan. Targeted data adaptive estimation of the causal dose response curve. Technical Report 306, Division of Biostatistics, University of California, Berkeley, submitted to JCI, 2013.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6:503–556, 2005.

R. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

Y. Goldberg and M. Kosorok. Q-learning with censored data. *Ann. Stat.*, 40: 529–60, 2012.

M.A. Hernan, E. Lanoy, D. Costagliola, and J.M. Robins. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol*, 98:237–242, 2006.

P.W. Holland. Statistics and causal inference. *J Am Stat Assoc*, 81(396): 945–960, 1986.

H. Jones. Reinforcement-based treatment for pregnant drug abusers. *ClinicalTrials.gov data base, updated October 19, 2012*, Natl. Inst. Health., Bethesda, MD, http:// clinicaltrials.gov/ct2/show/NCT01177982:accessed July 24, 2013, 2010.

C. Kasari. Developmental and augmented intervention for facilitating expressive language. *ClinicalTrials.gov database, updated Apr. 26, 2012*, Natl. Inst. Health., Bethesda, MD, http:// clinicaltrials.gov/ct2/show/NCT01013545: accessed July 24, 2013, 2009.

P. Lavori and R. Dawson. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society, Series A*, 163, 2000.

P. Lavori and R. Dawson. Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1, 2004.

P. Lavori and R. Dawson. Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.*, 59:443–453, 2008.

H. Lei, I. Nahum-Shani, K. Lynch, D. Oslin, and S. Murphy. A smart design for building individualized treatment sequences. *Annu. Rev. Clin. Psychol.*, 8:21–48, 2011.

E. Moodie, R. Platt, and M. Kramer. Estimating response-maximized decision rules with applications to breastfeeding. *J. Am. Stat. Assoc.*, 104:155–65, 2009.

E. Moodie, B. Chakraborty, and M. Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.*, 40:629–45, 2012.

S. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24, 2005.

S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–66, 2003.

I. Nahum-Shani, M. Qian, D. Almiral, W. Pelham, and et al. B. Gnagy B. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychol. Methods*, 17:457–77, 2012a.

I. Nahum-Shani, M. Qian, D. Almiral, W. Pelham, and et al. B. Gnagy B. Q-learning: a data analysis method for constructing adaptive interventions. *Psychol. Methods*, 17:478–94, 2012b.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480, 1990.

L. Orellana, A. Rotnitzky, and J.M. Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *Int. J. Biostat.*, page 6:8, 2010a.

D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Mach. Learn.*, 49:161–78, 2002.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2000.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M.J. van der Laan. Targeted minimum loss based estimation of marginal structural working models. *Journal of Causal Inference*, submitted, technical report http://biostats.bepress.com/ucbbiostat/paper312/, 2013.

M.L Petersen, S.G. Deeks, J.N. Martin, and M.J. van der Laan. History-adjusted marginal structural models to estimate time-varying effect modification. *Am J Epidemiol*, 166(9):985–993, 2007.

M.L. Petersen, M.J. van der Laan, S. Napravnik, J.J. Eron, R.D. Moore, and S.G. Deeks. Long-term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification. *AIDS*, 22 (16):2097–106, 2008.

E.C. Polley and M.J. van der Laan. Predicting optimal treatment assignment based on prognostic factors in cancer patients. In Karl E. Peace, editor, *in Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints*. Chapman and Hall, 2009.

E.C. Polley, Sherri Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2012.

M. Qian and S. Murphy. Performance guarantees for individualized treatment rules. *Ann. Stat.*, 39:1180–210, 2011.

J.M. Robins. Discussion of "optimal dynamic treatment regimes" by susan a. murphy. *Journal of the Royal Statistical Society: Series B*, 65(2):355–66, 2003.

J.M. Robins. Optimal structural nested models for optimal sequential decisions. *Proc. Seattle Symp. Biostat.*, 2nd, ed. D Lin, P Heagerty, New York: Springer:189–326, 2004.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

J.M. Robins. 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" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987a. ISSN 0097-4943.

J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis (40, Supplement)*, 2:139s–161s, 1987b.

J.M. Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceeding of the Biopharmaceutical section*, pages 24–33. American Statistical Association, 1993.

J.M. Robins. Causal inference from complex longitudinal data. In Editor M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer-Verlag, New York, 1997.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN, 1997)*, pages 95–133. Springer-Verlag, New York, 2000.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser, 1992.

J.M. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.*, 27:4678–721, 2008.

S. Rosthj, C. Fullwood, R. Henderson, and S. Stewart. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Stat. Med.*, 88:4197–215, 2006.

D. Rubin and M.J. van der Laan. A doubly robust censoring unbiased trans- formation. *The International Journal of Biostatistics*, Vol. 3 (1): http://www.bepress.com/ijb/vol3/iss1/4, 2007.

D.B. Rubin. *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 2006.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 64:688–701, 1974.

D.B. Rubin and M.J. van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *International Journal of Biostatistics*, 8:Issue 1, Article 18, 2012.

D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, 94(448):1096–1120 (1121–1146), 1999.

S. Shortreed and E. Moodie. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential-multiple assignment randomized catie schizophrenia study. *J.R. Stat. Soc. C.*, 61:577–99, 2012.

R. Sutton and H. Sung. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

P. Thall, R. Millikan, and H. Sung. Evaluating multiple treatment courses in clinical trials. *Stat. Med.*, 30:1011–128, 2000.

P. Thall, H. Sung, and E. Estey. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Am. Stat. Assoc.*, 39: 29–39, 2002.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan. Statistical inference when using data adaptive estimators of nuisance parameters. Technical Report 302, Division of Biostatistics, University of California, Berkeley, submitted to IJB, 2012.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.

M.J. van der Laan and S. Gruber. Targeted minimum loss based estimation of an intervention specific mean. *The International Journal of Biostatistics*, in press, 2012.

M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3, Issue 1, Article 3, 2007.

M.J. van der Laan and M.L. Petersen. Targeted learning. In *Ensemble Machine Learning*, chapter pages 117–156, ISBN 978-1-4419-9326-7. Springer, New York, 2012.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2003.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2012.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

M.J. van der Laan, A.E. Hubbard, and S. Kherad. Statistical inference for data adaptive target parameters. Technical Report 314, Division of Biostatistics, University of California, Berkeley, 2013.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

E. Wagner, B. Austin, C. Davis, M. Hindmarsh, J. Schaefer, and A. Bonomi. Improving chronic illness care: translating evidence into action. *Health Aff.*, 20:64–78, 2001.

Z. Yu and M.J. van der Laan. Measuring treatment effects using semiparametric models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003.

Y. Zhao, D. Zeng, A. Rush, and M. Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67:1422–33, 2011.

Y. Zhao, D. Zeng, A. Rush, and M. Kosorok. Estimating individual treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, 107:1106–18, 2012.

W. Zheng and M.J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. Technical Report 273, Division of Biostatistics, University of California, Berkeley, 2010.

W. Zheng and M.J. van der Laan. Cross-validated targeted minimum loss based estimation. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*. Springer, New York, 2012.

Appendix

A Formulation of optimal dynamic treatment estimation problem: multiple time point treatment

Suppose we observe n i.i.d. copies $O_1, \ldots, O_n$ of

$$O = (L(0), A(0), L(1), A(1), \ldots, L(K), A(K), Y = L(K+1)) \sim P_0,$$

where $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ is a binary treatment and $A_2(j)$ is a missing or right-censoring indicator at "time" j, $j = 0, 1 \ldots, K$. For a time-dependent process $X()$, we will use the notation $\bar{X}(t) = (X(s) : s \leq t)$. Let $\mathcal{M}$ be a statistical model that makes no assumptions on the marginal distribution $Q_{0,L(0)}$ of $L(0)$, and the conditional distributions $Q_{0,L(j)}$ of $L(j)$, given $\bar{A}(j-1), \bar{L}(j-1)$, $j = 1, \ldots, K+1$, but might make assumptions on the conditional distributions $g_{0,A(j)}$ of $A(j)$, given $\bar{A}(j-1), \bar{L}(j)$, $j = 0, \ldots, K$. We refer to g_0 as the censoring and treatment mechanism or simply intervention mechanism. We note that we can factorize g_0 as follows in a treatment mechanism $g_{0A(1)}$ and censoring mechanism $g_{0A(2)}$:

$$\begin{aligned} g_0(O) &= \prod_{j=0}^{K} g_{0,A_1(j)}(A_1(j) \mid \bar{L}(j), \bar{A}(j-1)) \\ &\quad \prod_{j=0}^{K} g_{0,A_2(j)}(A_2(j) \mid \bar{L}(j), \bar{A}_1(j), \bar{A}_2(j-1)) \\ &\equiv g_{0A(1)}(O) g_{0A(2)}(O). \end{aligned}$$

The data might have been generated by a sequential multiple assignment randomized trial (SMART) in which case $g_{0A(1)}$ is known. In addition, in such a SMART it might be known that right-censoring at time j only depends on the past through the past censoring and treatment history in which case nonparametric and root-n-consistent estimators of $g_{0A(2)}$ are directly available.

Let $V(0)$ be a function of $L(0)$ and let $(\bar{A}(j-1), V(j))$ be a function of $(\bar{A}(j-1), \bar{L}(j))$, $j = 1, \ldots, K$. Let $V = \bar{V} = (V(0), V(1), \ldots, V(K))$. Consider dynamic treatment rules $V(0) \rightarrow d_{A(0)}(V(0)) \in \{0,1\} \times \{1\}$, $(\bar{A}(j-1), V(j)) \rightarrow d_{A(j)}(\bar{A}(j-1), V(j)) \in \{0,1\} \times \{1\}$, $j = 1, \ldots, K$, that are restricted to make the treatment rule only depend on $\bar{L}(j)$ through $V(j)$, and to deterministically

set the censoring indicators $A_2(j) = 1$, $j = 0, \ldots, K$. Let $\mathcal{D}$ be the set of all such dynamic treatments.

We will also assume that $V(0)$ is a function of $V(1)$ (i.e., observing $V(1)$ includes observing $V(0)$), and $V(j)$ is a function of $V(j+1)$, $j = 1, \ldots, K-1$. Under this assumption $d_{A(j)}$ is only a function of $V(j)$ since $\bar{A}(j-1)$ is itself a function of $V(j)$. Therefore, we will also use notation $d_{A(j)}(V(j))$ instead of $d_{A(j)}(\bar{A}(j-1), V(j))$. For any rule $d \in \mathcal{D}$, let

$$\Psi_d(P) \equiv E_{P_d} Y_d$$

denote the mean outcome of Y_d under dynamic treatment d, where $L_d = (L_d(0), \ldots, Y_d = L_d(K+1))$ is a random variable with probability density

$$\begin{aligned} &P_d(L(0), A(0), \ldots, L(K), A(K), Y) \\ &= I(A = d(V)) Q_{L(0)}(L(0)) \textstyle\prod_{k=1}^{K+1} Q_{L(k)}(L(k) \mid \bar{L}(k-1), \bar{A}(k-1)), \end{aligned}$$

with respect to some dominating measure μ. This probability distribution P_d is the G-computation formula for the counterfactual O_d representing the probability distribution O would have had, if contrary to the fact, A would have been assigned according to the dynamic intervention $d = (d_{A(0)}, \ldots, d_{A(K)}) \in \mathcal{D}$. Thus,

$$E_{P_d} Y_d = \int_y y P_d(y) d\mu(y),$$

where

$$P_d(y) = \sum_{\bar{l}(K)} P_d(l(0), d_{A(0)}(v(0)), \ldots, l(K), d_{A(K)}(v(K)), y)$$

is the marginal density of Y_d under the joint distribution P_d of O_d. We are concerned with estimation of the V-optimal rule defined as

$$d_0 = \arg\max_{d \in \mathcal{D}} E_{P_{0,d}} Y_d.$$

That is, d_0 is the rule that maximizes the mean outcome under rule d over all treatment rules $d \in \mathcal{D}$. We are also concerned with statistical inference for the statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ defined by

$$\Psi(P_0) = E_{P_{0,d_0}} Y_{d_0} = \Psi_{d_0}(P_0).$$

These two estimation problems define the statistical estimation problem addressed in this appendix.

If we assume a structural equation model stating that $L(0) = f_{L(0)}(U_{L(0)})$, $A(k) = f_{A(k)}(\bar{L}(k), \bar{A}(k-1), U_{A(k)})$, $L(k+1) = f_{L(k+1)}(\bar{L}(k), \bar{A}(k), U_{L(k+1)})$, $k = 0, \ldots, K$, we can define counterfactuals Y_d defined by the modified system in which the equations for $A(k)$ are replaced by $A(k) = d_{A(k)}(V(k))$, $k = 0, \ldots, K$. One can now define the causally optimal rule as $d_0^* = \arg\max_{d \in \mathcal{D}} E_{P_0} Y_d$. If we assume a sequential randomization assumption stating that $A(k)$ is independent of $(U_{L(j)} : j = k+1, \ldots, K+1)$, given $\bar{L}(k), \bar{A}(k-1)$, then we have that $E_0 Y_d = E_{P_{0,d}} Y_d$ defined above, for all rules d, and thereby that the statistical rule d_0 defined above equals this causally optimal rule d_0^*. In this case, $E_0 Y_{d_0^*} = \Psi(P_0)$. Similarly, we have such an identifiability result under the Neyman-Rubin causal model (Robins (1987a)).

In the remainder of the article, if for a static or dynamic intervention d, we use notation L_d (or Y_d, O_d) we mean the random variable with probability distribution P_d, so that all our quantities are statistical parameters. For example, the quantity $E_{P_0}(Y_{\bar{a}(K)} \mid V_{\bar{a}(K-1)}(K))$ defined in the next theorem denotes the conditional expectation of $Y_{\bar{a}(K)}$, given $V_{\bar{a}(K-1)}(K)$, under the probability distribution $P_{0,\bar{a}(K)}$ (i.e., G-computation formula presented above for the static intervention $\bar{a}(K)$). In addition, if we write down these parameters, we will automatically assume the positivity assumption required for the G-computation formula to be well defined. For that it will suffice to assume

$$P_0\left(0 < \min_{\delta \in \{0,1\}} g_{0,A(k)}(\delta, 1 \mid \bar{L}(k), \bar{A}(k-1))\right) = 1, \ k = 0, \ldots, K. \tag{11}$$

B Characterization of V-optimal rule in terms of blip-functions

The next theorem presents an explicit functional form of the V-optimal individualized treatment rule d_0 as a function of P_0. We will use notation $d_{0,A(j:k)}$ for $(d_{0,A(l)} : l = j, \ldots, k)$ and for a process X we use $X(k:l) = (X(s) : k \leq s \leq l)$.

Theorem 13 *We assumed $V(k)$ is a function of $V(k+1)$, $k = 0, \ldots, K-1$. The V-optimal rule d_0 can be represented as the following explicit parameter*

of P_0:

$$
\begin{aligned}
&\bar{Q}_{K+10}(\bar{a}(K-1), v(K)) = \\
&E_{P_0}(Y_{\bar{a}(K-1),A(K)=(1,1)} \mid V_{\bar{a}(K-1)}(K) = v(K)) \\
&\qquad\qquad -E_{P_0}(Y_{\bar{a}(K-1),A(K)=(0,1)} \mid V_{\bar{a}(K-1)}(K) = v(K)) \\
&d_{0,A(K)}(\bar{A}(K-1), V(K)) = (I(\bar{Q}_{K+10}(\bar{A}(K-1), V(K)) > 0), 1) \\
&\bar{Q}_{K0}(\bar{a}(K-2), v(K-1)) = \\
&E_{P_0}(Y_{\bar{a}(K-2),A(K-1)=(1,1),d_{0,A(K)}} \mid V_{\bar{a}(K-2)}(K-1) = v(K-1)) \\
&\qquad -E_{P_0}(Y_{\bar{a}(K-2),A(K-1)=(0,1),d_{0,A(K)}} \mid V_{\bar{a}(K-2)}(K-1) = v(K-1)) \\
&d_{0,A(K-1)}(\bar{A}(K-2), V(K-1)) = (I(\bar{Q}_{K0}(\bar{A}(K-2), V(K-1)) > 0), 1) \\
&\bar{Q}_{k+10}(\bar{a}(k-1), v(k)) = \\
&E_{P_0}(Y_{\bar{a}(k-1),A(k)=(1,1),d_{0,A(k+1:K)}} \mid V_{\bar{a}(k-1)}(k) = v(k)) \\
&\qquad\qquad -E_{P_0}(Y_{\bar{a}(k-1),A(k)=(0,1),d_{0,A(k+1:K)}} \mid V_{\bar{a}(k-1)}(k) = v(k)) \\
&d_{0,A(k)}(\bar{A}(k-1), V(k)) = (I(\bar{Q}_{k+10}(\bar{A}(k-1), V(k)) > 0), 1) \\
&\ k = K, \ldots, 0.
\end{aligned}
$$

Recall that $a(k) \in \{0,1\} \times \{1\}$ for all $k = 0, \ldots, K$. If $V(k)$ does not include $V(k-1)$, but, for all $\bar{a}(K) \in \{\{0,1\} \times \{1\}\}^K$,

$$E(Y_{\bar{a}(K)} \mid V(0), \ldots, V_{\bar{a}}(K)) = E(Y_{\bar{a}(K)} \mid V_{\bar{a}}(K)), \tag{12}$$

then the above expression for the V-optimal rule d_0 is still true.

The proof is analogue to the proof of Theorem 5.

C Sequential super learning of V-optimal rule

Estimation of d_0 requires estimation of $\bar{Q}_{K+1,0}$, which then yields an estimator $d_{n,A(K)}$ of $d_{0,A(K)}$, and, subsequently, we need to estimate $\bar{Q}^{d_n}_{K,0}$ treating $d_{n,A(K)}$ as given. This process is iterated: given $d_{n,A(j:K)}$, we estimate $\bar{Q}^{d_n}_{j,0}$, $j = K, \ldots, 1$. Finally, the estimator of $\bar{Q}^{d_n}_{1,0}$ maps into an estimator of $d_{0,A(0)}$. We refer to such an estimation procedure as a sequential estimation procedure and our estimator of d_0 will follow this approach.

As a consequence, the estimation problem that needs to be addressed is given by: for a given $k \in \{1, K+1\}$, the estimation of $\bar{Q}^d_{k,0}$ for a given dynamic rule d, where $\bar{Q}^d_k$ only depends on d through $d_{A(k:K)}$. For that purpose we use the super-learning framework. This relies on the specification of a risk function $R_{\bar{Q}^d_k}(P_0)$ which uniquely characterizes the true parameter $\bar{Q}^d_{k,0}$ as the minimizer: $\bar{Q}^d_{k,0} = \arg\min_{\bar{Q}^d_k} R_{\bar{Q}^d_k}(P_0)$. In some cases we use a representation $R_{\bar{Q}^d_k}(P_0) = E_{P_0} L_{Q_0,g_0}(\bar{Q}^d_k)$ for a specified loss function L_{Q_0,g_0} indexed by

nuisance parameters Q_0, g_0. In addition, we need to construct a library of estimators $\hat{\bar{Q}}^d_{k,j}$ of $\bar{Q}^d_{k,0}$, $j = 1, \ldots, J$. This generates a family of candidate estimators $\hat{\bar{Q}}^d_{k,\alpha} = \sum_j \alpha_j \hat{\bar{Q}}^d_{k,j}$ obtained by taking linear combinations of these estimators using a weight-vector α, but the user can decide on the kind of parametric family to combine the library of estimators. We now also need a cross-validated estimator $R_{\hat{\bar{Q}}^d_\alpha,n}$ of $1/B \sum_{b=1}^B R_{\bar{Q}^d_{\alpha,n,b}}(P_0)$ in order to select among the candidate estimators $\hat{\bar{Q}}_{k,\alpha}$. Here b indicates a sample split in a training sample T_b and validation sample V_b, $P^1_{n,b}, P^0_{n,b}$ are the empirical distribution functions of the validation and training sample, respectively, and $\bar{Q}_{\alpha,n,b} = \hat{\bar{Q}}_\alpha(P^0_{n,b})$ is the estimate based on the training sample. We can now define the cross-validation selector

$$\alpha_n = \arg\min_\alpha R_{\hat{\bar{Q}}^d_{k,\alpha},n}.$$

It can be decided to restrict α to be a vector of positive numbers and sum up till 1. The proposed super-learner of $\bar{Q}^d_{k,0}$ is defined as $\hat{\bar{Q}}^d_{k,\alpha_n}(P_n)$, or, one could, define it as $\frac{1}{B}\sum_{b=1}^B \hat{\bar{Q}}^d_{k,\alpha_n}(P^0_{n,b})$. For example, if the risk function allows the loss-function representation, then we have

$$R_{\hat{\bar{Q}}^d_\alpha,n} = \frac{1}{B}\sum_{b=1}^B P^1_{n,b} L_{Q_{n,b},g_{n,b}}(\hat{\bar{Q}}^d_{k,\alpha}(P^0_{n,b})).$$

As we will see in the case that g_0 is known it will be possible to select an IPCW-loss function $L_{g_0}(\bar{Q}^d_k)$ that does not depend on unknown nuisance parameters, and this IPCW-loss function is just a weighted squared error or weighted log-likelihood loss so that the cross-validated risk estimator is just a weighted cross-validated sum of squared residuals or weighted cross-validated log-likelihood.

In the following sections we will propose risk functions $R_{\bar{Q}^d_k}(P_0)$, determine the efficient influence curve of these risk functions, and propose IPCW, DR-IPCW, and (double robust) TMLE of this risk function, and its corresponding cross-validated versions $CV - IPCW$, $CV - DR - IPCW$, and CV-TMLE. The CV-IPCW and CV-DR-IPCW estimators are defined as cross-validated empirical means of an IPCW and DR-IPCW loss function, respectively. The CV-IPCW estimators rely on a consistent estimator of g_0, which is certainly appropriate for SMART, while the CV-DR-IPCW and CV-TMLE are double robust in the sense that these estimators are consistent if either g_0 or a part of Q_0 is consistently estimated. The latter two estimators are also asymptotically efficient estimators of the risk if both nuisance parameters are consistently estimated. The CV-DR-IPCW can become an unstable estimator in finite

samples, while the CV-TMLE is a substitution estimator respecting the global constraints of the model, resulting in potentially important improvements in practical performance. In addition, we discuss how to generate a library of candidate estimators, and we will see that IPCW-loss function can be utilized for that purpose, allowing the incorporation of standard software, and we develop TMLE of projections on parametric working models as alternative candidate estimators to be included in the library of estimators.

D Risk function for V-optimal rule

The following theorem presents a squared error risk function of $\bar{Q}^d_{k,0}$.

Theorem 14 *Define*

$$\begin{aligned}&D_1(Q,g)(O_{\bar{a}(k-1)\underline{d}_{k+1}}) = I(A_2(k)=1)\tfrac{2A_1(k)-1}{g_{A(k)}(O)}\{Y_{\bar{a}(k-1)\underline{d}_{k+1}}\\ &-E_Q(Y_{\bar{a}(k-1)\underline{d}_{k+1}} \mid \bar{L}(k), \bar{A}(k-1)=\bar{a}(k-1)), A(k)\}\\ &+E(Y_{\bar{a}(k-1)\underline{d}_{k+1}} \mid \bar{L}(k), \bar{A}(k-1)=\bar{a}(k-1), A(k)=(1,1))\\ &-E(Y_{\bar{a}(k-1)\underline{d}_{k+1}} \mid \bar{L}(k), \bar{A}(k-1)=\bar{a}(k-1), A(k)=(0,1)).\end{aligned}$$

We have

$$E_{P_0}(D_1(Q,g)(O_{\bar{a}(k-1)\underline{d}_{k+1}} \mid V_{\bar{a}}(k)) = \bar{Q}^d_{0,k+1}(\bar{a}(k-1), V_{\bar{a}}(k)),$$

if either $D_1(Q,g) = D_1(Q_0,g)$ *or* $D_1(Q,g) = D_1(Q,g_0)$. *Define*

$$L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1}) = (D_1(Q,g) - \bar{Q}^d_{k+1})^2,$$

and

$$\begin{aligned}&R_{\bar{Q}^d_{k+1}}(D_1(Q,g),P_0) =\\ &\textstyle\sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1}(\bar{a}(k-1), V_{\bar{a}}(k)) L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1})(O_{\bar{a}(k-1)\underline{d}_{k+1}}).\end{aligned}$$

If either $D_1(Q,g) = D_1(Q_0,g)$ *or* $D_1(Q,g) = D_1(Q,g_0)$, *then*

$$\begin{aligned}&R_{\bar{Q}^d_{k+1}}(D_1(Q,g),P_0)\\ &= \textstyle\sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1}\left(D_1(Q,g)^2 - 2\bar{Q}^d_{0,k+1}\bar{Q}^d_{k+1} + \bar{Q}^{d2}_{k+1}\right)\\ &= \textstyle\sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1}\{\bar{Q}^d_{0,k+1} - \bar{Q}^d_{k+1}\}^2\\ &\qquad + \textstyle\sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1}\{D_1(Q,g)^2 - \bar{Q}^{d2}_{0,k+1}\}.\end{aligned}$$

Therefore, if either $D_1(Q,g) = D_1(Q_0,g)$ *or* $D_1(Q,g) = D_1(Q,g_0)$, *then*

$$\bar{Q}^d_{0,k+1} = \arg\min_{\bar{Q}^d_{k+1}} R_{\bar{Q}^d_{k+1}}(D_1(Q,g),P_0).$$

Define

$$Z_{\bar{a}(k-1)\underline{d}_{k+1}} \equiv h_{k+1}\left(D_1(Q,g) - \bar{Q}^d_{k+1}\right)^2 (O_{\bar{a}(k-1)\underline{d}_{k+1}}).$$

Then, this squared error risk can be represented as:

$$R_{\bar{Q}^d_{k+1}}(D_1(Q,g), P_0) = \sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)\underline{d}_{k+1}}.$$

The following theorem presents a log-likelihood risk function of $\bar{Q}^d_{k,0}$.

Theorem 15 *Suppose $\bar{Q}^d_{k+1,0} \in (a,b)$ for a known $a < b$. Define $D_1^{a,b}(Q,g)(Q,g) = (D_1(Q,g)-a)/(b-a)$. If either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then*

$$E_{P_0}(D_1^{a,b}(Q,g)(O_{\bar{a}(k-1)\underline{d}_{k+1}}) \mid V_{\bar{a}}(k)) = \bar{Q}^{a,b,d}_{0,k+1},$$

where $\bar{Q}^{a,b,d}_{0,k+1} = (\bar{Q}^d_{0,k+1}-a)/(b-a) \in (0,1)$. For notational convenience, in our presentation below $\bar{Q}^d_{0,k+1}, D_1(Q,g)$ are already standardized so that $\bar{Q}^d_{0,k+1} \in (0,1)$ and $E_{P_0}(D_1(Q,g)(O_{\bar{a}(k-1)\underline{d}_{k+1}}) \mid V_{\bar{a}}(k)) = \bar{Q}^d_{0,k+1}$ if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$.

Define

$$-L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1}) = D_1(Q,g)\log \bar{Q}^d_{k+1} + (1 - D_1(Q,g))\log(1-\bar{Q}^d_{k+1}).$$

Define

$$R_{\bar{Q}^d_{k+1}}(D_1(Q,g), P_0) = \sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1} L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1})(O_{\bar{a}(k-1),\underline{d}_{k+1}}).$$

If either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then

$$R_{\bar{Q}^d_{k+1}}(D_1(Q,g), P_0) \;=\; \sum_{\bar{a}(k-1)} E_{P_{0,\bar{a}(k-1)\underline{d}_{k+1}}} h_{k+1} L^F_{\bar{Q}^d_{0,k+1}}(\bar{Q}^d_{k+1})(O_{\bar{a}(k-1),\underline{d}_{k+1}}).$$

Therefore, if either $D_1(Q,g) = D_1(Q_0,g)$ or $D_1(Q,g) = D_1(Q,g_0)$, then

$$\bar{Q}^d_{0,k+1} = \arg\min_{\bar{Q}^d_{k+1}} R_{\bar{Q}^d_{k+1}}(D_1(Q,g), P_0).$$

Define

$$Z_{\bar{a}(k-1)\underline{d}_{k+1}} \equiv h_{k+1} L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1}).$$

Then, this quasi-log-likelihood risk can be represented as:

$$R_{\bar{Q}^d_{k+1}}(D_1(Q,g), P_0) = \sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)\underline{d}_{k+1}}.$$

E Sequential regression representation of risk function treating D_1 as given

If we treat $D_1 = D_1(Q, g)$ as given, then the risk parameter can be represented as the following statistical parameter of P_0:

$$R_{\bar{Q}^d_{k+1}}(D_1, P_0) = \sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)\underline{d}_{k+1}}.$$

In other words, $R_{\bar{Q}^d_{k+1}}(D_1, P_0)$ is an average over $\bar{a}(k-1)$ of the expectation of a counterfactual outcome of $Z = h_{k+1} L^F_{D_1}(\bar{Q}^d_{k+1})$ under intervention $(\bar{a}(k-1)\underline{d}_{k+1})$. Our proposed estimators of risk are two stage estimators in the sense that we first estimate $D_1(Q_0, g_0)$ and given this estimator D_{1n}, we estimate $R_{\bar{Q}^d_{k+1}}(D_{1n}, P_0)$ with estimators developed for the latter parameter treating D_{1n} as given. In a later section, we will also define the risk as a parameter $P_0 \rightarrow R_{\bar{Q}^d_{k+1}}(D_1(Q(P_0), g(P_0)), P_0)$ and develop estimators directly targeting this parameter (up till a term constant in $\bar{Q}^d_{k+1}$). In the sequel we will use the notation $d_{i:j}$ to indicate the rules $\{d_{A(l)} : l = i, \ldots, j\}$ and similarly $A(k : K) = (A(l) : l = k, \ldots, K)$. We also use the short-hand notation $\underline{d}_k = (d_l : l = k, \ldots, K)$. Finally, if we write $I(\bar{A}(k) = d_{1:k})$, then that represents short-hand notation for the indicator that $\bar{A}(k)$ equals the values assigned by the rule d as function of $\bar{L}(k)$.

Theorem 16 *Let $d = d(\bar{a}(k-1)) = (\bar{a}(k-1) g_{A(k)} \underline{d}_{k+1})$ be this particular intervention on $(A(0) \ldots, A(K))$ which leaves $A(k)$ random as under P, and note that this intervention is indexed by a choice $\bar{a}(k-1)$. Define $Z = h_{k+1}(\bar{A}(k-1), V(k)) L^F_{D_1}(\bar{Q}^d_{k+1})(O)$. Let*

$$\begin{aligned}
\bar{Q}^{\bar{a}(k-1)}_{Z,K+1} &= E_P(Z \mid \bar{A}(K-1), A(K) = d_K, \bar{L}(K)) \\
\bar{Q}^{\bar{a}(k-1)}_{Z,K} &= E_P(\bar{Q}^{\bar{a}(k-1)}_{Z,K+1} \mid \bar{A}(K-2), A(K-1) = d_{K-1}, \bar{L}(K-1)) \\
\bar{Q}^{\bar{a}(k-1)}_{Z,j} &= E_P(\bar{Q}^{\bar{a}(k-1)}_{Z,j+1} \mid \bar{A}(j-2), A(j-1) = d_{j-1}, \bar{L}(j-1)) \\
& \quad j = K, \ldots, 1. \\
\bar{Q}^{\bar{a}(k-1)}_{Z,0} &= E_{L(0)} \bar{Q}^{\bar{a}(k-1)}_{Z,1}(L(0)).
\end{aligned}$$

Note that for $j - 1 = k$,

$$E_P(\bar{Q}^{\bar{a}(k-1)}_{Z,k+2} \mid \bar{A}(k-1), A(k) = d_k, \bar{L}(k)) = E_P(\bar{Q}^{\bar{a}(k-1)}_{Z,k+2} \mid \bar{A}(k-1), A(k), \bar{L}(k)).$$

We have

$$R_{\bar{Q}^d_{k+1}}(D_1, P) = \sum_{\bar{a}(k-1)} \bar{Q}^{\bar{a}(k-1)}_{Z,j=0}.$$

This follows from the fact that $\bar{Q}_{Z,j}^{\bar{a}(k-1)} = E(Z_d \mid \bar{A}(j-2), \bar{L}(j-1))$, and, thus $\bar{Q}_{Z,1}^{\bar{a}(k-1)} = E(Z_d \mid L(0))$, and $\bar{Q}_{Z,0}^{\bar{a}(k-1)} = E_P Z_d$.

This provides us with a representation of the risk-parameter $R_{\bar{Q}_{k+1}^d}(D_1, P) = R_{\bar{Q}_{k+1}^d}(D_1, \bar{Q}_Z)$ as a function of P through $\bar{Q}_Z(P) = (\bar{Q}_{Z,j}^{d(\bar{a}(k-1))} : j = 0, \ldots, K+1, \bar{a}(k-1))$. The TMLE will be a plug-in estimator obtained with a targeted estimator $\bar{Q}_{Z,n}^*$ of $\bar{Q}_{Z,0}$.

F Efficient influence curve of risk function

In order to develop an efficient estimator of $R_{\bar{Q}_{k+1}^d}(D_1, P_0)$ for a given D_1, we need to know the efficient influence function. The following theorem presents this efficient influence function.

Theorem 17 *Let g^* be the intervention mechanism defined by $d = (\bar{a}(k-1) g_{A(k)} \underline{d}_{k+1})$. The efficient influence curve of $R_{\bar{Q}_{k+1}^d}(D_1, P_0)$ is given by $D^*_{\bar{Q}_{k+1}^d}(D_1, P_0) = \sum_{j=0}^{K+1} D^*_{\bar{Q}_{k+1}^d, j}(D_1, P_0)$, where*

$$\begin{aligned}
D^*_{\bar{Q}_{k+1}^d, K+1}(D_1, P_0) &= \sum_{\bar{a}(k-1)} \frac{g^*_{0:K}}{g_{0:K}(O)} (Z - \bar{Q}^d_{Z,K+1}) \\
D^*_{\bar{Q}_{k+1}^d, j}(D_1, P_0) &= \sum_{\bar{a}(k-1)} \frac{g^*_{0:j-1}(O)}{g_{0:j-1}(O)} (\bar{Q}^d_{Z,j+1} - \bar{Q}^d_{Z,j}) \\
j &= K+1, \ldots, 1 \\
D^*_{\bar{Q}_{k+1}^d, 0}(D_1, P_0) &= \sum_{\bar{a}(k-1)} \{\bar{Q}^d_{Z,1} - \bar{Q}^d_{Z,0}\}.
\end{aligned}$$

Note that for $j - 1 < k$, $g^*_{0:j-1}/g_{0:j-1} = I(\bar{A}(j-1) = d_{0:j-1})/g_{0:j-1}(O)$, and for $j - 1 \geq k$, $g^*_{0:j-1}/g_{0:j-1} = I(\bar{A}(k-1) = \bar{a}(k-1), A(k+1:K) = d_{k+1:K})/\prod_{l \neq k, l=0}^{j-1} g_{A(j)}$.

For $j \geq k$, we have

$$D^*_{\bar{Q}_{k+1}^d, j}(D_1, P_0) = \frac{g^*_{k:j-1}(O))}{g_{0:j-1}(O)} (\bar{Q}^d_{Z,j+1} - \bar{Q}^d_{Z,j}).$$

For $j < k$,

$$D^*_{\bar{Q}_{k+1}^d, j}(D_1, P_0) = \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)} (\bar{Q}^d_{Z,j+1} - \bar{Q}^d_{Z,j}).$$

G CV-IPCW and CV-DR-IPCW estimators of risk of candidate estimator.

In this section we define an IPCW and DR-IPCW estimator of $R_{\bar{Q}^d_{k+1}}(D_1, P_0)$, and, subsequently, we define their cross-validated counterparts for a candidate estimator $\hat{\bar{Q}}^d_{k+1}$.

G.1 IPCW-estimator of risk

Recall the represntation $R_{\bar{Q}^d_{k+1}}(P_0) = \sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)\underline{d}_{k+1}}$, where

$$Z = h_{k+1}(\bar{A}(k-1), V(k))(D_1(Q,g)(O) - \bar{Q}^d_{k+1})^2(\bar{A}(k-1), V(k)).$$

Let $g_{-k} = \prod_{j=0, j\neq k} g_{A(j)}$. An IPCW-estimator is given by:

$$\begin{aligned} R_{\bar{Q}^d_{k+1}, IPCW, n} &= \frac{1}{n}\sum_{i=1}^n \sum_{\bar{a}(k-1)} \frac{I(\bar{A}_i(k-1) = \bar{a}(k-1), A_i(k+1:K) = \underline{d}_{k+1})}{g_{n,-k}(O_i)} Z_i \\ &= \frac{1}{n}\sum_{i=1}^n \frac{I(A_i(k+1:K) = \underline{d}_{k+1})}{g_{n,-k}(O_i)} Z_i. \end{aligned}$$

Notice that this estimator is nothing else than an IPCW-empirical mean of squared errors. One can also use the stabilized IPCW-estimator $R_{\bar{Q}^d_{k+1}, SIPCW, n}$ obtained by dividing the above IPCW-estimator $R_{\bar{Q}^d_{k+1}, IPCW, n}$ by $\frac{1}{n}\sum_{i=1}^n \frac{I(A_i(k+1:K)=\underline{d}_{k+1})}{g_{n,-k}(O_i)}$.

G.2 IPCW-loss function, and cross-validated IPCW-estimator of risk of candidate estimator

Define

$$L_{IPCW, D_1, g})\bar{Q}^d_{k+1})(O) \equiv \frac{I(A(k+1:K) = \underline{d}_{k+1})}{g_{-k}(O)} Z(D_1, \bar{Q}^d_{k+1}). \qquad (13)$$

We refer to this as the IPCW-loss function indexed by nuisance parameters $D_1(Q,g), g$. The above IPCW-estimator can be represented as

$$R_{\bar{Q}^d_{k+1}, IPCW, n} = P_n L_{IPCW, D_{1n}, g_n}(\bar{Q}^d_{k+1}).$$

The cross-validated IPCW-estimator for a given estimator $\hat{\bar{Q}}^d_{k+1}$ is given by

$$R_{\hat{\bar{Q}}^d_{k+1}, CV-IPCW, n} = \frac{1}{B}\sum_{b=1}^B P^1_{n,b} L_{IPCW, D_{1,n,b}, g_{n,b}}(\bar{Q}^d_{k+1,n,b}).$$

If $g_n = g_0$, then, under very weak regularity conditions, we have that $R_{\hat{\bar{Q}}^d_{k+1}, CV-IPCW,n} - \frac{1}{B}\sum_{b=1}^{B} P_0 L_{IPCW,D_{10},g_0}(\bar{Q}^d_{k+1,n,b})$ is asymptotically linear with influence curve $L_{IPCW,D_{10},g_0}(\bar{Q}^d_{k+1}) - P_0 L_{IPCW,D_{10},g_0}(\bar{Q}^d_{k+1})$, where $D_{10} = D_1(Q, g_0)$, and thereby converges at $1/\sqrt{n}$-rate to a normal distribution. Note that $\frac{1}{B}\sum_{b=1}^{B} P_0 L_{IPCW,D_{10},g_0}(\bar{Q}^d_{k+1,n,b}) = \frac{1}{B}\sum_{b=1}^{B} R_{\bar{Q}^d_{k+1,n,b}}(D_1(Q, g_0), P_0)$, so that this is indeed the desired result.

G.3 DR-IPCW loss-function, and estimator of risk.

Consider the efficient influence curve $D^*_{\bar{Q}^d_{k+1}}(D_1, P_0)$ of $R_{\bar{Q}^d_{k+1}}(D_1, P_0)$ defined in Theorem 17, and note that it can be represented as $L_{D_1,Q_0,g_0}(\bar{Q}^d_{k+1}) - R_{\bar{Q}^d_{k+1}}(D_1, P_0)$. We refer to $L_{D_1,Q,g}(\bar{Q}^d_{k+1})$ as the DR-IPCW loss function indexed by nuisance parameters (D_1, Q, g). Therefore, an estimating equation based estimator based on solving the efficient influence curve estimating equation in the target parameter is given by:

$$R_{\bar{Q}^d_{k+1}, DR-IPCW,n} = P_n L_{D_{1n},Q_n,g_n}(\bar{Q}^d_{k+1}).$$

G.4 CV-DR-IPCW estimator of risk

The CV-DR-IPCW for a given estimator $\hat{\bar{Q}}^d_{k+1}$ is defined as:

$$R_{\hat{\bar{Q}}^d_{k+1}, CV-EE,n} = \frac{1}{B}\sum_{b=1}^{B} P^1_{n,b} L_{D_{1,n,b},Q_{n,b},g_{n,b}}(\bar{Q}^d_{k+1,n,b}).$$

H CV-TMLE of risk of candidate estimator.

H.1 TMLE of risk.

Define the intervention $d = d(\bar{a}(k-1)) = (\bar{a}(k-1), g_{A(k)}, \underline{d}_{k+1})$ and let $g^* = g^*(\bar{a}(k-1)$ be the corresponding stochastic intervention on $\bar{A}(K)$. Suppose that Z is standardized so that $E(Z \mid \bar{L}(K), \bar{A}(K)) \in (0, 1)$. Let $Z = Z_n$ be an estimator obtained by plugging in an estimator D_{1n} of $D_1(Q_0, g_0)$. Let $\bar{Q}^d_{Z,K+2} = Z$. Firstly, fit a logistic regression of $\bar{Q}^d_{Z,K+2}$ on $\bar{A}(K)$ and $\bar{L}(K)$, and set $A(K) = d_K$. This is an initial estimator $\bar{Q}^d_{Z,K+1,n}$ of $\bar{Q}^d_{Z,K+1} = E(Z_d \mid \bar{A}(K-1), \bar{L}(K))$. Let $\text{Logit}\bar{Q}^d_{Z,K+1,n}(\epsilon) = \text{Logit}\bar{Q}^d_{Z,K+1,n} + \epsilon C_{K+1}(g_n)$, where

$$C_{K+1}(g) = \frac{g^*_{k:K}(O))}{g_{0:K}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} P_n L(\bar{Q}^d_{Z,K+1,n}(\epsilon)),$$

where

$$-L(\bar{Q}^d_{Z,K+1}) = \\ I(A(K) = d_K) \left\{ \bar{Q}^d_{Z,K+2,} \log \bar{Q}^d_{Z,K+1} + (1 - \bar{Q}^d_{Z,K+2}) \log(1 - \bar{Q}^d_{Z,K+1}) \right\}.$$

The update $\bar{Q}^*_{Z,K+1,n} = \bar{Q}_{Z,K+1,n}(\epsilon_n)$ is the targeted estimator of $\bar{Q}^d_{Z,K+1}$.

For $j = K, \ldots, k+1$, fit a logistic regression of $\bar{Q}^{d*}_{Z,j+1,n}$ on $\bar{A}(j-1), \bar{L}(j-1)$ and set $A(j-1) = d_{j-1}$. This yields an initial estimator $\bar{Q}^d_{Z,j,n}$ of $\bar{Q}^d_{Z,j} = E(Z_d \mid \bar{A}(j-2), \bar{L}(j-1))$. Let $\text{Logit}\bar{Q}^d_{Z,j,n}(\epsilon) = \text{Logit}\bar{Q}^d_{Z,j,n} + \epsilon C_j(g_n)$, where

$$C_j(g) = \frac{g^*_{k:j-1}}{g_{0:j-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} P_n L_{\bar{Q}^{d*}_{Z,j+1,n}}(\bar{Q}^d_{Z,j,n}(\epsilon)),$$

where

$$-L_{\bar{Q}^{d*}_{Z,j+1,n}}(\bar{Q}^d_{Z,j,n}) = \\ I(A(j-1) = d_{j-1}) \left\{ \bar{Q}^{d*}_{Z,j+1,n} \log \bar{Q}^d_{Z,j,n} + (1 - \bar{Q}^{d*}_{Z,j+1,n} \log(1 - \bar{Q}^d_{Z,j,n}) \right\}.$$

Define $\bar{Q}^*_{Z,j,n} = \bar{Q}_{Z,j,n}(\epsilon_n)$. This results in a targeted estimator $\bar{Q}^{d*}_{Z,k+1,n}$ of $E(Z_d \mid \bar{A}(k-1), \bar{L}(k))$. Regress $\bar{Q}^{d*}_{Z,k+1,n}$ on $\bar{A}(k-1), \bar{L}(k-1)$ and set $A(k-1) = a(k-1)$. This results in an initial estimator $\bar{Q}^{a(k-1)d}_{Z,k,n}$ of $\bar{Q}^{a(k-1)d}_{Z,k} = E(Z_{a(k-1)\underline{d}_{k+1}} \mid \bar{A}(k-2), \bar{L}(k-1))$ for each $a(k-1) \in \{0,1\}$. For each $a(k-1) \in \{0,1\}$, let $\text{Logit}\bar{Q}^{a(k-1)d}_{Z,k,n}(\epsilon) = \text{Logit}\bar{Q}^{a(k-1)d}_{Z,k,n} + \epsilon C^{a(k-1)}_k(g_n)$, where

$$C^{a(k-1)}_k(g) = \frac{1}{g_{0:k-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{a(k-1)} P_n L_{\bar{Q}^{d*}_{Z,k+1,n}}(\bar{Q}^{a(k-1)d}_{Z,k,n}(\epsilon)),$$

where

$$-L_{\bar{Q}^{d*}_{Z,k+1,n}}(\bar{Q}^{a(k-1)d}_{Z,k,n}) = \\ I(A(k-1) = a(k-1)) \left\{ \bar{Q}^{d*}_{Z,k+1,n} \log \bar{Q}^{a(k-1)d}_{Z,k,n} + (1 - \bar{Q}^{d*}_{Z,k+1,n} \log(1 - \bar{Q}^{a(k-1)d}_{Z,k,n}) \right\}.$$

This defines $\bar{Q}^{a(k-1)d*}_{Z,k,n} = \bar{Q}^{a(k-1)d}_{Z,k,n}(\epsilon_n)$ for each $a(k-1) \in \{0,1\}$.

Let $j = k-1$. Given the collection of targeted estimators $\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}$ for each $a(j:k-1) \in \{0,1\}^{k-j}$, we fit a logistic regression of $\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}$ onto $\bar{A}(j-1), \bar{L}(j-1)$ and set $A(j-1) = a(j-1)$ for each $a(j-1) \in \{0,1\}$. This yields an initial estimator $\bar{Q}^{a(j-1:k-1)d}_{Z,j,n}$ of $\bar{Q}^{a(j-1:k-1)d}_{Z,j} = E(Z_{a(j-1:k-1)\underline{d}_{k+1}} \mid \bar{A}(j-2), \bar{L}(j-1))$ for each $a(j-1:k-1)$. Given $a(j:k-1)$, for each $a(j-1) \in \{0,1\}$, let $\text{Logit}\bar{Q}^{a(j-1:k-1)d}_{Z,j,n}(\epsilon) = \text{Logit}\bar{Q}^{a(j-1:k-1)d}_{Z,j,n} + \epsilon C^{a(j-1:k-1)}_j(g_n)$, where

$$C^{a(j-1:k-1)}_j(g) = \frac{1}{g_{0:j-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{a(j-1:k-1)} P_n L_{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}}(\bar{Q}^{a(j-1:k-1)d}_{Z,j,n}(\epsilon)),$$

where

$$-L_{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}}(\bar{Q}^{a(j-1:k-1)d}_{Z,j,n}) = I(A(j-1) = a(j-1))$$
$$\left\{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n} \log \bar{Q}^{a(j-1:k-1)d}_{Z,j,n} + (1 - \bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}) \log(1 - \bar{Q}^{a(j-1:k-1)d}_{Z,j,n})\right\}$$

This defines $\bar{Q}^{a(j-1:k-1)d*}_{Z,j,n} = \bar{Q}^{a(j-1:k-1)d}_{Z,j,n}(\epsilon_n)$ for each $a(j-1:k-1) \in \{0,1\}^{k-j+1}$.

This process is iterated from $j = k-1$ to $j = 1$, giving us $\bar{Q}^{a(0:k-1)d*}_{Z,1,n}$ for each $a(0:k-1)$.

Finally, $\bar{Q}^{a(0:k-1)d*}_{Z,0,n} = \frac{1}{n}\sum_{i=1}^n \bar{Q}^{a(0:k-1)d*}_{Z,1,n}(L_i(0))$, for each $a(0:k-1)$.

Our final TMLE of $R_{\bar{Q}^d_{k+1}}(P_0)$ is given by $\sum_{\bar{a}(k-1)} \bar{Q}^{\bar{a}(k-1)d*}_{Z,0,n}$.

By construction, this TMLE solves the efficient influence curve equation $P_n D^*_{\bar{Q}^d_{k+1}}(D_{1n}, \bar{Q}^*_{Z,n}, g_n) = 0$, for the risk-parameter $\bar{Q}_Z \to R_{\bar{Q}^d_{k+1}}(D_1, \bar{Q}_Z)$, where $\bar{Q}_Z = (\bar{Q}^{d(\bar{a}(k-1))*}_{Z,j} : j = 0, \ldots, K+1, \bar{a}(k-1))$.

H.2 CV-TMLE of risk

Split the sample, and let $Q_{n,b}$ denote the estimator of Q based on the b-th training sample, $b = 1, \ldots, B$, and let $P^1_{n,b}, P^0_{n,b}$ denote the empirical distributions of the b-th validation and training sample, respectively. The following describes the CV-TMLE of $\sum_b R_{\hat{\bar{Q}}_{k+1}(P^0_{n,b})}(D_1, \bar{Q}_{Z,0})$. Define the outcome Z with the nuisance parameters fitted on the training sample:

$$Z_{n,b} = h(\bar{A}(k-1), V(k))(D_1(Q_{n,b}, g_{n,b})(O) - \bar{Q}^d_{k+1,n,b})^2(\bar{A}(k-1), V(k)).$$

Firstly, based on $P^0_{n,b}$, fit a logistic regression of $\bar{Q}^d_{Z,K+2}$ on $\bar{A}(K)$ and $\bar{L}(K)$, and set $A(K) = d_K$. This is an initial estimator $\bar{Q}^d_{Z,K+1,n,b}$ of $\bar{Q}^d_{Z,K+1} = E(Z_d \mid \bar{A}(K-1), \bar{L}(K))$. Let $\text{Logit}\bar{Q}^d_{Z,K+1,n,b}(\epsilon) = \text{Logit}\bar{Q}^d_{Z,K+1,n,b} + \epsilon C_{K+1}(g_{n,b})$, where

$$C_{K+1}(g) = \frac{g^*_{k:K}(O))}{g_{0:K}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{b=1}^{B} P^1_{n,b} L(\bar{Q}^d_{Z,K+1,n,b}(\epsilon)),$$

where

$$-L(\bar{Q}^d_{Z,K+1,n.b}) = \\ I(A(K) = d_K) \left\{ \bar{Q}^d_{Z,K+2,n,b} \log \bar{Q}^d_{Z,K+1,n,b} + (1 - \bar{Q}^d_{Z,K+2,n,b}) \log(1 - \bar{Q}^d_{Z,K+1,n,b}) \right\}.$$

The update $\bar{Q}^*_{Z,K+1,n,b} = \bar{Q}_{Z,K+1,n,b}(\epsilon_n)$ is the targeted estimator of $\bar{Q}^d_{Z,K+1}$ for each $b = 1, \ldots, B$.

For $j = K, \ldots, k+1$, based on $P^0_{n,b}$, fit a logistic regression of $\bar{Q}^{d*}_{Z,j+1,n,b}$ on $\bar{A}(j-1), \bar{L}(j-1)$ and set $A(j-1) = d_{j-1}$. This yields an initial estimator $\bar{Q}^d_{Z,j,n,b}$ of $\bar{Q}^d_{Z,j} = E(Z_d \mid \bar{A}(j-2), \bar{L}(j-1))$. Let $\text{Logit}\bar{Q}^d_{Z,j,n,b}(\epsilon) = \text{Logit}\bar{Q}^d_{Z,j,n,b} + \epsilon C_j(g_{n,b})$, where

$$C_j(g) = \frac{g^*_{k:j-1}}{g_{0:j-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{b} P^1_{n,b} L_{\bar{Q}^{d*}_{Z,j+1,n,b}}(\bar{Q}^d_{Z,j,n,b}(\epsilon)),$$

where

$$-L_{\bar{Q}^{d*}_{Z,j+1,n,b}}(\bar{Q}^d_{Z,j,n,b}) = \\ I(A(j-1) = d_{j-1}) \left\{ \bar{Q}^{d*}_{Z,j+1,n,b} \log \bar{Q}^d_{Z,j,n,b} + (1 - \bar{Q}^{d*}_{Z,j+1,n,b} \log(1 - \bar{Q}^d_{Z,j,n,b}) \right\}.$$

Define $\bar{Q}^*_{Z,j,n,b} = \bar{Q}_{Z,j,n,b}(\epsilon_n)$. This results in a targeted estimator $\bar{Q}^{d*}_{Z,k+1,n,b}$ of $E(Z_d \mid \bar{A}(k-1), \bar{L}(k))$. Regress $\bar{Q}^{d*}_{Z,k+1,n,b}$ on $\bar{A}(k-1), \bar{L}(k-1)$ and set $A(k-1) = a(k-1)$ based on $P^0_{n,b}$. This results in an initial estimator $\bar{Q}^{a(k-1)d}_{Z,k,n,b}$ of $\bar{Q}^{a(k-1)d}_{Z,k} = E(Z_{a(k-1)\underline{d}_{k+1}} \mid \bar{A}(k-2), \bar{L}(k-1))$ for each $a(k-1) \in \{0,1\}$ and b. For each $a(k-1) \in \{0,1\}$ and split b, let $\text{Logit}\bar{Q}^{a(k-1)d}_{Z,k,n,b}(\epsilon) = \text{Logit}\bar{Q}^{a(k-1)d}_{Z,k,n,b} + \epsilon C^{a(k-1)}_k(g_{n,b})$, where

$$C^{a(k-1)}_k(g) = \frac{1}{g_{0:k-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{a(k-1)} \sum_{b} P^1_{n,b} L_{\bar{Q}^{d*}_{Z,k+1,n,b}}(\bar{Q}^{a(k-1)d}_{Z,k,n,b}(\epsilon)),$$

where

$$-L_{\bar{Q}^{d*}_{Z,k+1,n,b}}(\bar{Q}^{a(k-1)d}_{Z,k,n,b}) = I(A(k-1) = a(k-1))$$
$$\left\{\bar{Q}^{d*}_{Z,k+1,n,b} \log \bar{Q}^{a(k-1)d}_{Z,k,n,b} + (1 - \bar{Q}^{d*}_{Z,k+1,n,b} \log(1 - \bar{Q}^{a(k-1)d}_{Z,k,n,b})\right\}.$$

This defines $\bar{Q}^{a(k-1)d*}_{Z,k,n,b} = \bar{Q}^{a(k-1)d}_{Z,k,n,b}(\epsilon_n)$ for each $a(k-1) \in \{0,1\}$ and b.

Let $j = k-1$. Given the collection of targeted estimators $\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n}$ for each $a(j:k-1) \in \{0,1\}^{k-j}$ and sample split b, based on $P^0_{n,b}$, we fit a logistic regression of $\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n,b}$ onto $\bar{A}(j-1), \bar{L}(j-1)$ and set $A(j-1) = a(j-1)$ for each $a(j-1) \in \{0,1\}$. This yields an initial estimator $\bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b}$ of $\bar{Q}^{a(j-1:k-1)d}_{Z,j} = E(Z_{a(j-1:k-1)\underline{d}_{k+1}} \mid \bar{A}(j-2), \bar{L}(j-1))$ for each $a(j-1:k-1)$. Given $a(j:k-1)$, for each $a(j-1) \in \{0,1\}$, let $\text{Logit}\bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b}(\epsilon) = \text{Logit}\bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b} + \epsilon C^{a(j-1:k-1)}_j(g_{n,b})$, where

$$C^{a(j-1:k-1)}_j(g) = \frac{1}{g_{0:j-1}(O)}.$$

Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_{a(j-1:k-1)} \sum_{b} P^1_{n,b} L_{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n,b}}(\bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b}(\epsilon)),$$

where

$$-L_{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n,b}}(\bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b}) = I(A(j-1) = a(j-1))$$
$$\left\{\bar{Q}^{a(j:k-1)d*}_{Z,j+1,n,b} \log \bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b} + (1 - \bar{Q}^{a(j:k-1)d*}_{Z,j+1,n,b}) \log(1 - \bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b})\right\}.$$

This defines $\bar{Q}^{a(j-1:k-1)d*}_{Z,j,n,b} = \bar{Q}^{a(j-1:k-1)d}_{Z,j,n,b}(\epsilon_n)$ for each $a(j-1:k-1) \in \{0,1\}^{k-j+1}$ and b.

This process is iterated from $j = k-1$ to $j = 1$, giving us $\bar{Q}^{a(0:k-1)d*}_{Z,1,n,b}$ for each $a(0:k-1)$ and b. Finally, $\bar{Q}^{a(0:k-1)d*}_{Z,0,n,b} = \frac{1}{n}\sum_{i=1}^n \bar{Q}^{a(0:k-1)d*}_{Z,1,n,b}(L_i(0))$, for each $a(0:k-1)$ and b. Our final CV-TMLE of $\frac{1}{B}\sum_b R_{\bar{Q}^d_{k+1}(P^0_{n,b})}(D_1, P_0)$ is given by $\frac{1}{B}\sum_b \sum_{\bar{a}(k-1)} \bar{Q}^{\bar{a}(k-1)d*}_{0,n,b}$.

I Candidate estimators for super-learner

Consider a parametric working model m_β for $\bar{Q}^d_{k+1}$. We could fit this working model by minimizing the empirical risk of the IPCW-loss function:

$$\begin{aligned}\beta_n &= \arg\min_\beta P_n L_{IPCW,D_n,g_n}(m_\beta) \\ &= \arg\min_\beta \frac{1}{n}\sum_{i=1}^n \frac{I(A_i(k+1:K)=\underline{d}_{k+1})}{g_{n,-k}(O_i)} h_{k+1}\left\{D_1(Q_n,g_n)-\bar{Q}^d_{k+1}\right\}(O_i)^2.\end{aligned}$$

This can be fitted with standard software since it is just a regression of the outcome $D_1(Q_n,g_n)(O_i)$ on $\bar{A}_i(k-1), V_i(k)$ with weights $w_i \equiv I(A_i(k+1:K)=\underline{d}_{k+1})/g_{n,-k}(O_i)h_{k+1}(\bar{A}_i(k-1),V_i(k))$, $i=1,\ldots,n$. Similarly, if we use the quasi-log-likelihood loss L^F in the definition of $Z = h_{k+1}L^F_{D_1(Q,g)}(\bar{Q}^d_{k+1})$, it follows that the working model m_β can be fitted with weighted logistic regression. Thus for each candidate parametric working model, this generates a candidate estimator of $\bar{Q}^d_{k+1}$. These candidate estimators are aiming to estimate m_{β_0} defined as the minimizer of the true squared error risk $R_{m_\beta}(D_{10},P_0)$ between m_β and $\bar{Q}^d_{0,k+1}$. The consistency of these candidate estimators as an estimator of this m_{β_0} relies on consistency of g_n.

One could also minimize the DR-IPCW empirical risk or the TMLE of the risk of m_β. In this case, the consistency of the candidate estimators as estimators of m_{β_0} relies on the consistency of either g_n or Q_n, but implementation may require some programming. However, if m_β is linear, then the minimizer β_n will still exist in closed form, and, even when m_β is not linear, if one utilizes typical iterative algorithms, then one will only need to use the estimate of the β-specific risk at limited number of candidate values. The advantage of using the TMLE estimator of the risk of m_β instead of the IPCW or DR-IPCW estimator of risk is that it results in a more robust and efficient estimator of the desired m_{β_0} defined as the minimizer of the true risk. In the next subsection, we develop an actual TMLE of β_0, and also contrast it to using a TMLE for the β-specific risk and then minimizing this risk.

In addition, consider any machine learning algorithm for fitting a regression (i.e, conditional mean) of an outcome (i.e. D_{1n}) on covariates (i.e., $\bar{A}_i(k+1:K), V_i(k)$). By simply assigning this algorithm weights w_i, it results in a candidate estimator of $\bar{Q}^d_{k+1}$ based on the IPCW-loss function. In this manner, we can generate a library of candidate estimators of $\bar{Q}^d_{0,k+1}$ ranging from estimators based on a large variety of parametric working models and highly data adaptive estimators. They form the library of candidate estimators in the super-learning algorithm for fitting $\bar{Q}^d_{0,k+1}$.

I.1 TMLE of blip-function projected on a working model

Consider a working model $\{m_\beta : \beta\}$ for $\bar{Q}^d_{0,k+1} = E_{P_0}(Y_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}} \mid V_{\bar{a}}(k))$. Define

$$\beta_0 = \arg\min_\beta \sum_{\bar{a}(k-1)} E_{P_0} h_{k+1}(\bar{a}(k-1), V_{\bar{a}}(k))(\bar{Q}^d_{0,k+1} - m_\beta)^2(\bar{a}(k-1), V_{\bar{a}}(k)).$$

In this subsection, we will develop a TMLE of β_0. The parameter can be represented as follows:

$$\beta_0 = \arg\min_\beta \sum_{\bar{a}(k-1)} E_{P_0} h_{k+1}(\bar{a}(k-1), V_{\bar{a}}(k))\{Y_{\bar{a}(k-1)1\underline{d}_{k+1}} - Y_{\bar{a}(k-1)0\underline{d}_{k+1}} - m_\beta(\bar{a}(k-1), V_{\bar{a}}(k))\}^2.$$

Let's assume that $m_\beta = \sum_{j=0}^p \beta_j e_j$ for a set of basis functions e_j. Let $h = (e_j : j = 0, \ldots, p)^\top$ and let $h^*_{k+1} = h h_{k+1}$. The parameter β_0 is defined by the following equation:

$$0 = \sum_{\bar{a}(k-1)} E_{P_0} h^*_{k+1}(Y_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}} - m_{\beta_0}(\bar{a}(k-1), V_{\bar{a}}(k))$$

Equivalently, it is defined as the solution

$$\sum_{\bar{a}(k-1)} E_{P_0} h^*_{k+1} m_{\beta_0}(\bar{a}(k-1), V_{\bar{a}}(k)) = \sum_{\bar{a}(k-1)} E_{P_0} h^*_{k+1}(\bar{a}(k-1), V_{\bar{a}}(k)) Y_{\bar{a}(k-1(1-0)\underline{d}_{k+1}}.$$

Given our linear model for m_{β_0}, the latter equation can be solved explicitly:

$$\beta_0 = C(P_0)^{-1}\Phi(P_0),$$

where

$$\Phi(P_0) = \sum_{\bar{a}(k-1)} E_{P_0} h^*_{k+1}(\bar{a}(k-1), V_{\bar{a}}(k)) Y_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}},$$

and

$$C(P_0) = \sum_{\bar{a}(k-1)} E_{P_0} h^*_{k+1} h^\top(\bar{a}(k-1), V_{\bar{a}}(k)) = \sum_{\bar{a}(k-1)} E_{P_0} h_{k+1} h h^\top(\bar{a}(k-1), V_{\bar{a}}(k)).$$

Note that $C(P_0)$ is a $(p+1)\times(p+1)$-matrix. This presents β_0 as a parameter of $\{P_{\bar{a}(k)\underline{d}_{k+1}} : \bar{a}(k)\}$. Let $D^*_\phi(P_0)$ be the efficient influence curve of $\Phi(P_0)$ and note that

$$\Phi(P_0) = \textstyle\sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}},$$

where $Z = h^*_{k+1}Y$.

One component of the efficient influence curve of β_0 is given by $C(P_0)^{-1}D^*_\phi(P_0)$. Let $D^*_C(P_0)$ be the $(p+1)\times(p+1)$ efficient influence curve of $C(P_0)$, which is nothing else than the matrix whose (i,j)-th element is the efficient influence curve $D^*_{C,(i,j)}(P_0)$ of $\sum_{\bar{a}(k-1)} E_{P_0}h_{k+1}e_ie_j(\bar{a}(k-1), V_{\bar{a}}(k))$. The efficient influence curve of the matrix-valued parameter $C(P)^{-1}$ is given by $-C(P_0)^{-1}D^*_C(P_0)C(P_0)^{-1}$. This shows that the efficient influence curve of the parameter $\beta = C(P)^{-1}\Phi(P)$ at $P = P_0$ is given by:

$$D^*_\beta(P_0) = C(P_0)^{-1}D^*_\phi(P_0) - C(P_0)^{-1}D^*_C(P_0)C(P_0)^{-1}\Phi(P_0).$$

This proves the following theorem.

Theorem 18 *Let $d_1 = d_1(\bar{a}(k-1)) = (\bar{a}(k-1)1\underline{d}_{k+1})$ and $d_0 = d_0(\bar{a}(k-1)) = (\bar{a}(k-1)0\underline{d}_{k+1})$. Let $Z = h^*_{k+1}Y$ ($(p+1)$-vector) and $\mathbf{Z} = h_{k+1}\phi\phi^\top$ ($(p+1)\times(p+1)$-matrix). Define $\bar{Q}^d_{Z,j} = E_P(Z_d \mid \bar{A}(j-2), \bar{L}(j-1))$, $j = K+1,\ldots,0$, and $\bar{Q}^d_{\mathbf{Z},j} = E_P(\mathbf{Z}_{\bar{a}} \mid \bar{A}(j-2), \bar{L}(j-1))$, $j = k,\ldots,0$. The $(p+1)$-dimensional efficient influence curve of $\beta_0 = C(P_0)^{-1}\Phi(P_0)$ is given by*

$$D^*_\beta(P_0) = C(P_0)^{-1}D^*_\phi(P_0) - C(P_0)^{-1}D^*_C(P_0)C(P_0)^{-1}\Phi(P_0),$$

where the $(p+1)$-dimensional $D^*_\phi(P_0) = \sum_{j=0}^{K+1}\{D^*_{\phi,j,1}(P_0) - D^*_{\phi,j,0}(P_0)\}$, and the $(p+1)\times(p+1)$ matrix $D^*_C(P_0) = \sum_{j=0}^{k} D^*_{C,j}(P_0)$, are defined as follows: for $\delta \in \{0,1\}$

$$\begin{aligned}
D^*_{\phi,K+1,\delta}(P_0) &= \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(K) = d_\delta(\bar{a}(k-1)))}{g_{0:K}(O)}(Z - \bar{Q}^{d_\delta}_{Z,K+1}) \\
D^*_{\phi,j,\delta}(P_0) &= \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(j-1) = d_\delta(\bar{a}(k-1)))}{g_{0:j-1}(O)}(\bar{Q}^{d_\delta}_{Z,j+1} - \bar{Q}^{d_\delta}_{Z,j}) \\
j &= K+1,\ldots,1 \\
D^*_{\phi,0,\delta}(P_0) &= \sum_{\bar{a}(k-1)} \{\bar{Q}^{d_\delta}_{Z,1} - \bar{Q}^{d_\delta}_{Z,0}\},
\end{aligned}$$

and

$$\begin{aligned}
D^*_{C,k}(P_0) &= \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(k-1) = \bar{a}(k-1))}{g_{0:k-1}(O)}(\mathbf{Z} - \bar{Q}^{\bar{a}(k-1)}_{\mathbf{Z},k}) \\
D^*_{C,,j}(P_0) &= \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(j-1) = \bar{a}(j-1))}{g_{0:j-1}(O)}(\bar{Q}^{\bar{a}(k-1)}_{\mathbf{Z},j+1} - \bar{Q}^{\bar{a}(k-1)}_{\mathbf{Z},j}) \\
j &= k,\ldots,1 \\
D^*_{C,0}(P_0) &= \sum_{\bar{a}(k-1)} \{\bar{Q}^{\bar{a}(k-1)}_{\mathbf{Z},1} - \bar{Q}^{\bar{a}(k-1)}_{\mathbf{Z},0}\},
\end{aligned}$$

For $j \geq k+1$, we have

$$D^*_{\phi,j,\delta}(P_0) = \frac{I(\bar{A}(k+1:j-1) = d_\delta(\bar{a}(k-1)))}{g_{0:j-1}(O)}(\bar{Q}^{d_\delta}_{Z,j+1} - \bar{Q}^{d_\delta}_{Z,j}).$$

For $j < k+1$, we have

$$D^*_{\phi,j,\delta}(P_0) = \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)}(\bar{Q}^{d_\delta}_{Z,j+1} - \bar{Q}^{d_\delta}_{Z,j}).$$

Above we already presented a TMLE for $C_{ij}(P_0) = \sum_{\bar{a}(k-1)} E_{P_0}\mathbf{Z}(i,j)_{\bar{a}(k-1)}$ and $\Phi_{a(k)}(P_0) = \sum_{\bar{a}(k-1)} E_{P_0} Z_{\bar{a}(k-1)a(k)\underline{d}_{k+1}}$ and thus $\Phi(P_0) = \Phi_1(P_0) - \Phi_0(P_0)$. Let C^*_n and ϕ^*_n be these two TMLEs. In a later Section K we also present the TMLE directly targeting $\Phi(P_0)$ instead of plugging in separate TMLEs of $\Phi_{a(k)}(P_0)$ for each $a(k) \in \{0,1\}$. We propose to estimate β with the plug-in TMLE:

$$\beta^*_n = \{C^*_n\}^{-1}\phi^*_n.$$

The TMLEs C^*_n and ϕ^*_n are constructed so that the efficient influence curve equations $P_n D^*_C(Q^*_n, g_n) = 0$ and $P_n D^*_\phi(Q^*_n, g_n) = 0$, and, as a consequence, the efficient influence curve $P_n D_\beta(Q^*_n, g_n) = 0$ is solved as well. This TMLE is not really a substitution estimator since we used separate TMLE for the components $C(P_0)$ and $\Phi(P_0)$ and also separate TMLE for each $C_{ij}(P_0)$, but each of these components is estimated with a double robust efficient substitution estimator.

I.2 Estimation of working model by minimizing TMLE of risk

Recall that

$$\beta_0 = \arg\min_\beta \sum_{\bar{a}(k)} E_{P_0} Z(\beta)_{\bar{a}(k)\underline{d}_{k+1}},$$

where

$$Z(\beta) \equiv h_{k+1}\{Y - m_\beta\}^2.$$

For a given β, let $R^*_n(\beta)$ be a TMLE, as presented above, of $R_0(\beta) = \sum_{\bar{a}(k)} E_{P_0} Z(\beta)_{\bar{a}(k)\underline{d}_{k+1}}$. One can now define $\beta^*_n = \arg\min_\beta R^*_n(\beta)$. Even though this might seem to be a very computer intensive method, typical iterative algorithms for minimizing $R^*_n(\beta)$ will only require knowing the function at the current value and past-values (so that one has a sense of slope). Therefore, one would compute the values $R^*_n(\beta)$ on the fly, thereby minimizing the number of times one needs

to compute the β-specific TMLE. The advantage of this approach seems to be that it allows us to use a single TMLE targeting $R_0(\beta)$, while the above method relies on separate TMLEs for the normalizing matrix and $\Phi(P_0)$.

J Oracle results

Let's consider a randomized controlled trial so that g_0 is known (assuming no censoring). In that case, we can use the known IPCW-loss-function $L_{IPCW,g_0}(\bar{Q}^d_{k+1})$ obtained by setting $D_1(g_0) = (2A(k)-1)/g_{0,A(k)}Y$. The loss-based dissimilarity of this loss-function is given by

$$P_0\{L_{IPCW,g_0}(\bar{Q}^d_{k+1}) - L_{IPCW,g_0}(\bar{Q}^d_{0,k+1})\} = \sum_{\bar{a}(k-1)} P_{0,\bar{a}(k-1)} h_{k+1}(\bar{Q}^d_{k+1} - \bar{Q}^d_{0,k+1})^2(O_{\bar{a}(k-1)}),$$

for the squared error loss, and similarly for the log-likelihood version. Consider the super-learner based on this loss function. Due to the oracle inequality for the cross-validation selector α_n, if none of the candidate estimators $\hat{\bar{Q}}^d_{k+1,\alpha}$ converges at the parametric rate $1/\sqrt{n}$ to $\bar{Q}^d_{0,k+1}$, then we have that $\hat{\bar{Q}}^d_{k+1,\alpha_n}(P_n)$ is asymptotically equivalent (i.e. ratio of loss-based dissimilarities with $\bar{Q}_0$ converges to 1) with the oracle selected estimator $\hat{\bar{Q}}_{\tilde{\alpha}_n}(P_n)$, where the oracle selector is defined as

$$\begin{aligned}\tilde{\alpha}_n &= \arg\min_{\alpha} E_{B_n} P_0 L_{IPCW,g_0}(\hat{\bar{Q}}^d_{\alpha}(P^0_{n,B_n})) \\ &= \arg\min_{\alpha} E_{B_n} \sum_{\bar{a}(k-1)} P_0 h_{k+1}(\left(\hat{\bar{Q}}^d_{k+1,\alpha}(P^0_{n,B_n}) - \bar{Q}^d_{0,k+1}\right)^2 (\bar{a}(k-1), V_{\bar{a}}(k)).\end{aligned}$$

This result relies on the loss-function $L_{IPCW,g_0}(\bar{Q}^d_{k+1})$ to be uniformly bounded in O and $\bar{Q}^d_{k+1}$, which is arranged by assuming the strong version of the positivity assumption. If one of the candidate estimators converges at rate $1/\sqrt{n}$ (e.g., one of candidate estimators is based on a correctly specified parametric model), then the super-learner also converges at rate $1/\sqrt{n}$, but in this case, it is not asymptotically equivalent with the oracle selector. These results still hold if $J = J(n)$ converges to infinity as fast as a polynomial power in n. So this proves that the super-learner is asymptotically optimal in the sense that it outperforms any competitor estimator of the blip-function by simply including this competitor in the library of estimators that defines the super-learner.

We could improve the cross-validated risk estimators, and thereby the cross-validation selector, by using the estimated loss $L_{IPCW,Q_n,g_0}(\bar{Q}^d_{k+1})$ based on an estimator $D_1(Q_n, g_0)$ of $D_1(Q_0, g_0)$.

In an observational study, we could use the estimated DR-IPCW loss $L_{DR-IPCW,D_{1n},Q_n,g_n}(\bar{Q}^d_{k+1})$. Finite sample and asymptotic results for the resulting cross-validation selector are presented in (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006): in essence, one still obtains powerful oracle results for the cross-validation selector but the rate of convergence is upper-bounded by the product of the rates at which g_n converges to g_0 and Q_n converges to Q_0. Thus in observational studies in which one has strong knowledge about the treatment assignment mechanism or one knows that there is a single time-dependent covariate (e.g., the outcome process at that time point) that blocks the effect of the past on the outcome so that it is sufficient to only adjust for this single time-dependent covariate when fitting the treatment mechanism (and the past treatment regimen), the cross-validation selector may still be asymptotically equivalent with the oracle selector above that treats g_0 as known, even if Q_n converges to a misspecified Q.

We also obtained oracle inequalities for the super-learner based on the CV-TMLE of the risk ((van der Laan and Petersen, 2012; Diaz and van der Laan, 2013)).

K CV-TMLE of risk of candidate estimator of blip function.

Above we treated D_1 as given in the definition of the risk, resulting in the two stage estimator that first estimates $D_1(Q_0, g_0)$ and then applies the CV-TMLE for the risk parameter that treats D_1 as given. In this section, we directly target the risk parameter as a parameter of the data distribution. The following theorem presents the risk function as a parameter of P_0.

Theorem 19 *Consider*

$$\begin{aligned}\bar{Q}^d_{0,k+1} = &E_{P_0}(Y_{\bar{a}(k-1),A(k)=(1,1),d_{A(k+1:K)}} \mid V_{\bar{a}(k-1)}(k) = v(k)) \\ &-E_{P_0}(Y_{\bar{a}(k-1),A(k)=(0,1),d_{A(k+1:K)}} \mid V_{\bar{a}(k-1)}(k) = v(k)).\end{aligned}$$

Define

$$Z_0 = h_{k+1}\bar{Q}^{d2}_{k+1} \text{ and } Z = \bar{Q}^d_{k+1}Y.$$

Define the following risk function for $\bar{Q}^d_{0,k+1}$*:*

$$R_{\bar{Q}^d_{k+1}}(P_0) \equiv E_{P_0} \sum_{\bar{a}(k-1)} Z_{0,\bar{a}(k-1)} - 2\left\{Z_{\bar{a}(k-1)1\underline{d}_{k+1}} - Z_{\bar{a}(k-1)0\underline{d}_{k+1}}\right\}.$$

We have

$$\sum_{\bar{a}(k-1)} E_0 h_{k+1}(\bar{Q}^d_{k+1} - \bar{Q}^d_{0,k+1})^2(O_{\bar{a}(k-1)}) = R_{\bar{Q}^d_{k+1}}(P_0) + E_0 \sum_{\bar{a}(k-1)} h_{k+1}\{\bar{Q}^d_{0,k+1}\}^2(O_{\bar{a}(k-1)}).$$

Therefore,

$$\bar{Q}^d_{0,k+1} = \arg\min_{\bar{Q}^d_{k+1}} R_{\bar{Q}^d_{k+1}}(P_0).$$

Let $D^*_{\bar{a}(k-1)}(P_0) = \sum_{j=0}^k D^*_{\bar{a}(k-1),j}(P_0)$ be the efficient influence curve of $E_0 Z_{0,\bar{a}(k-1)}$. Let $\bar{Q}^{\bar{a}(k-1)}_{Z_0,j} = E(Z_{0,\bar{a}(k-1)} \mid \bar{L}(j-1), \bar{A}(j-2))$ for $j = 0, \ldots, k$. We have

$$D^*_{\bar{a}(k-1),j}(P_0) = \frac{I(\bar{A}(j-1) = \bar{a}(j-1))}{g_{0:j-1}}(\bar{Q}^{\bar{a}(k-1)}_{Z_0,j+1} - \bar{Q}^{\bar{a}(k-1)}_{Z_0,j}),$$

$j = 0, \ldots, k$. Let $D^*_{\bar{a}(k-1)1\underline{d}_{k+1}}(P_0) = \sum_{j=0}^{K+1} D^*_{\bar{a}(k-1)0\underline{d}_{k+1}}$ be the efficient influence curves of $E_0 Z_{\bar{a}(k-1)1\underline{d}_{k+1}}$ and $E_0 Z_{\bar{a}(k-1)0\underline{d}_{k+1}}$, respectively. Let $\bar{Q}^{\bar{a}(k)\underline{d}_{k+1}}_{Z,j} = E(Z_{\bar{a}(k)\underline{d}_{k+1}} \mid \bar{L}(j-1), \bar{A}(j-2))$. We have

$$D^*_{\bar{a}(k)\underline{d}_{k+1},j}(P_0) = \frac{I(\bar{A} = (\bar{a}(k), \underline{d}_{k+1}))}{g_{0:j-1}(O)}(\bar{Q}^{d(\bar{a})}_{Z,j+1} - \bar{Q}^{d(\bar{a})}_{Z,j}),$$

$j = 0, \ldots, K+1$. The efficient influence curve of $R_{\bar{Q}^d_{k+1}}(P_0)$ is given by $D^*_{\bar{Q}^d_{k+1}}(P_0) = \sum_{j=0}^{K+1} D^*_{\bar{Q}^d_{k+1},j}(P_0)$, where

$$
\begin{aligned}
D^*_{\bar{Q}^d_{k+1},j}(P_0) &= \sum_{\bar{a}(k-1)} D^*_{\bar{a}(k-1),j}(P_0) + D^*_{\bar{a}(k-1)1\underline{d}_{k+1},j}(P_0) - D^*_{\bar{a}(k-1)0\underline{d}_{k+1},j}(P_0) \\
&= \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(j-1) = \bar{a}(j-1))}{g_{0:j-1}(O)} (\bar{Q}^{\bar{a}(k-1)}_{Z_0,j+1} - \bar{Q}^{\bar{a}(k-1)}_{Z_0,j}) \\
&\quad + \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(j-1) = \bar{a}(j-1))}{g_{0:j-1}(O)} (\bar{Q}^{\bar{a}(k-1)1\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{\bar{a}(k-1)1\underline{d}_{k+1}}_{Z,j}) \\
&\quad - \sum_{\bar{a}(k-1)} \frac{I(\bar{A}(j-1) = \bar{a}(j-1))}{g_{0:j-1}(O)} (\bar{Q}^{\bar{a}(k-1)0\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{\bar{a}(k-1)0\underline{d}_{k+1}}_{Z,j}) \\
&= \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)} (\bar{Q}^{a(j:k-1)}_{Z_0,j+1} - \bar{Q}^{a(j:k-1)}_{Z_0,j}) \\
&\quad + \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)} (\bar{Q}^{a(j:k-1)(1-0)\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{a(j:k-1)(1-0)\underline{d}_{k+1}}_{Z,j}) \\
&\quad j = 0, \ldots, k \\
D^*_{\bar{Q}^d_{k+1},j}(P_0) &= \sum_{\bar{a}(k-1)} \{D^*_{\bar{a}(k-1)1\underline{d}_{k+1},j}(P_0) - D^*_{\bar{a}(k-1)0\underline{d}_{k+1},j}(P_0)\} \\
&= \sum_{\bar{a}(k-1)} \frac{I(\bar{A} = d(\bar{a}(k-1)))A_2(k)(2A_1(k)-1)}{g_{0:j-1}(O)} \\
&\qquad (\bar{Q}^{\bar{a}(k-1)A(k)\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{\bar{a}(k-1)A(k)\underline{d}_{k+1}}_{Z,j}) \\
&= \frac{I(A(k+1,j-1) = \underline{d}_{k+1}))A_2(k)(2A_1(k)-1)}{g_{0:j-1}(O)} (\bar{Q}^{A(k)\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{A(k)\underline{d}_{k+1}}_{Z,j}) \\
&= \frac{I(A(k+1,j-1) = \underline{d}_{k+1}))A_2(k)(2A_1(k)-1)}{g_{0:j-1}(O)} (\bar{Q}^{\underline{d}_{k+1}}_{Z,j+1} - \bar{Q}^{\underline{d}_{k+1}}_{Z,j}) \\
&\quad j = k+1, \ldots, K+1.
\end{aligned}
$$

Thus, we have shown the following result.

Theorem 20 *The efficient influence curve of $R_{\bar{Q}^d_{k+1}}(P_0)$ is given by $D^*_{\bar{Q}^d_{k+1}}(P_0) =$*

$\sum_{j=0}^{K+1} D^*_{\bar{Q}^d_{k+1},j}(P_0)$, *where*

$$\begin{aligned} D^*_{\bar{Q}^d_{k+1},j}(P_0) &= \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)}(\bar{Q}^{a(j:k-1)}_{j+1} - \bar{Q}^{a(j:k-1)}_{j}) \\ &\quad + \sum_{a(j:k-1)} \frac{1}{g_{0:j-1}(O)}(\bar{Q}^{a(j:k-1)(1-0)\underline{d}_{k+1}}_{j+1} - \bar{Q}^{a(j:k-1)(1-0)\underline{d}_{k+1}}_{j}) \\ &\quad j = 0, \ldots, k \\ D^*_{\bar{Q}^d_{k+1},j}(P_0) &= \frac{I(A(k+1, j-1) = \underline{d}_{k+1})A_2(k)(2A_1(k)-1)}{g_{0:j-1}(O)}(\bar{Q}^{\underline{d}_{k+1}}_{j+1} - \bar{Q}^{\underline{d}_{k+1}}_{j}) \\ &\quad j = k+1, \ldots, K+1 \end{aligned}$$

K.1 Sequential regression representation of risk

We have

$$R_{\bar{Q}^d_{k+1}}(P_0) = \sum_{\bar{a}(k-1)\in\mathcal{A}(k-1)} \left\{ E_0 Z_{0,\bar{a}(k-1)} + E_0 Z_{\bar{a}(k-1)1\underline{d}_{k+1}} - E_0 Z_{\bar{a}(k-1)0\underline{d}_{k+1}} \right\}.$$

Consider

$$\sum_{\bar{a}(k-1)} \{E_0 Z_{\bar{a}(k-1)1\underline{d}_{k+1}} - E_0 Z_{\bar{a}(k-1)0\underline{d}_{k+1}}\}.$$

We now present a sequential regression representation of the latter quantity. Firstly, regress $Z = 2h_{k+1}\bar{Q}^d_{k+1}Y$ on $\bar{A}(K), \bar{L}(K)$, and set $A(K) = d_K$, which yields $\bar{Q}^d_{Z,K+1} = E_P(Z \mid \bar{A}(K-1), \bar{L}(K))$. Now, regress $\bar{Q}^d_{Z,k+1}$ on $\bar{A}(K-1), \bar{L}(K-1)$, and set $A(K-1) = d_{K-1}$. This yields $\bar{Q}_{Z,K} = E_P(Z_d \mid \bar{A}(K-2), \bar{L}(K-1))$. Now, regress $\bar{Q}_{Z,K}$ on $\bar{A}(K-2)$ and $\bar{L}(K-2)$ and set $A(K-2) = d_{K-2}$, giving $\bar{Q}^d_{Z,K-1} = E(Z_d \mid \bar{A}(K-2), \bar{L}(K-2))$. So in this way, we obtain $\bar{Q}_{Z,k+2} = E(Z_d \mid \bar{A}(k), \bar{L}(k+1))$. Now, regress the latter on $\bar{A}(k), \bar{L}(k)$ to obtain $E(Z_d \mid \bar{A}(k), \bar{L}(k))$ and compute

$$\bar{Q}^{1-0d}_{Z,k+1} = E(Z_d \mid A(k) = 1, \bar{A}(k-1), \bar{L}(k)) - E(Z_d \mid A(k) = 0, \bar{A}(k-1), \bar{L}(k)).$$

This represents $\bar{Q}^d_{Z,k+1} = E(Z_{1-0\underline{d}_{k+1}} \mid \bar{A}(k-1), \bar{L}(k))$. Now regress the latter on $\bar{A}(k-1), \bar{L}(k-1)$. Now, set $A(k-1) = a(k-1)$ giving $\bar{Q}^{a(k-1)(1-0)\underline{d}_{k+1}}_{Z,k}$ for each $a(k-1)$. Regress $\bar{Q}^{a(k-1)(1-0)\underline{d}_{k+1}}_{Z,k}$ on $\bar{A}(k-2), \bar{L}(k-2)$ and set $A(k-2) = a(k-2)$ for each $a(k-2)$. This results in $\bar{Q}^{a(k-2:k-1)(1-0)d}_{Z,k-1} = E(Z_{a(k-2:k-1)(1-0)\underline{d}_{k+1}} \mid \bar{A}(k-3), \bar{L}(k-2))$. Iterate this, giving us for each

$\bar{a}(k-1)$, $\bar{Q}_{Z,1}^{a(0:k-1)d} = E_P(Z_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}} \mid L(0))$ and finally averaging over $L(0)$ yields $E_P(Z_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}})$.

Note that in this sequential regression algorithm we started regressing a difference $\bar{Q}_{Z,k+1}^{(1-0)d}$. Alternatively, one uses the iterative regression to evaluate $E_P(Z_{\bar{a}(k-1)\delta\underline{d}_{k+1}})$ separately for $\delta \in \{0,1\}$, and one takes the difference at at the end.

We have a separate sequential regression representation for $\sum_{\bar{a}(k-1)\in\mathcal{A}(k-1)} E_0 Z_{0,\bar{a}(k-1)}$.

K.2 Corresponding TMLE based on sequential regression representation of risk

Consider the term

$$\sum_{\bar{a}(k-1)} E_0 Z_{\bar{a}(k-1)(1-0)\underline{d}_{k+1}}.$$

Firstly, regress $Z = 2h_{k+1}\bar{Q}_{Z,k+1}^d Y$ on $\bar{A}(K), \bar{L}(K)$ and set $A(K) = d_K$, which yields an initial estimator of $\bar{Q}_{Z,K+1}^d = E(Y_d \mid \bar{A}(K-1), \bar{L}(K))$. Consider a submodel $\text{Logit}\bar{Q}_{Z,K+1,n}^d(\epsilon) = \text{Logit}\bar{Q}_{Z,K+1,n}^d + \epsilon C_{K+1}(g)$, where $C_{K+1}(g) = I(A(k+1:K) = \underline{d}_{k+1})A_2(K)(2A_1(K)-1)/g_{0:K}(O)$, and let ϵ_n be the MLE obtained with logistic regression of Z onto C_{K+1}, using $\bar{Q}_{K+1,n}^d$ as off-set, and only using observations with $A_K = d_K$. Define the targeted estimator $\bar{Q}_{Z,K+1,n}^{d*} = \bar{Q}_{Z,K+1,n}^d(\epsilon_n)$ of $E_P(Y_{\underline{d}_{k+1}} \mid \bar{A}(K-1), \bar{L}(K))$. Set $j = K$. Now, regress $\bar{Q}_{Z,j+1,n}^{d*}$ on $\bar{A}(j-1), \bar{L}(j-1)$, and set $A(j-1) = d_{j-1}$, which yields an initial estimator $\bar{Q}_{Z,j,n}^d$ of $\bar{Q}_{Z,j}^d$. Consider the submodel $\text{Logit}\bar{Q}_{Z,j,n}(\epsilon) = \text{Logit}\bar{Q}_{Z,j,n} + \epsilon C_j(g)$, where $C_j(g) = I(A(k+1:j-1) = \underline{d}_{k+1})A_2(k)(2A_1(k)-1)/g_{0:j-1}(O)$. Let ϵ_n be the MLE as defined above. This yields a targeted estimator $\bar{Q}_{Z,j,n}^* = \bar{Q}_{Z,j,n}(\epsilon_n)$ of $E(Y_{\underline{d}_{k+1}} \mid \bar{A}(j-2), \bar{L}(j-1))$. Iterate this from $j = K, \ldots, k+1$, resulting in an initial estimator of $\bar{Q}_{Z,k+1}^{a(k)d} = E(Y_{\underline{d}_{k+1}} \mid A(k) = a(k), \bar{A}(k-1), \bar{L}(k))$ for each choice of $a(k)$. Consider submodel $\text{Logit}\bar{Q}_{Z,k+1,n}^{a(k)d}(\epsilon) = \text{Logit}\bar{Q}_{Z,k+1,n}^{a(k)d} + \epsilon C_{k+1}(g)$, where $C_{k+1}(g) = A_2(k)I(A_1(k) = a(k))/g_{0:k}(O)$. Let ϵ_n be the MLE, and define $\bar{Q}_{Z,k+1,n}^{a(k)d*} = \bar{Q}_{k+1,n}^{a(k)d}(\epsilon_n)$. This has thus resulted in a targeted estimator of $\bar{Q}_{Z,k+1}^{a(k)=1\underline{d}_{k+1}}$ and $\bar{Q}_{Z,k+1}^{a(k)=0\underline{d}_{k+1}}$. The difference $\bar{Q}_{Z,k+1,n}^{(1-0)\underline{d}_{k+1}*}$ is a targeted estimator of $\bar{Q}_{Z,k+1}^{(1-0)\underline{d}_{k+1}} = E(Z_{a(k)=1,\underline{d}_{k+1}} \mid \bar{A}(k-1), \bar{L}(k)) - E(Z_{a(k)=0,\underline{d}_{k+1}} \mid \bar{A}(k-1), \bar{L}(k))$. Regressing the difference $\bar{Q}_{Z,k+1,n}^{(1-0)\underline{d}_{k+1}*}$ on $\bar{A}(k-1), \bar{L}(k-1)$ and setting $A(k-1) = a(k-1)$ yields in initial estimator $\bar{Q}_{Z,k,n}^{a(k-1)(1-0)\underline{d}_{k+1}}$. Construct $\bar{Q}_{Z,k,n}^{a(k-1)(1-0)d*} = \bar{Q}_{Z,k,n}^{a(k-1)(1-0)\underline{d}_{k+1}}(\epsilon_n)$ using clever covariate $C_k(g) = 1/g_{0:k-1}$ as above, using standardization and logistic regression. Iterate this, giving for each $\bar{a}(k-1)$, $\bar{Q}_{Z,1,n}^{\bar{a}(k-1)(1-0)\underline{d}_{k+1}*}$, and finally

$\bar{Q}_{Z,0,n}^{\bar{a}(k-1)(1-0)\underline{d}_{k+1}*} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_{Z,1,n}^{\bar{a}(k-1)(1-0)\underline{d}_{k+1}*}(L_i(0))$. Our estimator of this risk-term is thus $\sum_{\bar{a}(k-1)} \bar{Q}_{Z,0,n}^{\bar{a}(k-1)(1-0)\underline{d}_{k+1}*}$.

Similarly, we obtain the estimator $\sum_{\bar{a}(k-1)} \bar{Q}_{0,n}^{\bar{a}(k-1)*}$, so that the TMLE of our risk is given by the sum of these two estimators.

K.3 IPCW

We have

$$R_{\bar{Q}_{k+1}^d}(P_0) = \sum_{\bar{a}(k-1)\in\mathcal{A}(k-1)} \left\{ E_0 Z_{0,\bar{a}(k-1)} + E_0 Z_{\bar{a}(k-1)1\underline{d}_{k+1}} - E_0 Z_{\bar{a}(k-1)0\underline{d}_{k+1}} \right\}.$$

An IPCW-estimator is given by

$$\begin{aligned} &R_{\bar{Q}_{k+1}^d, IPCW, n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{a}(k-1)} \frac{I(\bar{A}_i(k-1)=\bar{a}(k-1)}{g_{0:k-1}(O_i)} Z_{0i}(\bar{Q}_{k+1}^d) \\ &+\frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{a}(k-1)} \frac{I(\bar{A}_i(k-1)=\bar{a}(k-1), A_i(k+1:K)=\underline{d}_{k+1})(2A_i(k)-1)}{g_{0:K}(O_i)} Z_i(\bar{Q}_{k+1}^d) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{g_{0:k-1}(O_i)} Z_{0,i}(\bar{Q}_{k+1}^d) + \frac{I(A_i(k+1:K)=\underline{d}_{k+1})(2A_i(k)-1)}{g_{0:K}(O_i)} Z_i(\bar{Q}_{k+1}^d). \end{aligned}$$

Let $L_{IPCW,g}(\bar{Q}_{k+1}^d) = \frac{1}{g_{0:k-1}(O)} Z_0(\bar{Q}_{k+1}^d) + \frac{I(A(k+1:K)=\underline{d}_{k+1})(2A(k)-1)}{g_{0:K}(O)} Z(\bar{Q}_{k+1}^d)$. Then this IPCW-estimator can be represented as $P_n L_{IPCW,g_n}(\bar{Q}_{k+1}^d)$. We can also use a stabilized IPCW-estimator. The CV-IPCW estimator is defined as $R_{\hat{\bar{Q}})_{k+1}^d, CV-IPCW, n} = 1/B \sum_b P_{n,b}^1 L_{IPCW,g_{n,b}}(\bar{Q}_{k+1,n,b}^d)$.

K.4 Double robust estimating equation based estimator of risk.

Consider the efficient influence curve $D_{\bar{Q}_{k+1}^d}^*(P_0)$ of $R_{\bar{Q}_{k+1}^d}(P_0)$ and note that it can be represented as $L_{Q_0,g_0}(\bar{Q}_{k+1}^d) - R_{\bar{Q}_{k+1}^d}(P_0)$. Therefore, an estimating equation based estimator based on solving the efficient influence curve estimating equation in the target parameter is given by:

$$R_{\bar{Q}_{k+1}^d, EE, n} = P_n L_{Q_n,g_n}(\bar{Q}_{k+1}^d).$$

The cross-validated estimating equation based estimator for a given estimator $\hat{\bar{Q}}_{k+1}^d$ is defined as:

$$R_{\hat{\bar{Q}}_{k+1}^d, CV-EE, n} = \frac{1}{B}\sum_{b=1}^{B} P_{n,b}^1 L_{Q_{n,b},g_{n,b}}(\bar{Q}_{k+1,n,b}^d).$$

K.5 Asymptotic linearity of CV-estimators of true conditional risk

Consider $R_{\hat{\bar{Q}}^d_{k+1},CV-TMLE}$. The true risk-parameter is defined as

$$R_{0,n} = \frac{1}{B}\sum_{b=1}^{B} R_{\bar{Q}^d_{k+1,n,b}}(P_0).$$

We obtained a representation $R_{\bar{Q}_{k+1,n,b}}(Q_0)$ for some Q_0 (parameter of P_0, i.e. sequential regressions) for this parameter, which has efficient influence curve $D^*_{\bar{Q}_{k+1,n,b}}(Q_0, g_0)$. We then constructed an estimator $Q^*_{n,b} = Q_{n,b}(\epsilon_n)$ of Q_0 based on training sample $P^0_{n,b}$ and MLEs ϵ_n, so that

$$0 = \frac{1}{B}\sum_{b=1}^{B} P^1_{n,b} D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b}).$$

Our proposed CV-TMLE was defined as $R^*_n = \frac{1}{B}\sum_{b=1}^{B} R_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b})$. We also have the identity:

$$R^*_n - R_{0,n} = -\frac{1}{B}\sum_{b=1}^{B} P_0 D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b}) + \frac{1}{B}\sum_{b=1}^{B} R_{P_0}(Q^*_{n,b}, Q_0, g_{n,b}, g_0).$$

Combining these two identities yields:

$$\begin{aligned} R^*_n - R_{0,n} &= \frac{1}{B}\sum_{b=1}^{B}(P^1_{n,b} - P_0) D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b}) \\ &\quad + \frac{1}{B}\sum_{b=1}^{B} R_{P_0}(Q^*_{n,b}, Q_0, g_{n,b}, g_0). \end{aligned}$$

For example, if $g_n = g_0$, then the remainder term equals zero. Or, if $Q_{n,b}$ and $g_{n,b}$ are both consistent, one might assume $R_{P_0}(Q^*_{n,b}, Q_0, g_{n,b}, g_0) = o_P(1/\sqrt{n})$. Under such an assumption, we have

$$R^*_n - R_{0,n} = \frac{1}{B}\sum_{b=1}^{B}(P^1_{n,b} - P_0) D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b}) + o_P(1/\sqrt{n}).$$

Since this cross-validated empirical process term is asymptotically normally distributed under minimal assumptions, this provides asymptotic linearity under very weak conditions. Specifically, if $P_0\{D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b}) - D^*_{\bar{Q}^d_{k+1}}(Q, g_0)\}^2$ converges to zero in probability, then

$$R^*_n - R_{0,n} = (P_n - P_0) D^*_{\bar{Q}^d_{k+1}}(Q, g_0) + o_P(1/\sqrt{n}).$$

In particular, we can estimate the asymptotic variance of $R_n^* - R_{0,n}$ with

$$\sigma_n^2 = \frac{1}{B}\sum_{b=1}^{B} P_{n,b}^1\{D^*_{\bar{Q}^d_{k+1,n,b}}(Q_{n,b}(\epsilon_n), g_{n,b})\}^2.$$

An asymptotic 95 % confidence interval is given by $R_n^* \pm 1.96\sigma_n/\sqrt{n}$.

L Pathwise differentiability of the mean outcome under V-optimal rule: multiple time-point treatment.

We already proved the following theorems for the two time-point treatment case. Since the proofs are a complete analogue, we will just state the theorems without proof.

Theorem 21 *For notational convenience, let suppress the $\bar{a}_2 = 1$ since it is 1 always. Recall the definitions of $\bar{Q}_{j0}$, $j = 1, \ldots, K+1$. We can represent $\Psi(P_0) = E_{P_{d_0}} Y_{d_0}$ as follows:*

$$\Psi(P_0) = E_{P_0} Y_{\bar{a}_1=0} + \sum_{j=0}^{K} E_{V_{\bar{a}}(j)} d_{0,A(j)}(\bar{a}(j-1), V_{\bar{a}}(j))\bar{Q}_{j+1,0}(\bar{a}(j-1), V_{\bar{a}}(j))\big|_{\bar{a}(j-1)=0}.$$

Theorem 22 *Assume that $P_0(|\ Y\ | < M) = 1$ for some $M < \infty$. The parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is pathwise differentiable with canonical gradient given by*

$$D^*(P_0) = \sum_{k=0}^{K+1} D_k^*(P_0),$$

where

$$\begin{aligned} D_0^*(P_0) &= E_{P_0}(Y_{d_0} \mid L(0), A(0) = d_{0,A(0)}(V(0))) - E_{P_0} Y_{d_0} \\ D_k^*(P_0) &= \frac{I(\bar{A}(k) = \bar{d}_k(V(k)))}{\prod_{j=0}^{k} g_{0,A(j)}(O)}(E_{P_0}(Y_{d_0} \mid \bar{L}_{d_0}(k)) - E_{P_0}(Y_{d_0} \mid \bar{L}_{d_0}(k-1)) \\ & \quad k = 1, \ldots, K+1. \end{aligned}$$

That is, $D^*(P_0)$ equals the efficient influence curve $D_0^*(d, P_0)$ for the parameter $\Psi_d(P) \equiv E_P Y_d$ treating d as given, at the V-optimal rule $d = d_0$: $D^*(P_0) = D_0^*(d_0, P_0)$.

L.1 TMLE of mean outcome under V-optimal rule: multiple time-point treatment.

Our proposed TMLE is to first estimate the optimal rule d_0, giving us an estimated rule $d_n(V) = (d_{n,A(0)}(V(0)), \ldots, d_{n,A(K)}(V(K))$, and subsequently apply the TMLE of EY_d for a fixed rule d at $d = d_n$ as presented in van der Laan and Gruber (2012). In a previous section we described a data adaptive estimator d_n of d_0, so that the TMLE presented in van der Laan and Gruber (2012) provides us with the TMLE of $E_0 Y_{d_0}$. The asymptotic linearity theorem for this TMLE is just a copy of Theorem 10.

L.2 TMLE and CV-TMLE of mean outcome under data-adaptively determined dynamic treatment: multiple time-point treatment

The presentation of the TMLE of $E_{P_0} Y_{d_n}$ and the CV-TMLE of $E_{B_n} E_{P_0} Y_{\hat{d}(P^0_{n,B_n})}$ (treating $\hat{d}(P^0_{n,B_n})$ as fixed) are presented in a complete analogue fashion as for the two time-point treatment case, and is therefore omitted here.