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

Year 2014 *Paper* 326

Super-Learning of an Optimal Dynamic Treatment Rule

Alexander R. Luedtke[*] Mark J. van der Laan[†]

[*]University of California, Berkeley, Division of Biostatistics, aluedtke@berkeley.edu
[†]University of California - Berkeley, laan@berkeley.edu

Super-Learning of an Optimal Dynamic Treatment Rule

Alexander R. Luedtke and Mark J. van der Laan

Abstract

We consider the estimation of an optimal dynamic two time-point treatment rule defined as the rule that maximizes the mean outcome under the dynamic treatment, where the candidate rules are restricted to depend only on 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 mechanisms. We propose data adaptive estimators of this optimal dynamic regime which are defined by sequential loss-based learning under both the blip function and weighted classification frameworks. Rather than \textit{a priori} selecting an estimation framework and algorithm, we propose combining estimators from both frameworks using a super-learning based cross-validation selector that seeks to minimize an appropriate cross-validated risk. One of the proposed risks directly measures the performance of the mean outcome under the optimal rule. The resulting selector is guaranteed to asymptotically perform as well as the best convex combination of candidate algorithms in terms of loss-based dissimilarity under conditions. We offer simulation results to support our theoretical findings. This work expands upon that of an earlier technical report (van der Laan, 2013) with new results and simulations, and is accompanied by a work which develops inference for the mean outcome under the optimal rule (van der Laan and Luedtke, 2014).

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. 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, 2009).

Researchers have aimed to learn optimal rules from data generated by sequential multiple assignment randomized trials (SMART) (Robins, 1986). Researchers have also aimed to learn dynamic treatments from observational studies: Cotton and Heagerty (2011); Orellana et al.; Robins et al. (2008a); Rosthøj 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-regime specific counterfactual outcome over a user supplied class of dynamic regimes. The reinforcement learning and statistical literature have made enormous advances in developing statistical methods that aim to learn such optimal rules.

We define the optimal dynamic multiple time-point treatment rule 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. 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 is an example of dynamic programming (Bellman, 1957). The optimal rule can be learned through fitting the sequential regressions, such as sequential linear least squares regression (see e.g., Murphy (2005)). 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 a 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.

Murphy (2003) and Robins (2003, 2004) develop structural nested mean models tailored to 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 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 until the first time point. These models are semi-parametric since they only rely on a parametric model of the blip function (at least in a SMART), but they 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. Structural nested mean models have also been generalized to learn of optimal rules that are restricted to only using a (counterfactual) subset of the past (Robins (2004) and Section 6.5 in van der Laan and Robins (2003)).

In Example 4 of Robins et al. (2008b), the authors propose selecting a data adaptive estimate of the optimal treatment rule by a particular cross-validation scheme over a set of basis functions, and show that this estimator achieves a data adaptive rate of convergence under smoothness assumptions on the blip function. This work only considers data generated by a point treatment randomized controlled trial (RCT), but makes no other model assumptions. Additionally, the library of estimators applied in this approach is limited and does not fully take advantage of the breadth of state of the art machine learning methods, though it is certainly an improvement over parametric approaches.

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) and Zhang et al. (2012) independently identify entire families of such reductions to classification.

Most of the above discussed estimation strategies rely on parametric assumptions. In a companion paper we develop inference for the mean outcome under the V-optimal rule, where the V-optimal rule is the optimal rule that relies only on a specified subset V of the covariate history (van der Laan and Luedtke, 2014). We develop the inference procedure under a semi-parametric model for a data distribution that is nonparametric, beyond possible knowledge about the treatment mechanism. For inference about the mean outcome under the optimal rule, it is crucial that we consistently estimate the optimal rule under this semi-parametric model at a sufficient rate.

Our proposed estimators of the V-optimal rule are based on sequential (analogous 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 an ensemble method known as super-learning to fit each rule after having estimated the optimal rule at future time points. 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 regression or classification machine learning algorithms. In this sense we have unified the more classical blip function approach with the recent optimal treatment classification literature.

By previously established oracle inequality results on the cross-validation selector established in the above references, our results guarantee that in a SMART the super-learner will be asymptotically equivalent with the estimator selected by the oracle selector and thereby outperform any of the parametric model based estimators and any of the other estimators in the family of candidate estimators, under the assumption that none of the parametric models are correctly specified. If one of the parametric models is correctly specified, the proposed method achieves an almost parametric $\log n/n$ rate. 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. Past findings and strongly suggest that this will also result in superior performance in most practical situations relative to *a priori* selecting one particular estimation procedure (Polley et al., 2012; van der Laan and Rose, 2012). We also outline how to develop a cross-validated targeted minimum loss-based estimator of the cross-validated risk to improve finite sample performance of this selector.

For the sake of presentation, we focus on two-time point treatments in this article. In the appendix of our earlier technical report (van der Laan, 2013) we generalize these results to general multiple time point treatments. We emphasize that this technical report is a distinct document from our companion paper, which focuses on inference for the mean outcome under the optimal rule in a model that is nonparametric beyond possible knowledge about the treatment mechanism (van der Laan and Luedtke, 2014).

1.1 Organization of article

Section 2 defines the optimal rule as a causal parameter, and gives identifiability assumptions under which the causal parameter is identified with a statistical parameter of the observed data distribution.

The remainder of the paper describes and evaluates strategies for estimating this statistical parameter that is identified with the optimal rule. Unless otherwise specified, all of the approaches presented in this paper aim to learn the optimal rule sequentially. That is, we first estimate the optimal treatment strategy at the second time point. Given an estimate of this second time point rule, we estimate the optimal rule at the first time point under (the G-computation distribution corresponding to) the counterfactual distribution in which the already estimated treatment rule is followed at the second time point.

Section 3 describes three classes of loss functions that can be used to estimate the optimal rule. Section 3.1 describes sequential estimation of blip functions based on any loss function that can give a valid estimate of a conditional mean (e.g. squared error loss), where the sign of the estimated conditional mean is used to estimate the optimal rule. Section 3.2 aims to directly estimate the optimal treatment by maximizing the sequential mean outcomes under the fitted rules, where the treatment at future time points is set according to the previously fit rule. Section 3.3 shows that maximizing an estimate of the mean outcome can be written as a weighted classification problem that includes a rich class of previously described classification loss functions. All loss functions presented in Section 3 rely on correct specification of the intervention mechanism, which is trivially true in an RCT without missingness. Double robust generalizations of the loss functions in Section 3 are presented in Appendix A.

Section 4 describes a cross-validation selector that can combine multiple estimation algorithms, including any of those which aim to minimize the empirical risks from the loss functions in Section 3. Section 4.1 gives the oracle inequality for the second time point treatment and gives examples of losses for which the oracle inequality will be satisfied. A finite sample oracle inequality is given to support the proposed methodology and the asymptotic implications of this inequality are described. Appendix B contains a proof that a particular

Section, Theorem	Loss function	Latent function?	Sequential?
Section 3.1 Theorem 2	Estimates blip functions using loss functions tailored to the estimation of means. Theorem 1 proves the validity of this estimation method for estimating V-optimal rule.	Yes	Yes
Section 3.2 Theorem 3	Risk function resulting from loss is the sequential mean outcome under the fitted rule at a given time point. Maximizing this quantity is directly targeted at our goal. A CV-TMLE of the risk resulting from this loss function is presented in Section 5.1.	No	Yes
Section 3.3 Theorem 4	Weighted $0-1$ loss function that we show is equivalent to the mean outcome loss from Section 3.2 up to an additive term that does not rely on the fitted rule.	Yes	Yes
Section 3.3 Theorem 5	Weighted surrogate loss functions that provide a convex approximations to the weighted $0-1$ loss function.	Yes	Yes
Section 5.2 N/A	Non-sequential super-learner that seeks to directly maximize the mean outcome under the two time point rule. Relies on sequential candidate estimators based on the losses above.	No	No

Table 1: Summary of all loss functions considered in this paper. All have risks that are related to the mean outcome under the optimal rule, so that the expected loss (risk) under the true distribution is minimized at either the optimal rule or some latent function whose sign gives the optimal rule. Note that all but one of the estimation procedures considered approach the problem sequentially, i.e. first estimating the optimal rule at the second time point and then, given this estimated rule, estimate the optimal rule at the first time point.

margin condition yields one of the conditions needed to establish our oracle inequality for the sequential mean outcome losses. Section 4.2 describes the super-learner for estimating the treatment rule at the first time point.

Section 5 outlines how we can replace the cross-validated empirical risk with a cross-validated targeted minimum loss-based estimator (CV-TMLE) of the risk. The CV-TMLE is a substitution estimator, and thus naturally respects the bounded nature of the data. Section 5.1 outlines a CV-TMLE for the sequential mean outcome losses presented in Section 3.2. Section 5.2 describes a non-sequential super-learner which directly uses the estimated mean outcome under the optimal rule to combine candidate estimators. This mean outcome based criteria is the only non-sequential strategy that we consider in this paper. The CV-TMLE in Section 5.2 is based on the CV-TMLE presented in our companion technical report.

Table 1 gives an overview of the loss functions we consider in this paper. Minimizing the risks resulting from all of these losses yields an estimate of the optimal rule, either directly for the loss in Section 3.2, or by the sign of the estimated latent function for all other losses. All but one of the losses considered estimates the optimal rule sequentially, and even this loss function relies on sequential candidate estimators.

Section 6 presents the simulation methods. The simulations compare our proposed super-learner to single choices of machine learning algorithms and misspecified parametric models. Section 7 presents the simulation results. Section 8 closes with a discussion and directions for future work.

All proofs are left to the appendix.

2 Formulation of optimal dynamic treatment estimation problem

We use the same formulation as is given in Section 2 of our companion technical report. We restate important notation here, but refer to the other paper for a more thorough discussion of the context and assumptions which identify our statistical parameter with a causal parameter.

For a discrete-time process $X(\cdot)$, we will use the notation $\bar{X}(t) = (X(s) : 0 \leq s \leq t)$, where $\bar{X}(-1) = \emptyset$. Suppose we observe n i.i.d. copies $O_1, \ldots, O_n \in \mathcal{O}$ of $O = (\bar{L}(1), \bar{A}(1), Y) \sim P_0$, where $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ is a binary treatment and $A_2(j)$ is an indicator of not being right-censored at "time" j, $j = 0, 1$. Each time point j has covariates $L(j)$ that precede treatment, $j = 0, 1$, and the outcome of interest is given by Y and occurs after time point 1. 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 distribution $Q_{0,L(1)}$ of $L(1)$, given $A(0), L(0)$, 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:

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

Throughout this article we will automatically assume the positivity assumption:

$$P_0\left(0 < \min_{a_1 \in \{0,1\}} g_{0,A(0)}(a_1, 1|L(0))\right) = 1$$
$$P_0\left(0 < \min_{a_1 \in \{0,1\}} g_{0,A(1)}(a_1, 1 \mid \bar{L}(1), A(0))\right) = 1. \quad (1)$$

The strong positivity assumption will be defined as this assumption (1), but where the 0 is replaced by a $\delta > 0$.

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 $V = (V(0), V(1))$. 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. Note that the rules for $A(0)$ and $A(1)$ are only a functions of $V(0)$ and $(A(0), V(1))$, respectively, and are restricted to set the observations to uncensored. Let $\mathcal{D}$ be the set of all such rules. We assume that $V(0)$ is a function of $V(1)$, but in the theorem below we indicate an alternative assumption. At times we abuse notation and let $a(0) \in \{0,1\} \times \{1\}$ and $a(1) \in \{0,1\} \times \{1\}$ represent the static rules at the first and second time points in which everyone receives treatment $a(0)$ or $a(1)$.

Define the distribution $P_{0,d}$ as the distribution with density

$$\begin{aligned}
&p_{0,d}(L(0), A(0), L(1), A(1), Y) \\
&\equiv I(A = d(V)) q_{0,L(0)}(L(0)) q_{0,L(1)}(L(1) \mid L(0), A(0)) q_{0,Y}(Y \mid \bar{L}(1), \bar{A}(1)),
\end{aligned}$$

where $q_{0,L(0)}$, $q_{0,L(1)}$, and $q_{0,Y}$ are the densities for $Q_{0,L(0)}$, $Q_{0,L(1)}$, and $Q_{0,Y}$ and all densities are absolutely continuous with respect to some dominating measure μ. This probability distribution $P_{0,d}$ is the G-computation formula (Robins (1987b,b, 1997, 1999); Gill and Robins (2001); Yu and van der Laan (2003)). In our companion paper we give identifiability results to relate $P_{0,d}$ to a counterfactual distribution in which the treatment rule d is, contrary to fact, implemented for the entire population. We use notation L_d (or Y_d, O_d) to mean the random variable with distribution $P_{0,d}$.

In this article we are concerned with estimation of the V-optimal rule defined as

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

The next theorem states an explicit form of the V-optimal individualized treatment rule d_0 as a function of P_0. We prove the theorem in our companion paper.

Theorem 1. *Suppose $V(0)$ is a function of $V(1)$. The V-optimal rule d_0 can be represented as the following explicit parameter of P_0:*

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

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

$$E_{P_0}(Y_{a(0),a(1)} \mid V(0), V_{a(0)}(1)) = E_{P_0}(Y_{a(0),a(1)} \mid V_{a(0)}(1)), \tag{2}$$

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

3 Sequential loss functions for the V-optimal rule

We will derive three approaches to estimate the V-optimal rule. The first is based on blip function methodology (Murphy, 2003; Robins, 2003, 2004). The second aims to directly maximize an estimate of the mean outcome under the optimal rule. The third is based on previously described weighted classification approaches (Zhao et al., 2012; Rubin and van der Laan, 2012; Zhang et al., 2012).

We will generally assume that $d_{A(1)} = d_{0,A(1)}$ when stating results in this section. Nonetheless, it is straightforward to show that the first time point loss functions in this section are valid for estimating the optimal fitted rule given correct specification of the intervention mechanism under the constraint that the second time point treatment must follow the possibly suboptimal rule $d_{A(1)}$.

For presentation purposes, all of the loss functions described in this section are inverse probability of censoring weighted (IPCW) loss functions. That is, these loss functions are correct if the intervention mechanism is specified correctly, which is trivially true in an RCT without missingness. We denote a (possibly misspecified) intervention mechanism estimate with g. We take $g_{A(0)}$ and $g_{A(1)}$ to be the resulting first and second time point intervention mechanisms.

The simplicity of the IPCW formulations comes at the expense of robustness and efficiency. In the appendix we present double robust versions all of the loss functions and theorems given in this section so that the loss functions will be correct if either the intervention mechanism is correctly specified or if particular conditional expectations of the outcome are correctly specified. Because the IPCW versions of the theorems are special cases of the double robust versions, we give proofs for the double robust case in the appendix and omit proofs for the IPCW case in this section.

3.1 Blip functions

We first give a formulation which aims to sequentially learn the blip functions at each time point. That is, we aim to sequentially learn the V-strata-specific average treatment effect at each time point. For the second time point, we find this strata-specific average treatment effect under the counterfactual distribution in which the first time point treatment is fixed at $a(0) \in \{0,1\} \times \{1\}$. For the first time point, we find this under the counterfactual distribution in which the second time point follows the estimated second time point rule.

Define

$$D_2(g)(O) = A_2(1)\frac{2A_1(1)-1}{g_{A(1)}(O)}Y. \tag{3}$$

Let $P_{0,a(0)}$ denote the static-intervention specific G-computation distribution $P_{0,a(0)}$ and $O_{a(0)}$ represents a counterfactual observation under this distribution. Let $L^F_{2,D_2(g)}(\bar{Q}_2)(O)$ denote a valid loss function for estimating $E_{P_{0,a(0)}}[D_2(g) \mid V_{a(0)}(1) = v_{a(0)}(1))]$, in the sense that

$$(a(0), v(1)) \mapsto E_{P_{0,a(0)}}[D_2(g)(O_{a(0)}) \mid V_{a(0)}(1) = v(1)]$$

minimizes

$$\sum_{\tilde{a}(0)\in\{0,1\}\times\{1\}} E_{P_{0,\tilde{a}(0)}}\left[L^F_{2,D_2(g)}(\bar{Q}_2)(O_{\tilde{a}(0)})\right] \tag{4}$$

over all measurable functions $\bar{Q}_2$ of $a(0)$ and $v(1)$. Because the minimum is over all measurable functions, one can split the above sum and minimize the expected loss (risk) first for $\tilde{a} = (0, 1)$, and then for $\tilde{a} = (1, 1)$. At the end of this section we provide two examples of loss functions satisfying this property. In fact, one can construct a valid $L^F_{2,D_2(g)}$ from any loss that can be used to fit a conditional mean. To identify the resulting risk function with the observed data distribution, we apply the IPCW mapping (van der Laan and Dudoit, 2003):

$$L_{2,g}(\bar{Q}_2)(O) = \frac{A_2(0)}{g_{A(0)}(O)} L^F_{2,D_2(g)}(\bar{Q}_2)(O), \tag{5}$$

Note that we inverse weight by the entire first time point intervention mechanism, not just the censoring mechanism at the first time point. We will use the sign of the $\bar{Q}_2$ which minimizes $L_{2,g}$ to estimate $d_{0,A(1)}$. For a given rule at the second time point $d_{A(1)}$, define

$$D_1(g)(O) = A_2(0)\frac{2A_1(0) - 1}{g_{A(0)}(O)} Y. \tag{6}$$

Let $L^F_{1,D_1(g)}$ be some loss that satisfies:

$$E_{P_{0,d_{A(1)}}}\left[D_1(g)(O_{d_{A(1)}}) \mid V(0) = \cdot\,\right] = \arg\min_{\bar{Q}_1} P_{0,d_{A(1)}} L^F_{1,D_1(g)}(\bar{Q}_1), \tag{7}$$

where $P_{0,d_{A(1)}}$ represents the post-intervention distribution corresponding with the dynamic intervention $d_{A(1)}$ and $O_{d_{A(1)}}$ represents a counterfactual observation under this distribution. Our proposed loss function is obtained by applying the IPCW mapping to the above loss function:

$$L_{1,d_{A(1)},g}(\bar{Q}_1)(O) = \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} L^F_{1,D_1(g)}(\bar{Q}_1). \tag{8}$$

We now state a theorem that gives conditions under which the above loss functions allow us to learn the optimal rule d_0.

Theorem 2. *Suppose the positivity assumption holds at g_0. Then:*

$$\begin{aligned}
&P_0\{L_{2,g_0}(\bar{Q}_2) - L_{2,g_0}(\bar{Q}_{20})\} = \sum_{a(0)} P_{0,a(0)} \left(L^F_{2,D_2(g_0)}(\bar{Q}_2) - L^F_{2,D_2(g_0)}(\bar{Q}_{20})\right) \\
&P_0\{L_{1,d_{0,A(1)},g_0}(\bar{Q}_1) - L_{1,d_{0,A(1)},g_0}(\bar{Q}_{10})\} = P_{0,d_{0,A(1)}} \left(L^F_{1,D_1(g_0)}(\bar{Q}_1) - L^F_{1,D_1(g_0)}(\bar{Q}_{10})\right),
\end{aligned} \tag{9}$$

where the sum is over $a(0) \in \{0,1\} \times \{1\}$*. As a consequence:*

$$\bar{Q}_{20} = \arg\min_{\bar{Q}_2} P_0 L_{2,g_0}(\bar{Q}_2)$$
$$\bar{Q}_{10} = \arg\min_{\bar{Q}_1} P_0 L_{1,d_{0,A(1)},g_0}(\bar{Q}_1) \tag{10}$$

A double robust generalization of the above theorem appears with proof in the appendix. We will refer to the quantities in (9) as loss-based dissimilarities for L_{2,g_0} and $L_{1,d_{0,A(1)},g_0}$, which represent the difference between the P_0-expected loss (risk) at a candidate function and the P_0-expected loss (risk) at the true parameter value. The loss-based dissimilarity is defined analogously for general losses.

Algorithm 1 outlines how to implement the proposed method for the second time point treatment in an RCT in which the missing mechanism is known (e.g. no missingness) given some loss $L^F_{2,D_2(g)}$. The class $\mathcal{F}$ that appears in the algorithm is defined by the choice of regression algorithm, e.g. a linear combination of $A_1(0)$ and covariates in $V(1)$. The estimation procedure for the rule at the first time point given a rule at the second time point is similar.

Algorithm 1 Blip function based estimation of $d_{0,A(1)}$

function BLIP(O_1,...,O_n)
    for $i = 1$ to n **do**
        Assign $A_2(1)_i \frac{2A_1(1)_i - 1}{g_{0,A(1)}(O_i)} Y_i$ to element i of the vector $D_2(g_0)$
    Regress $(D_2(g_0)_i : i = 1, ..., n)$ on $((A(0)_i, V(1)_i) : i = 1, ..., n)$ using loss function $L^F_{2,D_2(g)}$, where observation i receives weight $\frac{A_2(0)}{g_{0,A(0)}(O_i)}$. The goal is to find $\bar{Q}_{2n}$ in some class $\mathcal{F}$ such that: $\bar{Q}_{2n} = \arg\min_{\bar{Q}_2 \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} \frac{A_2(0)_i}{g_{0,A(0)}(O_i)} L^F_{2,D_2(g)}(\bar{Q}_2)(O_i)$.
    return $\bar{Q}_{2n}$

In an observational study or an RCT with missingness, one must also estimate the treatment and/or censoring mechanism g_0. The rate of convergence of the final estimate when g_0 is estimated will be upper bounded by the rate at which the estimate g converges to g_0. For this reason we suggest using the more efficient double robust inverse probability of censoring weighted (DR-IPCW) loss function presented in the appendix. The implementation of the double robust loss function can be slightly more difficult for the loss functions presented in this section, the details of which are explored in the appendix. The double robust losses based on those in Section 3.3, on the other hand, are straightforward to implement (details in appendix).

The expressions in (10) make L_{2,g_0} and $L_{1,d_{0,A(1)},g_0}$ valid losses. Even if the estimated rule is not the optimal rule, one can show that the blip function at the first time point will maximize the mean outcome under the constraint of the suboptimal second time point rule. In an observational study, we have access to an empirical rather than the true observed data distribution. Hence it may be important to consider the smoothness of $L^F_{2,D_2(g)}$ and $L^F_{1,D_1(g)}$ in the neighborhood of the minimizers in (4) and (7) so that reasonable estimation of the sequential risk functions is possible. It may also be desirable to have empirical process conditions on the class over which the blip functions are fit suffice to yield a well-behaved empirical risk minimization problem. Nonetheless, empirical process conditions are not needed for the cross-validation based super-learner algorithm proposed in Section 4 to be valid.

We close this section with two examples of loss functions that have many desirable statistical properties.

Example 1. Squared error loss.

$$L^F_{2,D_2(g),MSE}(\bar{Q}_2)(o) = h_2(a(0), v(1)) \left[D_2(g)(o) - \bar{Q}_2(a(0), v(1))\right]^2$$
$$L^F_{1,D_1(g),MSE}(\bar{Q}_1)(o) = h_1(v(0))(D_1(g)(o) - \bar{Q}_1(v(0)))^2,$$

where h_2 and h_1 represent positive user-supplied weight functions of $(a(0), v(1))$ and $v(0)$, respectively. By Theorem 2:

$$P_0\{L_{2,g_0,MSE}(\bar{Q}_2) - L_{2,g_0,MSE}(\bar{Q}_{20})\} = \sum_{a(0)} P_0 \left\{h_2(\bar{Q}_2 - \bar{Q}_{20})^2(a(0), V_{a(0)}(1))\right\}$$
$$P_0\{L_{1,d_{0,A(1)},g_0,MSE}(\bar{Q}_1) - L_{1,d_{A(1)},g_0,MSE}(\bar{Q}_{10})\} = P_0 \left\{h_1(\bar{Q}_1 - \bar{Q}_{10})^2(V(0))\right\}.$$

Using $L^F_{2,D_2(g_0),MSE}$ in Algorithm 1 shows that the optimal rule at the second time point can be found using a weighted squared-error loss function. The same holds for estimating the optimal rule at the first time point. □

Example 2. Quasi-log likelihood loss. Suppose it is known that $\bar{Q}_{20}$ and $\bar{Q}_{10}$ fall in some interval (a, b). For example, if $Y \in (0, 1)$ then we can take $a = -b = 1$. Define $\eta : \mathbb{R} \to \mathbb{R}$ as the function $\eta(x) = \frac{x-a}{b-a}$. For any real-valued function f we define f^η to be $\eta \circ f$, where $\circ$ denotes function composition. Define:

$$-L^F_{2,D_2(g),KL}(\bar{Q}_2)$$
$$= D_2(g)^\eta \log\left(\bar{Q}_2^\eta(a(0), v(1))\right) + (1 - D_2(g)^\eta) \log\left(1 - \bar{Q}_2^\eta(a(0), v(1))\right)$$
$$-L^F_{1,D_1(g),KL}(\bar{Q}_1)$$
$$= D_1(g)^\eta \log\left(\bar{Q}_1^\eta(v(0))\right) + (1 - D_1(g)^\eta) \log\left(1 - \bar{Q}_1^\eta(v(0))\right),$$

for all $\bar{Q}_2$, $\bar{Q}_1$ with range in (a, b).

The loss-based dissimilarities that result after applying the IPCW mapping are equal to Kullback-Leibler dissimilarities:

$$P_0\{L_{2,g}(\bar{Q}_2) - L_{2,g}(\bar{Q}_{20})\}$$
$$= -\sum_{a(0)} P_0\left[\left\{\bar{Q}_{20}^{\eta} \log\left(\bar{Q}_2^{\eta}\right) + (1-\bar{Q}_{20}^{\eta})\log\left(1-\bar{Q}_2^{\eta}\right)\right\}(a(0), V_{a(0)}(1))\right]$$
$$P_0\{L_{1,d_{0,A(1)},g}(\bar{Q}_1) - L_{1,d_{0,A(1)},g}(\bar{Q}_{10})\}$$
$$= -P_0\left[\left\{\bar{Q}_{10}^{\eta} \log\left(\bar{Q}_1^{\eta}\right) + (1-\bar{Q}_{10}^{\eta})\log\left(1-\bar{Q}_1^{\eta}\right)\right\}(V(0))\right].$$

We can add weight functions h_2 and h_1 as in the previous example and the loss functions are still valid. □

3.2 Performance of rule

We now describe a risk function which sequentially targets the performance of the fitted rule in terms of mean outcome as is done in the Q-learning literature. By definition, $d_0 = \arg\max_{d\in\mathcal{D}} E_{P_0,d}Y_d$. It follows immediately that $-E_{P_0}Y_d$ is a valid risk function for a candidate rule d. In our companion paper we discuss two estimates of $-E_{P_0}Y_d$. Rather than restate these results, we state a single theorem which summarizes these findings. We refer the reader to the companion paper for a thorough discussion of the proposed methods.

Define:

$$\widetilde{L}_{2,g}(d_{A(1)})(O) = -\frac{A_2(0)}{g_{A(0)}(O)}\frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)}Y. \tag{11}$$

Let $d_{A(1)}$ be a treatment rule for the second time point. Define:

$$\widetilde{L}_{1,d_{A(1)},g}(d_{A(0)})(O) = -\frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)}\frac{I(A(0) = d_{A(0)}(V(0)))}{g_{A(0)}(O)}Y$$

Theorem 3. *Suppose the positivity assumption holds at g_0. Then:*

$$P_0\left\{\widetilde{L}_{2,g_0}(d_{A(1)}) - \widetilde{L}_{2,g_0}(d_{0,A(1)})\right\} = \sum_{a(0)} P_0 I\left(d_{A(1)} \neq d_{0,A(1)}\right)\left|\bar{Q}_{20}\right|(a(0), V_{a(0)})$$

$$P_0\left\{\widetilde{L}_{1,d_{0,A(1)},g_0}(d_{A(0)}) - \widetilde{L}_{1,d_{0,A(1)},g_0}(d_{0,A(0)})\right\} = P_0 I\left(d_{A(0)} \neq d_{0,A(0)}\right)\left|\bar{Q}_{10}\right|(V(0))$$

where the sum is over $a(0) \in \{0,1\} \times \{1\}$. It follows that:

$$d_{0,A(1)} = \arg\min_{d_{A(1)}} P_0\widetilde{L}_{2,g_0}(d_{A(1)})$$
$$d_{0,A(0)} = \arg\min_{d_{A(0)}} P_0\widetilde{L}_{1,d_{0,A(1)},g_0}(d_{A(0)})$$

A double robust generalization of the above theorem appears with proof in the appendix.

To estimate $P_0\widetilde{L}_{2,g}(d_{A(1)})$ and $P_0\widetilde{L}_{1,d_{0,A(1)},g}(d_{A(0)})$ we recommend using either the empirical distribution, which results in an empirical risk minimization problem, or the CV-TMLE's of the respective parameters, which we describe in Section 5.

Though the risks described in this section are desirable in that they directly target the measure of performance of interest, the computational tractability of these problems needs to be explored further. In the next section we show that the optimization problems associated with the above risks can be solved by empirical risk minimization using weighted $0-1$ loss functions.

3.3 Weighted classification

We now show that maximizing $E_{P_0}Y_d$ can be viewed as a risk minimization problem resulting from using a weighted $0-1$ loss function. This result is a longitudinal extension to that of Zhang et al. (2012). We then show that a rich class of smooth surrogate loss functions can be used to improve computational tractability. We will use the definitions of D_1 and D_2 from the Section 3.1.

Let $Z:\mathbb{R}\rightarrow\mathbb{R}$ represent the function $Z(x)=I(x\geq 0)$. Define:

$$\begin{aligned}
K_{2,g}(O) &= \frac{A_2(0)}{g_{A(0)}(O)}D_2(g)(O)\\
\widehat{L}_{2,g}(d_{A(1)})(O) &= |K_{2,g}(O)|I\left(d_{A(1)}(A(0),V(0))\neq(Z\circ K_{2,g}(O),1)\right),
\end{aligned}$$

where $\circ$ denotes function composition. For some fixed $d_{A(1)}$, define:

$$\begin{aligned}
K_{1,d_{A(1)},g}(O) &= \frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}D_1(d_{A(1)},g))(O)\\
\widehat{L}_{1,d_{A(1)},g}(d_{A(0)})(O) &= |K_{1,d_{A(1)},g}(O)|I\left(d_{A(0)}(V(0))\neq(Z\circ K_{1,d_{A(1)},g}(O),1)\right).
\end{aligned}$$

The following theorem shows that the optimal rule can be learned through a sequential classification problem using $Z\circ K_{2,g}(O)$ and $Z\circ K_{1,d_{A(1)},g}(O)$ as outcomes and weighted $0-1$ loss functions with weights $|K_{2,g}|$ and $|K_{1,d_{A(1)},g}|$, where the weights respectively do not rely on $d_{A(1)}$ or $d_{A(0)}$, i.e. the current rule to the routine aims to learn.

Theorem 4. *Suppose the positivity assumption holds at g and g_0. Then for any $(d_{A(0)},d_{A(1)})\in\mathcal{D}$:*

$$\begin{aligned}
\widehat{L}_{2,g}(d_{A(1)}) &= \widetilde{L}_{2,g}(d_{A(1)})+C_{2,g}\\
\widehat{L}_{1,d_{A(1)},g}(d_{A(0)}) &= \widetilde{L}_{1,d_{A(1)},g}(d_{A(0)})+C_{1,d_{A(1)},g}
\end{aligned}$$

where $C_{2,g}(O)$ and $C_{1,d_{A(1)},g}(O)$ do not rely on $d_{A(1)}$ or $d_{A(0)}$, respectively. It follows that $\widehat{L}_{2,g}$ and $\widehat{L}_{1,d_{0,A(1)},g}$ are valid loss functions for sequentially estimating $d_{0,A(1)}$ and $d_{0,A(0)}$ if $g = g_0$.

A double robust generalization of the above theorem appears with proof in the appendix. The above theorem shows that the weighted classification losses yield the same loss-based dissimilarities as the corresponding mean performance based losses.

We now present a simple result which motivates future work to apply general results on surrogate loss functions like those in Bartlett et al. (2006) to the above weighted classification problem. Zhang et al. (2012) present a specific result with a weighted hinge loss function in the single time point case. The result below can be extended naturally using the methods in Bartlett et al., but the result below covers many interesting cases.

Theorem 5. *Suppose the positivity assumption holds at g and $0 < E|K_{2,g}(O)| < \infty$. Let $\phi : \mathbb{R} \to [0, \infty)$ be some convex function that is differentiable at 0 with $\phi'(0) < 0$. Define:*

$$L_{2,\phi,g}(f)(O) = |K_{2,g}(O)|\, \phi \left(f(A(0), V(1))(2Z \circ K_{2,d_{A(1)},g}(O) - 1) \right)$$

for some latent function f with range $\mathbb{R}$. Let f_i be some sequence of functions and $d_{A(1),i}$ be a sequence of functions such that $d^{(i)}_{A(1)}(A(0), V(1))$ gives treatment $I(f_i(A(0), V(1)) \geq 0)$ without censoring. Then:

$$P_0 L_{2,\phi,g}(f_i) \overset{i\to\infty}{\longrightarrow} \inf_{\tilde{f}} P_0 L_{2,\phi,g}(\tilde{f}) \implies P_0 \widehat{L}_{2,g}(d^{(i)}_{A(1)}) \overset{i\to\infty}{\longrightarrow} P_0 \widehat{L}_{2,g}(d_{0,A(1)}),$$

where the infimum is over all measurable functions $\tilde{f}$ that take $A(0), V(1)$ as input.

Examining the proof of the theorem in the appendix shows that an analogous result holds for $\widehat{L}_{1,d_{0,A(1)},g}$.

If g is correctly specified and the infimum of $P_0 L_{2,\phi,g}(\cdot)$ is achievable at some f^* then it follows immediately that $d_{0,A(1)}$ has the same performance as the rule $(A(0), V(1)) \to I(f^*(A(0), V(1)) > 0)$ under $P_{0,a(0)}$. This shows that weighted surrogate loss functions are valid for $d_{0,A(1)}$. A sufficient condition for $E|K_{2,g}(O)| < \infty$ is that g satisfies the strong positivity assumption and Y is bounded.

Because nonnegatively weighted linear combinations of convex functions are convex, $L_{2,\phi,g}$ is necessarily convex. Thus the proposed procedure yields an empirical risk that is easy to minimize via convex optimization techniques. In the appendix we show that the double robust extension of the above result holds. Thus, unlike the double robust blip function approach, the double robust weighted classification loss functions lead to a straightforward optimization routine in the longitudinal setting.

Though a convex surrogate loss has many desirable properties, recent advances in classification shows that there are also many interesting nonconvex surrogate losses (see, e.g., Masnadi-Shirazi and Vasconcelos (2009)). It would be interesting to extend these results to weighted classification and explore their performance.

Algorithm 2 shows how to learn a latent function f_{2n} for a particular surrogate loss L_{2,ϕ,g_0} when the intervention mechanism are taken to be known. The extension to the case where g_0 is unknown follows the same approach as for the blip function based loss functions (Section 3.1).

Algorithm 2 Weighted surrogate loss based estimation of $d_{0,A(1)}$

function WeightedSurrogate(O_1,...,O_n)
  for $i = 1$ to n **do**
    Assign $\frac{A_2(0)_i}{g_{0,A(0)}(O_i)} A_2(1)_i \frac{2A_1(1)_i - 1}{g_{0,A(1)}(O_i)} Y_i$ to element i of the vector K_{2,g_0}
  Run a classification algorithm to predict $(I(K_{2,g_0,i} \geq 0) : i = 1, ..., n)$ using predictors $((A(0)_i, V(1)_i) : i = 1, ..., n)$ and loss function L_{2,ϕ,g_0}, where observation i receives weight $|K_{2,g_0,i}|$. The goal is to find f_{2n} in some class $\mathcal{F}$ such that:
  $f_{2n} = \arg\min_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} |K_{2,g_0,i}| L_{2,\phi,g_0}(f)(O_i)$.
  return f_{2n}

We close this section with two examples of valid weighted surrogate loss functions, and refer the reader to Bartlett et al. (2006) for more examples. One can verify that ϕ is convex and differentiable at 0 in both of these examples.

Example 3. Weighted log loss: $\phi(x) = \log(1 + e^{-x})$. Then:

$$L_{2,\phi,g}(f)(O) = |K_{2,g}(O)| \log\left(1 + e^{-f(A(0),V(1))(2Z \circ K_{2,g}(O) - 1)}\right).$$

One can show that the above loss is closely related to the loss function in Example 2 but can yield different estimated rules. The unweighted log loss directly estimates class probabilities. This partly explains the similarities between the above and certain losses that are calibrated to estimate the sequential blip functions. □

Example 4. Weighted hinge loss: $\phi(x) = \max(1-x, 0)$. Then:

$$L_{2,\phi,g}(f)(O) = |K_{2,g}(O)| \max\left(1 - f(A(0), V(1))[2Z \circ K_{1,d_{A(1)},g}(O) - 1], 0\right).$$

Unlike the previous example, optimizing the unweighted hinge loss does not directly estimate class probabilities. Nonetheless, the above loss can be related to the weighted log loss using the notion of a soft maximum (Cook, 2011). □

4 Sequential super-learning of the optimal rule

We now present an ensemble method called super-learning that combines candidate estimators of the optimal rule at a particular time point into a single estimated rule for that time point. At each time point, the final estimator satisfies an oracle inequality stating that it will asymptotically perform at least as well as the best convex combination of candidates in the library in terms of loss-based dissimilarity under mild conditions. The super-learner methodology allows for data adaptive candidate estimators, by which we mean estimators that are consistent over a large semi-parametric model. Our super-learner can select the best resolution for a data set based on data adaptive and parametric candidate estimators.

4.1 Second time point

For the sake of presentation we will present these results for IPCW loss functions in an RCT without missingness, but the oracle inequalities for the double robust losses are straightforward extensions of the results in this section. At the end of this section we give examples of loss functions that satisfy the conditions for the oracle inequality derived from the blip function, mean performance, and weighted classification approaches.

We start by introducing the notation used in this section. Let $B_n \in \{0,1\}^n$ denote a random split of the data into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$ so that np_n of the elements in each realization of B_n have value 1 for some $p_n \in (0,1)$. Let P^0_{n,B_n} and P^1_{n,B_n} denote the corresponding empirical distributions of these two complementary subsamples. Section 2.4.1 of Dudoit and van der Laan (2005) shows that many commonly used cross-validation procedures yield sample splitting formulations in terms of a random variable B_n.

Let $\hat{f}_{2,j}$, $j = 1, \ldots, J$, denote an estimator of a latent function that takes a distribution P as input and outputs an estimate $\hat{f}_{2,j}(P)$ for which the indicator that $\hat{f}_{2,j}(P)(A(0), V(1))$ is nonnegative gives an estimate of the optimal rule $d_{0,A(1)}$ evaluated at $(A(0), V(1))$. In the examples following Theorem 6 we suppose that all of the latent functions under consideration have bounded range, which can be accomplished in practice via truncation. We do not expect reasonable truncation (not truncating too close to 0) of the latent function to greatly impact the mean performance of the fitted rule because it will not change the sign of the candidate.

All loss functions discussed in the previous section yield a latent function representation. For blip function based losses, the blip function itself can be taken as the latent function. For the mean performance of the estimated rule, $2d-1$ can be used as the latent function, where d represents the value estimated rule at a particular point. For the weighted classification approach, either the latent function resulting from many classification methods can be used directly or an artificial latent function can be constructed as for the mean performance loss.

We first give a general oracle inequality as presented in van der Laan et al. (2007) for estimating $d_{0,A(1)}$. Let α_n fall in a grid G_n of $K(n)$ points on Δ_{J-1}, where Δ_{J-1} represents the $(J-1)$-simplex is the set of all $\alpha \in [0,1]^J$ such that $\sum_j \alpha_j = 1$. The given setup yields the following finite sample result.

Theorem 6. *Let L_{g_0} be some loss function that relies on g_0 which takes as input a function $f : \mathcal{A}(0) \times \mathcal{V}(1) \to \mathbb{R}$ and yields a function of O. Let $f_{20} = \arg\min_f P_0 L_{g_0}(f)$. Suppose that:*

$$\sup_f \sup_{o \in O} |L_{g_0}(f)(o) - L_{g_0}(f_{20})(o)| < \infty \tag{12}$$

$$\sup_f \frac{Var_{P_0}\left(L_{g_0}(f)(O) - L_{g_0}(f_{20})(O)\right)}{E_{P_0}\left[L_{g_0}(f)(O) - L_{g_0}(f_{20})(O)\right]} < \infty. \tag{13}$$

where the supremums are over all measurable functions $f : \mathcal{A}(0) \times \mathcal{V}(1) \to \mathbb{R}$ and we take $0/0 = 0$. For all $\alpha \in \Delta_{J-1}$, define $\hat{f}_{2,\alpha}(P) = \sum_{j=1}^J \alpha_j \hat{f}_{2,j}(P)$. For a fixed sample of size n, define:

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

Then:

$$\begin{aligned} &E_{P_0^n} E_{B_n} P_0 \{L_{g_0}(\hat{f}_{2,\alpha_n}(P^0_{n,B_n})) - L_{g_0}(f_{20})\} \\ &\quad \le (1+\lambda) E_{P_0^n} \min_{\alpha \in G_n} E_{B_n} P_0 \{L_{g_0}(\hat{f}_{2,\alpha}(P^0_{n,B_n})) - L_{g_0}(f_{20})\} + C(\lambda) \frac{\log K(n)}{n p_n} \end{aligned}$$

for all $n \in \mathbb{N}$ and $\lambda > 0$, where $C(\lambda) \ge 0$ is a constant that may rely on P_0 and P_0^n represents the distribution of the observed n i.i.d. draws from P_0.

The above theorem is a special case of Corollary 3.2 in van der Laan et al. (2006) so the proof is omitted. In this article we focus on V' cross-validation. Note the distinction between V' and V from the V-optimal rule in the notation. In V'-fold cross-validation, the data is split into V' mutually exclusive and exhaustive sets of size approximately n/V' uniformly at random. Each set is then used as the validation set once, with the union of all other sets serving as the training set. The fact that n may not be divisible by V' so that the validation sets are not all exactly the same size will not matter asymptotically and will make little difference in finite samples.

We can choose G_n so that any point on the simplex can be arbitrarily well approximated by a point on the grid of polynomial size $K(n)$ asymptotically. Given a Lipschitz condition on the loss-based dissimilarity, the approximation error by using points on $K(n)$ instead of the entire simplex is asymptotically negligible. Such a Lipschitz condition will hold under bounding conditions for all loss functions previously discussed except the mean outcome and weighted $0-1$ loss functions. For these two losses we posit that the finite sample result on the grid G_n still gives a useful asymptotic result over the entire simplex under reasonable conditions, e.g. consistency of at least one of the candidates.

The limiting result is referred to as an oracle inequality because we asymptotically do as well as the oracle in selecting α (up to an almost parametric $O(\log n/n)$ term) in terms cross-validated loss-based dissimilarity averaged across training samples. Letting V' go to infinity at a slow enough rate also shows that we do a well as the oracle who can see the entire data set rather than just training samples, again up to the $O(\log n/n)$ term. Thus the ratio of loss-based dissimilarities converges to 1 as $n \to \infty$ whenever the oracle dissimilarity converges slower than $\log n/n$.

The methodology shows that there is no need to *a priori* decide on a single loss function or algorithm to fit the optimal rule – simply including all candidate methods of interest in the super-learner library guarantees that we asymptotically do at least as well as the best of the algorithms in terms of the cross-validated loss-based dissimilarity resulting from the chosen L_{g_0}.

In Algorithm 3 we describe how to implement the super-learner algorithm using V'-fold cross-validation for a given data set a collection of prediction algorithms $\hat{f}_{2,1}, ..., \hat{f}_{2,J}$. Let L_{g_0} be a loss function satisfying the conditions of Theorem 6. For simplicity we assume that n is divisible by V'. Rather than optimize for α_n over G_n, we recommend (approximately) optimizing over the entire simplex.

For an observational study or an RCT with an unknown censoring mechanism, an estimate $g_{n,v}$ of g_0 can be estimated each training sample $v = 1, ..., V'$.

Algorithm 3 Super-learner estimation of $d_{0,A(1)}$

function SUPERLEARNER($O_1,...,O_n,\hat{f}_{2,1},...,\hat{f}_{2,J}$)
    Let F be a randomly ordered vector of length n containing n/V' 1s, n/V' 2s, ..., n/V' V's
    Initialize an empty matrix X of dimension $n \times J$
    for $v = 1$ to V' **do**
        for $j = 1$ to J **do**
            Fit the estimate $f_{2,v,j}$ by running $\hat{f}_{2,j}$ on the set $\{O_i : F_i \neq v\}$
            For all i such that $F_i = v$, let $X_{i,j} = f_{2,v,j}(A(0)_i, V'(1)_i)$
    Run an optimization routine to solve:

$$\alpha_n = \arg\min_{\alpha \in \Delta_{J-1}} \sum_{v=1}^{V'} \sum_{i:F_i=v} L_{g_0}\left(\sum_{j=1}^{J} \alpha_j X_{i,j}\right)$$

    for $j = 1$ to J **do**
        Fit the estimate f_j by running $\hat{f}_{2,j}$ on the $\{O_i : i = 1, ..., n\}$
    return $f_{\alpha_n} \equiv \sum_{j=1}^{J} \alpha_{n,j} f_j$

We then let:

$$\alpha_n = \arg\min_{\alpha \in \Delta_{J-1}} \sum_{v=1}^{V'} \sum_{i:F(i)=v} L_{g_{n,v}}\left(\sum_{j=1}^{J} \alpha_j f_{2,v,j}\right)(O_i). \tag{14}$$

We have described oracle inequalities showing that we asymptotically estimate the best candidate rule at each time point given our sample, subject to the implementation of possibly suboptimal rules at future time points. Getting an oracle result in terms of the mean under the optimal rule of the entire treatment regime is desirable, but we have not shown whether or not such a result holds in this section. One cannot fit the convex combinations at both time points simultaneously when the optimal rule is learned sequentially because the candidates at the first time point rely on the convex combination at the second time point.

We now give examples of loss functions to which the inequality in Theorem 6 can be applied. Each of the below examples makes use of a subset of following assumptions (each example specifies the subset it uses). The fourth assumption is only used in Example 5 and is discussed there.

A1. $Pr_{P_0}(|Y| < M) = 1$ for some $M < \infty$.

A2. The strong positivity assumption holds at g_0 for some $\delta > 0$.

A3. Each of the estimators in the candidate library produces estimates of uniformly bounded range, where the uniformity is over input distributions P.

A4. There exists some constant $c > 0$ that may rely on P_0 such that $|\bar{Q}_{20}(A(0), V_{a(0)}(1))| \geq cE_{P_0}[Y^2_{a(0),a(1)}|V_{a(0)}]$ almost surely with respect to the distribution in which the first treatment is set to $a(0)$ for $a(0), a(1) \in \{0,1\} \times \{1\}$.

Example 1 (continued)**.** Squared error loss. We consider the unweighted case so that $h_2 = 1$. If A1, A2, and A3 then $|L_{2,g_0,MSE}|$ is uniformly bounded and thus satisfies (12). For all f, it can be shown that:

$$
\begin{aligned}
Var_{P_0}(L_{2,g_0,MSE}(f) - L_{2,g_0,MSE}(\bar{Q}_{20})) &\leq E_{P_0}\left(L_{2,g_0,MSE}(f) - L_{2,g_0,MSE}(\bar{Q}_{20})\right)^2 \\
&\leq M_1 E_{P_0}\left[L_{2,g_0,MSE}(f) - L_{2,g_0,MSE}(\bar{Q}_{20})\right],
\end{aligned}
$$

where $M_1 = \sup_f \sup_{o\in\mathcal{O}} \left(2D_2(g_0)(o) - (f + \bar{Q}_{20})(a(0), v(1))\right)^2 \geq 0$ is bounded by the stated assumptions. Thus the condition in (13) holds. □

Example 3 (continued)**.** Weighted log loss. Suppose A1, A2, and A3. It follows that $|K_{2,g_0}|$ is almost surely bounded. These conditions immediately show that (12) holds. The result is obvious if $E_{P_0}|K_{2,g_0}| = 0$, so suppose $E_{P_0}|K_{2,g_0}| > 0$. To show that (13) holds, one can use a similar change of measure argument as the one applied in the proof of our Theorem 5 to account for the weighting and apply Corollary 5.4 in van der Laan et al. (2006) to:

$$\phi(x) = \log(1 + e^{-x}) = -\log\left(\frac{1}{1 + e^{-x}}\right).$$

□

Example 5. Mean performance. Define:

$$L_{g_0}(f)(O) = -\frac{A_2(0)}{g_{A(0)}(O)} \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} Y,$$

where we have modified the definition in (11) so that L_{g_0} depends directly on the latent function. Suppose A1, A2, and A4. The loss-based dissimilarity representation in Theorem 3 shows that $\widetilde{L}_{2,g_0}$ satisfies (12). We show that the stated conditions suffice for (13) in Appendix B.

Assumption A4 can be viewed as a margin condition that ensures that the classification problem is not too difficult. In particular, it requires that the strata-specific treatment effect be larger than both of the expected squared outcomes under the counterfactual distributions where $a(1)$ is fixed without censoring. For binary Y, this means that the absolute average treatment effect in each strata of $V_{a(0)}$ be larger than some fixed proportion of the counterfactual prevalence of the outcome in strata of $V_{a(0)}$ when we set $a(0)$ and $a(1)$. □

4.2 First time point

The approach for estimating the optimal rule at the first time point is entirely analogous to the second time point, with the caveat that it takes an estimate of $d_{0,A(1)}$ as a nuisance function. To incorporate the estimate of the nuisance function, we suggest using the same approach used to incorporate an estimate of g_0 when it is unknown as we do in (14). In particular, this means estimating the nuisance function $d_{0,A(1)}$ on training set v and using this estimate of the nuisance function to obtain an estimate of a latent function at the first time point based on training set v for each algorithm j. One can then learn the convex combination similarly to what is done in (14), and apply this convex combination to the candidates learned on the full data set, which take an estimate of $d_{0,A(1)}$ based on the entire data set as nuisance function. The rate of convergence of the estimated first time point rule to $d_{0,A(0)}$ will be upper bounded by the rate of convergence of the estimated second time point rule to $d_{n,A(1)}$, see Theorem 1 of van der Laan and Dudoit (2003) for a detailed exposition.

To estimate the nuisance function $d_{0,A(1)}$, we suggest using the super-learner procedure presented in Algorithm 3, leading to a nested cross-validation procedure. In terms of runtime, this can cost up to a factor of V. If this is a concern, one can simply use the estimate of $d_{n,A(1)}$ resulting from the entire data set as nuisance function for all folds. Such a practice is not advisable because it invalidates the oracle inequality and necessitates empirical process conditions on the candidates. In general we look to avoid such conditions since they limit the data adaptivity of the estimators. Thus we believe that an honest cross-validation scheme in which the candidate estimators are only functions of the training samples is extremely valuable for estimating rules in practice.

5 CV-TMLE of risk

The empirical risk estimates resulting from the loss functions provided in Section 3 are valid in the sense that they are minimized at the true optimal treatment regime. Nonetheless, the empirical risk resulting from the given loss functions (and the double robust losses presented in the appendix) are not substitution estimators and thus can fail to respect a key constraint of the model: the fact that the risk is bounded. To improve finite sample performance, we propose using a CV-TMLE based estimate of risk. The CV-TMLE is a substitution estimator and thus naturally respects the bounded nature of our data. The CV-TMLE was originally proposed in Zheng and van der Laan (2010). Diaz and van der Laan (2013) use a CV-TMLE to estimate the risk of the causal dose response curve. In our companion paper we presented a CV-TMLE for the cross-validated mean outcome under a fitted rule.

Here we present two CV-TMLE's. The first is a sequential CV-TMLE that estimates the risks resulting from Theorem 3. The second is a non-sequential CV-TMLE which aims to directly maximize the mean outcome under the fitted two time point rule, which is even more targeted towards our goal than the losses in Theorem 3.

To distinguish between the convex combinations for super-learner at the first and second time points in this section, we will use the notation $\alpha_{A(k)}$ for the convex combination used at time k, $k = 0, 1$.

5.1 Sequential CV-TMLE

Suppose we use the sequential negative mean performance risk function from Section 3.2 and conditions hold so that the risk is bounded. Consider selecting the convex combination $\alpha_{A(1)}$ and $\alpha_{A(0)}$ for the super-learner presented in the previous section when g_0 is known (e.g., in an RCT without missingness). Suppose the outcome Y is bounded. While the empirical risk is root-n consistent under conditions (and the double robust empirical risk is even asymptotically efficient under conditions), the given risk estimates may not respect the bounded nature of the data in finite samples.

Given an $\alpha_{A(1)}$, we can estimate the risk for the second time point rule indexed by this $\alpha_{A(1)}$. The CV-TMLE for the second time point is identical to the CV-TMLE presented in Appendix B.2 of our companion paper, with the exception that the covariate for ϵ_2 is replaced by

$$\frac{A_2(0)I(A(1) = I\left(\sum_j \alpha_{A(1),j} f_{2,v,j}(O) > 0)\right)}{g_0(O)},$$

and the covariate for ϵ_1 is replaced by $A_2(0)/g_{0,A(0)}(O)$. The conditions for the validity of the resulting risk estimate are not presented here, but are analogous to those presented in Diaz and van der Laan (2013). The CV-TMLE has the same double robustness and asymptotic efficiency properties as the cross-validated empirical mean of the double robust loss. For more details, we refer the reader to our companion paper.

Fitting the rule at the first time point is similar, with the covariate for ϵ_2 replaced by

$$\frac{I\left(A(0) = I(\sum_j \alpha_{A(0),j} f_{1,v,j}(O) > 0)\right) I\left(A(1) = d_{nv,A(1)}(A(0), V(1))\right)}{g_0(O)}$$

where $d_{nv,A(1)}$ is a nuisance parameter for the second time point rule learned only on training sample v. One could give empirical process conditions under which $d_{nv,A(1)}$ does not need to be learned only on training sample v, but we will not do so here. The covariate for ϵ_1 is then given by $I(A(0) = I(\sum_j \alpha_{A(0),j} f_{1,v,j}(O) > 0))/g_{0,A(0)}(O)$.

All risks presented in the previous examples are pathwise differentiable, and thus can yield CV-TMLE's of risk that may outperform the cross-validated empirical risk in finite samples.

5.2 Non-sequential super-learner targeted directly at mean outcome

We now sketch a non-sequential super-learner which seeks to maximize the mean outcome under the entire estimated rule $d_n = (d_{n,A(0)}, d_{n,A(1)})$. This estimator is a direct application of the CV-TMLE presented in Section 7.1 and Appendix B.2 of our companion paper. Suppose we have libraries of sequential candidate latent function estimators $(P \mapsto \hat{f}_{1,j}(P) : j = 1, ..., J_1)$ and $(P \mapsto \hat{f}_{2,j}(P) : j = 1, ..., J_2)$ for the first and second time points. The latent function estimators for the first time point rely on nuisance function fits for the second time point rule, but there is no requirement that this nuisance function be the same as the final output rule $d_{n,A(1)}$ at the second time point. For each fold v we can compute a sequential super-learner for the second time point on training set v, which yields an estimate $d^{nuis}_{n,v,A(1)}$ of $d_{0,A(1)}$. In learning each $d^{nuis}_{n,v,A(1)}$ we have estimated latent functions resulting from estimators $\hat{f}_{2,j_2}$, $j_2 = 1, ..., J_2$, applied to all of the training samples. We can get estimates resulting from applying the sequential estimators $\hat{f}_{1,j_1}$, $j_1 = 1, ..., J_1$, to each training set v, where the treatment at the second time point is set to $d^{nuis}_{n,v,A(1)}$.

We now have estimates resulting from estimators $\hat{f}_{1,j_1}$ and $\hat{f}_{2,j_2}$ applied to each training sample for all j_1, j_2. We can simultaneously optimize over $\alpha_{A(0)}$ and $\alpha_{A(1)}$ to maximize the CV-TMLE of the mean outcome under the fitted rule. The final estimated latent functions at the first and second time points are given by $\sum_j \alpha_{n,A(0),j}\hat{f}_{1,j}(P_n)$ and $\sum_j \alpha_{n,A(1),j}\hat{f}_{2,j}(P_n)$, respectively. This method seems to be most targeted towards our goal, namely maximizing the mean outcome under the estimated rule. We note that $\alpha_{A(1)}$ need not equal any of the convex combinations $\alpha^{nuis}_{v,A(1)}$ used to obtain each $d^{nuis}_{n,v,A(1)}$, but we can establish oracle inequalities that will ensure that $\alpha_{n,A(1)}$ performs at least as well as each $\alpha^{nuis}_{v,A(1)}$ in terms of mean outcome for the final output optimal rule. We leave deeper consideration of this cross-validation scheme to future work.

6 Simulation methods

Section 6.1 and Section 6.2 respectively introduce the data and methods for estimating the optimal rule d_0 in the one and two time point case.

6.1 Single time point

We start by presenting two single time point simulations. In our earlier technical report we directly describe the single time point problem (van der Laan, 2013). Here, we instead note that a single time point optimal treatment is a special case of a two time point treatment when only the second treatment is of interest. In particular, we can see this by taking $L(0) = V(0) = \emptyset$, estimating $\bar{Q}_{2,0}$ without any dependence on $a(0)$, and correctly estimating $\bar{Q}_{1,0}$ with the constant function zero. We can then let $I(A(0) = d_{n,A(0)}(V(0))) = 1$ for all $A(0), V(0)$ wherever the indicator appears in our calculations. Because the first time point is not of interest, we only describe $\bar{Q}_{2,0}$ and the second time point treatment mechanism for this simulation. We refer the interested reader to our earlier technical report for a thorough discussion of the single time point case.

6.1.1 Data

See Section 8.1.1 of our companion article. We remind the reader that static treatments (treating everyone or no one at the second time point) have approximately the same mean outcome of 0.464. The optimal rule has mean outcome $E_{P_0}Y_{d_0} \approx 0.536$ when $V(1) = W_3$ and the optimal rule has mean outcome $E_{P_0}Y_{d_0} \approx 0.563$ when $V(1) = (W_1, W_2, W_3, W_4)$.

6.1.2 Estimation methods

We assume that the treatment and censoring mechanisms are known. For ease of interpretation, we consider two estimates of $E_{P_0}\left[Y|\bar{A}(1),W\right]$: (i) a naive estimate of 1/2 for all $A(1), W$, and (ii) the true conditional expectation $E_{P_0}\left[Y|\bar{A}(1),W\right]$. We note that (i) is slightly different from an IPCW estimator in that it contains a term which stabilizes the inverse weighted outcome term in the (cross-validated) empirical or CV-TMLE estimate of risk. This stabilized approach should do slightly better in our simulation since the conditional mean of Y given $\bar{A}(1), W$ is approximately centered around 0.5. In practice we always recommend using a double robust approach, even if just an intercept-only best guess of the conditional mean as we do here. When the outcome is has mean $a \neq 0$, one can always (approximately) mean-center the outcome before estimating the optimal rule. This turns out to be equivalent to misspecifying $E_{P_0}\left[Y|\bar{A}(1),W\right]$ to be the constant function 1/2.

We estimate $\bar{Q}_{2,0}$ using both a misspecified parametric model and a library containg both parametric and machine learning methods. We always recommend using data adaptive methods to estimate $\bar{Q}_{2,0}$ in practice, but use the misspecified parametric model to demonstrate the robustness of using the mean outcome as the risk criterion. We only consider the misspecified parametric model when $V(1) = W_3$. In particular, we use the parametric fit:

$$\bar{Q}_{2,n}(a(0), w_3) \quad = \quad \beta_0 + \beta_1 w_3$$

where β is chosen to minimize either the empirical mean-squared error or the TMLE estimate of mean outcome. Neither of the risk estimates for the misspecified parametric model uses cross-validation. For the parametric fit we take the estimate of $E_{P_0}[Y|\bar{A}(1),W]$ to be the constant 1/2

We also use super-learner to estimate $\bar{Q}_{2,0}$. Table 2 shows the methods used from the SuperLearner package in R (Polley and van der Laan, 2012) and the corresponding estimating methodology with which they were estimated. The multivariate adaptive regression splines algorithm was only used for $V = W_1, ..., W_4$. We separately consider the candidates generated according to the squared error and surrogate log loss functions, and also consider a candidate library that includes both the squared error and surrogate log loss function methods.

To generate convex combinations of predictors we maximize the CV-TMLE or CV-DR-IPCW estimates of mean outcome (see our companion paper for a description of the estimating equation based CV-DR-IPCW estimator). We approximate solutions to the resulting optimization problems using the Subplex routine in the `nloptr` package in R (Ypma, 2014; Rowan, 1990). We use

Loss function	Method	R function
Squared error	Bayesian GLM	`SL.bayesglm`
	Generalized additive model	`SL.gam`
	Generalized linear model	`SL.glm`
	Generalized linear model, interactions	`SL.glm.interaction`
	Multivariate adaptive regression splines	`SL.earth`
	Sample mean	`SL.mean`
	Neural network	`SL.nnet`
	Stepwise regression	`SL.step`
	Forward stepwise regression	`SL.step.forward`
	Stepwise regression, interactions	`SL.step.interaction`
Weighted log	Generalized additive model	`SL.gam`
	Generalized linear model	`SL.glm`
	Generalized linear model, interactions	`SL.glm.interaction`
	Neural network	SL.nnet
	Recursive partitioning	SL.rpart

Table 2: Candidate estimators used to estimate $\bar{Q}_{2,0}$. See the SuperLearner package documentation for more details (Polley and van der Laan, 2012). SL.earth only used for $V = (W_1, ..., W_4)$.

thirty starting values selected randomly from the simplex to avoid sensitivity to intitial conditions, and also include the selection of α based on the weighted log loss criterion as an initial value. We also consider minimizing the cross-validated empirical risk functions derived from the squared error and weighted log loss functions. We do not truncate the latent functions, though we note only the empirical MSE blip function estimates can be unbounded, and this should not cause problems in our data set because the outcome is bounded. We compare the mean outcome under the rules generated by several combinations of candidate libraries and criteria for choosing the convex combination.

To evaluate the performance of the described methods we will use the mean performance of the estimated rule as a criterion, which is given by $E_{P_0} Y_{d_n}$ for a given rule d_n. We estimate E_{P_0} using 10^6 Monte Carlo simulations.

6.2 Two time points

Having already compared several different methodologies in the single time point setting, we use the two time point setting to show that our proposed method can sequentially learn a rule with good performance in practice.

6.2.1 Data

See Section 8.1.2 of the companion article. We remind the reader that static treatments yield mean outcomes $E_{P_0}Y_{(0,1),(0,1)} = 0.400$, $E_{P_0}Y_{(0,1),(1,1)} \approx 0.395$, $E_{P_0}Y_{(1,1),(0,1)} \approx 0.361$, and $E_{P_0}Y_{(1,1),(1,1)} \approx 0.411$. The true optimal treatment has mean outcome $E_{P_0}Y_{d_0} \approx 0.485$ when $V(0) = L(0)$ and $V(1) = (A(0), \bar{L}(1))$.

6.2.2 Estimation methods

As in the single time point case, we treat the intervention mechanism as known. As in the single time point case, we consider two stabilized classes risk estimates instead of the IPCW estimator (see Section 6.1.2). Rather than estimate $E_{P_0}[Y|\bar{A}(1), \bar{L}(1)]$ when estimating $d_{0,A(1)}$, we consider two extreme cases, namely plugging in either the truth or the constant function 1/2 for the desired expectation. Once the rule $d_{n,A(1)}$ at the second time point has been estimated, we estimate $E_{P_0}[Y_{d_{n,A(1)}}|A(0), L(0)]$ by either plugging in the truth, which can be computed analytically using the G-computation formula, or the constant function 1/2. In our simulations we only consider the cases where either both or neither of the sequential regressions is estimated correctly. All simulations use the IPCW mapping to relate the full data loss function to the observed data distribution (see Appendix A).

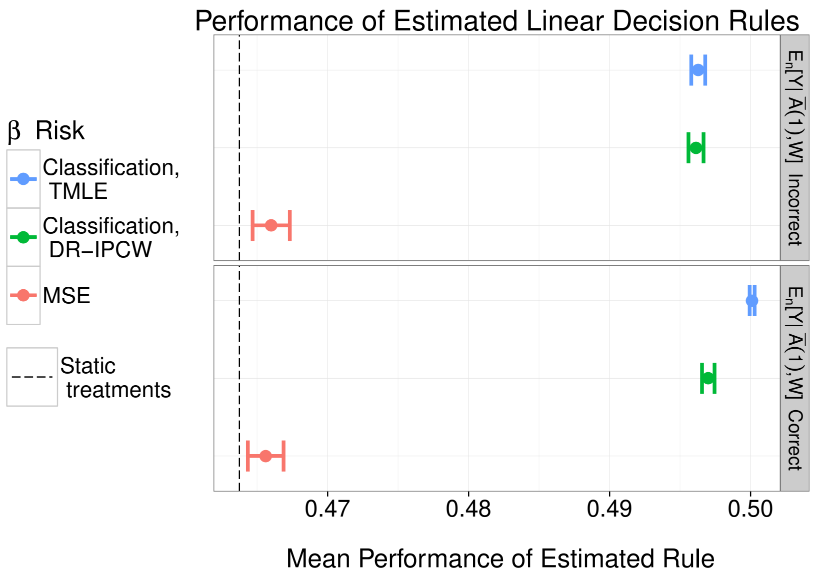
We use the candidate library in Table 2, with the exception that the Bayes GLM algorithm was excluded from these runs due to an occasional error from the software and the multivariate adaptive regression spline model was also excluded. The convex combinations for the sequential super-learners are selected using the cross-validated empirical risk resulting from the surrogate log loss function and the CV-TMLE estimate of the negative mean outcome risk. The weights $1/g_{0,A(0)}(O)$ and $I(A(1) = d_{n,A(1)}(O))/g_{0,A(1)}(O)$ were incorporated into the procedures for estimating $d_{0,A(1)}$ and $d_{0,A(0)}$ by weighting the candidate algorithms and the empirical risk optimization problem. The fitted rule $d_{n,A(1)}$ used to weight the losses for estimating $d_{0,A(0)}$ was not fitted on the training samples as we recommended in Section 4.2 due to time constraints.

7 Simulation results

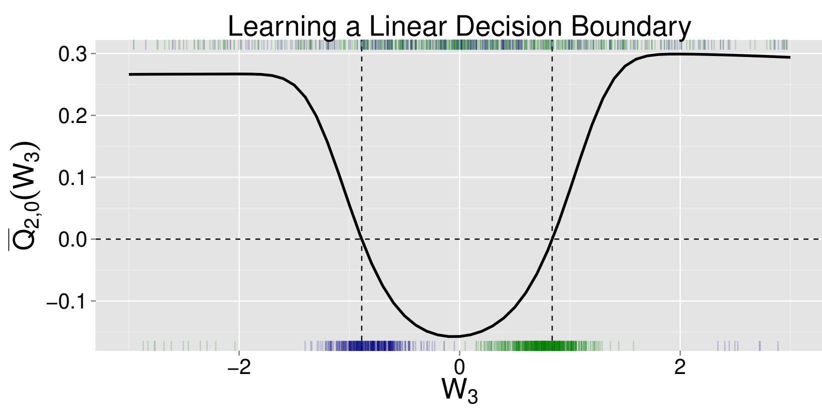
7.1 Single time point

7.1.1 Incorrectly specified parametric model, $V(1) = W_3$

Using the TMLE estimate of the mean outcome under the fitted rule as a risk estimate appears to be more robust to model misspecification than the



(a) Mean performance of the estimated rule when the estimate $E_n[Y|\bar{A}(1), W]$ of $E_{P_0}[Y|\bar{A}(1), W]$ is correctly and incorrectly specified. Both the TMLE and the DR-IPCW $-E_{P_0}Y_d$ risk estimates outperform the empirical mean-squared error risk criterion. Error bars indicate 95% confidence intervals to account for uncertainty from Monte Carlo draws.



(b) Linear boundaries learned by minimizing the TMLE of the negative mean outcome and the MSE when $E_{P_0}[Y|\bar{A}(1), W]$ is estimated with the constant 1/2. The x-intercepts from the least-squares fits are above the plot, and the x-intercepts from fits which maximize the expected mean outcome are below the plot. Green indicates positive slope, blue indicates negative slope. The TMLE nearly learns the linear classifier which maximizes the mean outcome under the fitted rule.

Figure 1

mean-squared error criterion.

Figure 1a shows that estimators which use the estimated mean outcome under the fitted rule as a the criteria to select β outperform the estimators which use the mean-squared error. Figure 1b demonstrates why the TMLE-based estimator outperforms the MSE-based estimator by such a large margin. In particular, we see that the near-quadratic shape of $\bar{Q}_{2,0}$ is not well-described by a linear fit. Nonetheless, linear classifiers which have x-intercepts near $W_3 = -1$ with negative slope or x-intercepts near $W_3 = 1$ with positive slope correctly estimate the optimal treatment in the interval between ± 1 and one of the two intervals $\pm(1, \infty)$. The TMLE-based estimator approximately learns one of these two decision boundaries, while the MSE-based estimator does not. In practice it is unlikely that such gross misspecification will occur for a one-dimensional $\bar{Q}_{2,0}$. Nonetheless, for more complex or higher dimensional $\bar{Q}_{2,0}$ it is likely that correctly specifying $\bar{Q}_{2,0}$ will be infeasible.

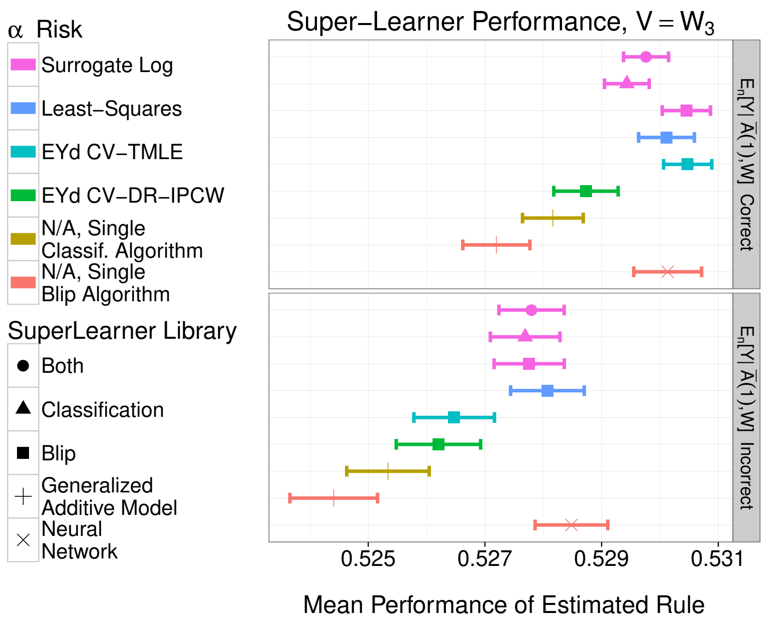
See the toy example presented in Qian and Murphy (2011) for another example of when using a classification approach performs better than using a blip function based approach.

7.1.2 Data adaptive methods

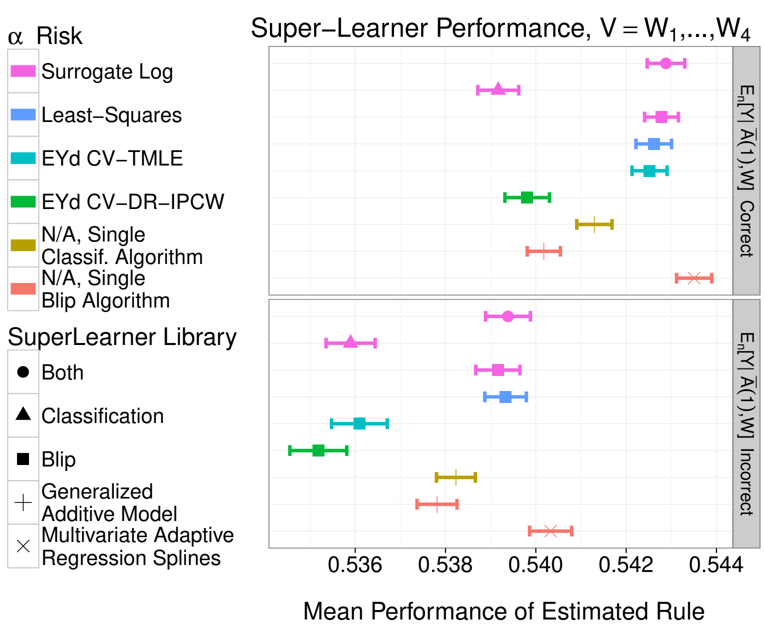
It remains to show that the mean outcome criterion performs well for selecting α when data adaptive methods are used to estimate $\bar{Q}_{2,0}$. Figure 2a and Figure 2b respectively give performance results of the super-learner based methods when $V(1) = W_3$ and $V(1) = W_1, ..., W_4$.

In this simulation the CV-TMLE for the mean outcome performs well when $E_{P_0}[Y|\bar{A}(1), W]$ is correctly specified, while the CV-DR-IPCW is outperformed by all other methods for selecting α regardless of the specification of $E_{P_0}[Y|\bar{A}(1), W]$. Combining both the weighted classification and the regression libraries perform well in all cases. The regression methods with the MSE risk criterion also performs well for all settings of our simulation. Correctly specifying the estimate of $E_{P_0}[Y|\bar{A}(1), W]$ improves performance for all candidate libraries and choices of the convex combination vector α. Comparing the weighted classification and blip function approaches is difficult given the different candidate library sizes, but both perform well overall.

Multivariate adaptive regression splines appear do the best of all algorithms in the super-learner library when $V(1) = W_1, ..., W_4$, though only slightly better than the super-learner fits which do not require *a priori* specification of a single algorithm. The super-learner outperformed all other algorithms in the candidate library. The super-learners perform similarly to the neural network algorithm when $V(1) = W_3$ and outperforms all other algorithms in



(a)



(b)

Figure 2: Mean performance of the estimated rule when the estimate $E_n[Y|\bar{A}(1), W]$ of $E_{P_0}[Y|\bar{A}(1), W]$ is correctly and incorrectly specified. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws in our simulation. (a) $V(1) = W_3$, (b) $V(1) = W_1, ..., W_4$.

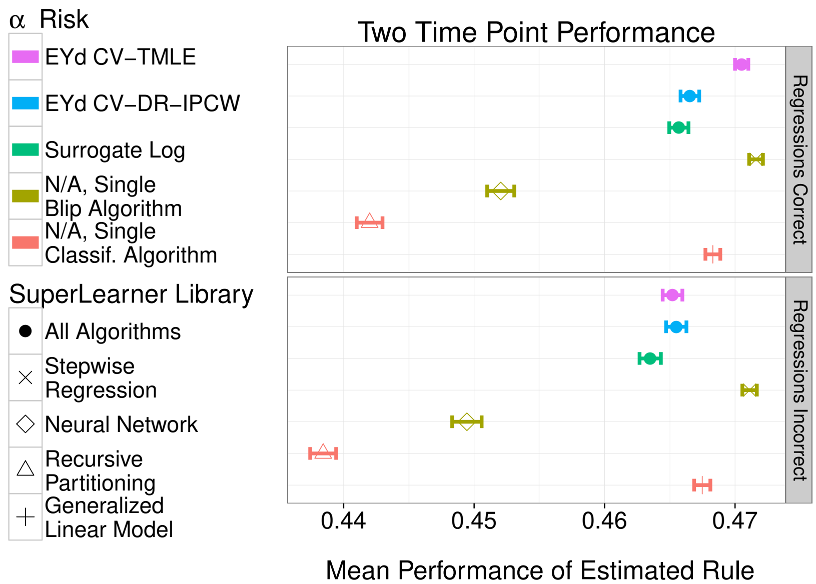


Figure 3: Mean performance of the estimated rule when $E_{P_0}[Y|\bar{A}(1), \bar{L}(1)]$ and $E_{P_0}[Y_d|A(0), L(0)]$ are specified correctly and incorrectly. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws in our simulation.

the candidate library.

All generalized linear model (GLM) methods performed poorly for all settings. For example, when a stepwise regression which includes interaction was used to estimate the blip function and $E_{P_0}[Y|\bar{A}(1), W]$ was correctly specified, the mean performance was respectively 0.465 and 0.483 when $V = W_3$ and $V = W_1, ..., W_4$. Thus here we see a setting where using data adaptive methods is important for good estimation of the optimal rule. Though we only show the generalized additive model in Figure 2, the super-learners outperformed all methods under consideration.

7.2 Two time points

Figure 3 shows that the the performance of several estimation methods in the two time point case. It appears that the optimal rule for our simulation can be well described by a generalized linear model. In particular, we see a stepwise regression with only main terms outperform all other methods under consideration, including our super-learners. Though the weighted classification based stepwise regression was not included in our model, we ran this algorithm alone to compare to the blip function based stepwise regression. The results were similar, with mean performance of approximately 0.470 for both settings considered.

Although the stepwise regression algorithm performed better for the given data generating distribution at this sample size, the super-learners which aim to maximize an estimate of the mean performance perform well overall. Note that some of the data adaptive methods, such as blip function based neural networks and classification based recursive partitioning perform poorly compared to the other methods. On average across the thousand runs, the super-learner which seeks to maximize the CV-TMLE of the mean outcome and has the conditional mean correctly specified gave the most weight at the first time point to the following algorithms: blip stepwise regression, 0.206; blip stepwise regression with interactions, 0.108; blip forward stepwise regression, 0.101; blip GLM, 0.080; and blip generalized additive model, 0.073. Thus our super-learner has naturally learned to select a linear as opposed to a more data adaptive estimator for the latent function.

The mean outcome based super-learners slightly outperformed the weighted log based super-learners in terms of mean performance for both settings.

8 Discussion

This article investigated nonparametric estimation of a V-optimal dynamic treatment. 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, this method is guaranteed to asymptotically outperform any competitor (with respect to loss-based dissimilarity) at each stage by simply including it in the library of candidate estimators. Some of the proposed sequential super-learners aim to minimize risks associated with learning some latent function which gives the optimal rule. One of these super-learners aims to optimize the performance of the fitted rule itself by maximizing the mean outcome. This seems to be more targeted towards our goal, but our theoretical claim suggests that stronger conditions are needed for the oracle inequality for this selector to hold.

Our simulation results support our theoretical findings. The super-learners always performed comparably to the best candidate in the library, and our theoretical results suggest that increasing sample size will improve their relative performance further. Further simulations are needed to fully understand the relationship between the weighted classification and blip function methods, and whether or not there are situations in which one will always perform better than the other. We demonstrated a misspecified linear classifier for which using the mean outcome criterion outperforms a misspecified linear blip function estimator in our simulation. We expect that such a situation can also occur with data adaptive methods, especially when none of the algorithms are correctly specified.

It would be interesting to compare the performance of our proposed super-learners against the fits of experts in the field who use a single fitting algorithm but perform variable selection and modify the tuning parameters based on the data. Such a (wo)man versus machine challenge can be done in practice with real or simulated data by proposing a fit on a training set and evaluating performance on a validation set. We expect that the human will tend to over- or underfit the data, while our proposed cross-validation method will select an appropriate level of smoothing. One might argue that the super-learner, which encourages using a large library of candidate estimators, requires an excessive amount of computing time. If runtime is a concern, one could use super-learners that use a rich class of parametric regression models with many variables (basis functions) as a candidate library. These algorithms can be optimized using stochastic gradient descent, so yield a computationally efficient super-learner algorithm. We leave it to the individual to decide how to best learn a dynamic treatment rule, but emphasize that theory, simulations, and the documented performance of super-learner algorithms in other contexts (see Introduction) suggest that our proposed method should perform well in almost any reasonably sized (not trivially small) sample.

In the current article we defined the treatment as binary at each time point. Consider now a treatment that has k possible values. We can then define a vector of binary indicators, ordered in a user-supplied manner, that identify the treatment. We can now apply the results for the multiple time-point treatment case in the appendix of our earlier technical report, since this represents a special case in which at some time-point there are no intermediate covariates between binary treatments (van der Laan, 2013). As a consequence, our results also apply to this case. Because the rate of convergence at each time point is upper bounded by the convergence rates at previously fitted time points, there may be better approaches when $\log_2 k \gg 1$. We leave such approaches to future work.

The sophistication of estimation and inference strategies for optimal treatment regimes has progressed dramatically in recent years thanks to the innovative work of many researchers. We look forward to continued statistical and computational advancements in this field, and to the eventual implementation of these treatment strategies on a large scale.

Acknowledgements

This research was supported by an NIH grant R01 AI074345-06. AL was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. The authors would like to thank Erin LeDell and Sam Lendle for their insights into developing a computationally efficient super-learner using stochastic gradient descent, and to thank Sam Lendle for providing them with code for Example 2. The authors would also like to thank the anonymous reviewers and Erica Moodie for their invaluable comments and suggestions to improve the quality of the paper.

Appendix

A Double robust loss functions

Below Q represents a parameter value, where the parameter maps from a distribution P to a collection of conditional distributions. Alternatively, we can set these estimates equal to 0 for IPCW-like risk estimates. We use Q_0 to denote the parameter mapping applied to P_0, i.e. the collection of conditional distributions under the observed data distribution P_0. All of the mappings used in this section only require expectations under the conditional distributions in Q. Thus in practice standard regression algorithms can be used to estimate the needed portions of Q_0. When we write conditional expectations under Q as E_Q, it will always be clear from context what parameter mapping (conditional distribution) of P_0 the appropriate part of Q is supposed to estimate.

Estimates for the optimal rule can be obtained using any regression or classification software, including data adaptive techniques. Because products of differences of Q and Q_0 and g and g_0 will serve as remainder terms for the final risk estimates, it is important to consistently estimate as many of these quantities of interest as possible, ideally at a reasonable rate. Note that the desire for consistent estimates of Q_0 likely precludes the use of parametric regressions for fitting Q, though parametric regressions can be taken as candidates in a cross-validation based algorithm such as SuperLearner. If known, any knowledge of Q_0 or g_0 may be incorporated into the estimates.

Throughout this section we introduce double robust versions of functions defined in the main text. Rather than introduce new notation to account for this, we simply add a Q next to the g in the notation, e.g. $D_2(g)$ becomes $D_2(Q, g)$ and $L_{2,g}$ becomes $L_{2,Q,g}$.

A.1 Blip functions

Define

$$\begin{aligned}D_2(Q,g)(O) &= A_2(1)\frac{2A_1(1)-1}{g_{0,A(1)}(O)}(Y - E_Q[Y \mid \bar{L}(1), \bar{A}(1))]\\ &+ E_Q[Y \mid \bar{L}(1), A(0), A(1) = (1,1)] - E_Q[Y \mid \bar{L}(1), A(0), A(1) = (0,1)],\end{aligned}$$

Let $L^F_{2,D_2(g)}(\bar{Q}_2)(O)$ denote a valid loss function for estimating $E_{P_{0,a(0)}}[D_2(Q,g) \mid V_{a(0)}(1) = v_{a(0)}(1))]$, in the sense that

$$(a(0), v(1)) \mapsto E_{P_{0,a(0)}}[D_2(Q,g)(O_{a(0)}) \mid V_{a(0)}(1) = v(1)]$$

minimizes

$$\sum_{\tilde{a}(0)\in\{0,1\}\times\{1\}} E_{P_{0,\tilde{a}(0)}}\left[L^F_{2,D_2(Q,g)}(\bar{Q}_2)(O_{\tilde{a}(0)})\right]$$

over all measurable functions $\bar{Q}_2$ of $a(0)$ and $v(1)$. Applying the DR-IPCW mapping (van der Laan and Dudoit, 2003) gives:

$$\begin{aligned}&L_{2,Q,g}(\bar{Q}_2)(O)\\ &\quad= \frac{A_2(0)}{g_{A(0)}(O)}\left(L^F_{2,D_2(Q,g)}(\bar{Q}_2) - E_Q\left[L^F_{2,D_2(Q,g)}(\bar{Q}_2) \mid A(0), L(0)\right]\right)\\ &\quad+ \sum_{a_1(0)=0}^{1} E_Q\left[L^F_{2,D_2(Q,g)}(\bar{Q}_2) \mid A(0) = (a_1(0), 1), L(0)\right], \qquad (15)\end{aligned}$$

We will use the sign of the $\bar{Q}_2$ which minimizes $L_{2,Q,g}$ to estimate $d_{0,A(1)}$. For a given $d_{A(1)}$, define

$$\begin{aligned}D_1(d_{A(1)},Q,g)(O) &= A_2(0)\frac{2A_1(0)-1}{g_{A(0)}(O)}\left(Y - E_Q\left[Y_{d_{A(1)}} \mid L(0), A(0)\right]\right)\\ &+E_Q\left[Y_{d_{A(1)}} \mid L(0), A(0) = (1,1)\right] - E_Q\left[Y_{d_{A(1)}} \mid L(0), A(0) = (0,1)\right].\end{aligned}$$

Let $L^F_{1,D_1(d_{A(1)},Q,g)}$ be some loss that satisfies:

$$E_{P_{0,d_{A(1)}}}\left[D_1(d_{A(1)},Q,g) \mid V(0) = \cdot\ \right] = \arg\min_{\bar{Q}_1} P_{0,d_{A(1)}} L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1),$$

Our proposed loss function is obtained by applying the DR-IPCW mapping to the above loss function:

$$\begin{aligned} L_{1,d_{A(1)},Q,g}(\bar{Q}_1)(O) &= \frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)} L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1) \\ &- \frac{I(A(1)=d_{A(1)}(A(0),V(1))}{g_{A(1)}(O)} E_Q\left(L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1) \mid \bar{A}(1),\bar{L}(1)\right) \\ &+ E_Q\left(L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1) \mid A(0),A(1)=d_{A(1)}(A(0),V(1)),\bar{L}(1)\right), \end{aligned} \qquad (16)$$

We now state a theorem that gives conditions under which the above loss functions allow us to learn the optimal rule d_0.

Theorem 2 (DR Version)**.** *Suppose the positivity assumption holds at g and g_0 and either $Q = Q_0$ or $g = g_0$. Then:*

$$\begin{aligned} &P_0\{L_{2,Q,g}(\bar{Q}_2) - L_{2,Q,g}(\bar{Q}_{20})\} \\ &\quad = \sum_{a(0)} P_{0,a(0)}\left(L^F_{2,D_2(Q,g)}(\bar{Q}_2) - L^F_{2,D_2(Q,g)}(\bar{Q}_{20})\right) \\ &P_0\{L_{1,d_{0,A(1)},Q,g}(\bar{Q}_1) - L_{1,d_{0,A(1)},Q,g}(\bar{Q}_{10})\} \\ &\quad = P_{0,d_{0,A(0)}}\left(L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_1) - L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_{10})\right), \end{aligned}$$

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

$$\begin{aligned} \bar{Q}_{20} &= \arg\min_{\bar{Q}_2} P_0 L_{2,Q,g}(\bar{Q}_2) \\ \bar{Q}_{10} &= \arg\min_{\bar{Q}_1} P_0 L_{1,d_{0,A(1)},Q,g}(\bar{Q}_1) \end{aligned}$$

The condition that $Q = Q_0$ can be weakened so that only the needed conditional expectations Q are equal to the analogous expectations under Q_0. We state a slightly stronger form of double robustness than stated in the above theorem in Section 9.1 of the earlier technical report (van der Laan, 2013). The stronger form shows that we have double robustness separately at each time point, so we need only have the portion of g_0 or that of Q_0 corresponding to each time point correctly specified. For example, we may have the intervention mechanism correctly specified at the first but not the second time point, but $L_{2,Q,g}$ is still a valid loss as long as the portion of Q corresponding to the second time point is correctly specified (even if Q is misspecified at the first time point!).

Proof of Theorem 2 *(DR Version)***.** Suppose $Q = Q_0$ or $g = g_0$. By the double robustness of DR-IPCW mapping:

$$E_{P_0} L_{2,Q,g}(\bar{Q}_2)(O) = \sum_{a(0)} E_{P_{0,a(0)}} L^F_{2,D_2(Q,g)}(\bar{Q}_2)$$

$$E_{P_0} L_{1,d_{0,A(1)},Q,g}\left(\bar{Q}_1\right) = E_{P_{0,d_{0,A(1)}}}\left[L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_1)\right].$$

All claims again follow immediately by the choice of $L^F_{2,D_2(Q,g)}$ and $L^F_{1,D_1(d_{0,A(1)},Q,g)}$.
□

Optimizing the double robust blip loss functions is not straightforward because of the final two terms in expressions in (15) and (16). Taking these terms to be 0, which is equivalent to misspecifying these needed conditional expectations under Q_0, allows for the use of weighted regression methods. We show in Section A.3 that optimizing the weighted classification losses does not encounter this difficulty.

A.2 Performance of rule

Define:

$$\begin{aligned}-\widetilde{L}^F_{2,Q,g}(d_{A(1)})(O) =& \frac{I(A(1) = d_{A(1)}(a(0), V(1)))}{g_{A(1)}(O)}(Y - E_Q[Y \mid \bar{L}(1), \bar{A}(1)]) \\ &+ E_Q[Y \mid \bar{L}(1), A(0), A(1) = d_{A(1)}(a(0), V(1))].\end{aligned}$$

Applying the DR-IPCW mapping (van der Laan and Dudoit, 2003) gives:

$$\begin{aligned}\widetilde{L}_{2,Q,g}(d_{A(1)})(O) =& \frac{A_2(0)}{g_{A(0)}(O)}\left(\widetilde{L}^F_{2,Q,g}(d_{A(1)})(O) - E_Q\left[\widetilde{L}^F_{2,Q,g}(d_{A(1)}) \mid A(0), L(0)\right]\right) \\ &+ \sum_{a_1(0)=0}^{1} E_Q\left[\widetilde{L}^F_{2,Q,g}(d_{A(1)}) \mid A(0) = (a_1(0), 1), L(0)\right].\end{aligned}$$

Let $d_{A(1)}$ be a treatment rule for the second time point. Define:

$$\begin{aligned}-\widetilde{L}^F_{1,d_{A(1)},Q,g}(d_{A(0)})(O) =& \frac{I(A(0) = d_{A(0)}(V(0)))}{g_{A(0)}(O)}(Y - E_Q[Y_{d_{A(1)}} \mid L(0), A(0)]) \\ &+ E_Q[Y_{d_{A(1)}} \mid L(0), A(0) = d_{A(0)}(V(0))].\end{aligned}$$

Applying the DR-IPCW mapping gives:

$$\widetilde{L}_{1,d_{A(1)},Q,g}(d_{A(0)})(O) = \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} \widetilde{L}^F_{1,d_{A(1)},Q,g}(d_{A(0)})$$
$$- \frac{I(A(1) = d_{A(1)}(A(0), V(1))}{g_{A(1)}(O)} E_Q\left(\widetilde{L}^F_{1,d_{A(1)},Q,g}(d_{A(0)}) \mid \bar{A}(1), \bar{L}(1)\right)$$
$$+ E_Q\left(\widetilde{L}^F_{1,d_{A(1)},Q,g}(d_{A(0)}) \mid A(0), A(1) = d_{A(1)}(A(0), V(1)), \bar{L}(1)\right)$$

Theorem 3 (DR Version)**.** *Suppose the positivity assumption holds at g and g_0 and either $Q = Q_0$ or $g = g_0$. Then:*

$$P_0\left\{\widetilde{L}_{2,Q,g}(d_{A(1)}) - \widetilde{L}_{2,Q,g}(d_{0,A(1)})\right\} = \sum_{a(0)} P_0 I\left(d_{A(1)} \neq d_{0,A(1)}\right)\left|\bar{Q}_{20}\right|(a(0), V_{a(0)}(1))$$

$$P_0\left\{\widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{A(0)}) - \widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{0,A(0)})\right\} = P_0 I\left(d_{A(0)} \neq d_{0,A(0)}\right)\left|\bar{Q}_{10}\right|(V(0))$$

where the sum is over $a(0) \in \{0,1\} \times \{1\}$. It follows that:

$$d_{0,A(1)} = \arg\min_{d_{A(1)}} P_0 \widetilde{L}_{2,Q,g}(d_{A(1)})$$
$$d_{0,A(0)} = \arg\min_{d_{A(0)}} P_0 \widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{A(0)})$$

Proof of Theorem 3 *(DR Version)***.** For all $d_{A(1)}$:

$$\begin{aligned}
&P_0\left(\widetilde{L}_{2,Q,g}(d_{A(1)}) - \widetilde{L}_{2,Q,g}(d_{0,A(1)})\right) \\
&= \sum_{a(0)} P_{0,a(0)}\left(\widetilde{L}^F_{2,Q,g}(d_{A(1)}) - \widetilde{L}^F_{2,Q,g}(d_{0,A(1)})\right) \\
&= \sum_{a(0)} P_{0,a(0)}\left(E_{P_{0,a(0)}}\left[\widetilde{L}^F_{2,Q,g}(d_{A(1)}) - \widetilde{L}^F_{2,Q,g}(d_{0,A(1)}) \mid V_{a(0)}\right]\right) \\
&= \sum_{a(0)} P_{0,a(0)} I\left(d_{A(1)} \neq d_{0,A(1)}\right)(a(0), V_{a(0)})\left|\bar{Q}_{20}(a(0), V_{a(0)})\right|,
\end{aligned}$$

where the sums are over $a(0) \in \{0,1\} \times \{1\}$. Because $|\bar{Q}_{20}| \geq 0$, the above is minimized at $d_{A(1)} = d_{0,A(1)}$. For any first time point treatment rule $d_{A(0)}$:

$$\begin{aligned}
&P_0\left\{\widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{A(0)}) - \widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{0,A(0)})\right\} \\
&= P_{0,d_{0,A(1)}}\left\{\widetilde{L}^F_{1,d_{0,A(1)},Q,g}(d_{A(0)}) - \widetilde{L}^F_{1,d_{0,A(1)},Q,g}(d_{0,A(0)})\right\} \\
&= P_0\left\{E_{P_{0,d_{0,A(1)}}}\left[\widetilde{L}^F_{1,d_{0,A(1)},Q,g}(d_{A(0)}) - \widetilde{L}^F_{1,d_{0,A(1)},Q,g}(d_{0,A(0)}) \mid V(0)\right]\right\} \\
&= P_0 I\left(d_{A(0)} \neq d_{0,A(0)}\right)(V(0))\left|\bar{Q}_{10}(V(0))\right|.
\end{aligned}$$

The above expression is minimized at $d_{A(0)} = d_{0,A(0)}$. $\square$

A.3 Weighted classification

We will use the definitions of Q, D_1, and D_2 from the Section A.1.

Define:

$$\begin{aligned} K_{2,Q,g}(O) =& \frac{A_2(0)}{g_{A(0)}(O)} \left(D_2(Q,g) - E_Q \left[D_2(Q,g) \mid A(0), L(0) \right] \right) \\ &+ \sum_{a_1(0)=0}^{1} E_Q \left[D_2(Q,g) \mid A(0) = (a_1(0), 1), L(0) \right]. \end{aligned}$$

Also define:

$$\widehat{L}_{2,Q,g}(d_{A(1)})(O) = |K_{2,Q,g}(O)| I(d_{A(1)}(A(0), V(0)) \neq (Z \circ K_{2,Q,g}(O), 1)).$$

Similarly, let:

$$\begin{aligned} K_{1,d_{A(1)},Q,g}(O) =& \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} D_1(d_{A(1)}, Q, g)) \\ &- \frac{I(A(1) = d_{A(1)}(A(0), V(1))}{g_{A(1)}(O)} E_Q \left(D_1(d_{A(1)}, Q, g) \mid \bar{A}(1), \bar{L}(1) \right) \\ &+ E_Q \left(D_1(d_{A(1)}, Q, g) \mid A(0), A(1) = d_{A(1)}(A(0), V(1)), \bar{L}(1) \right), \end{aligned}$$

and:

$$\widehat{L}_{1,d_{A(1)},Q,g}(d_{A(0)})(O) = |K_{1,d_{A(1)},Q,g}(O)| I(d_{A(0)}(V(0)) \neq (Z \circ K_{1,d_{A(1)},Q,g}(O), 1)).$$

We have the following theorem:

Theorem 4 (DR Version). *Suppose the positivity assumption holds at g and g_0. Then for any $(d_{A(0)}, d_{A(1)}) \in \mathcal{D}$:*

$$\begin{aligned} \widehat{L}_{2,Q,g}(d_{A(1)}) &= \widetilde{L}_{2,Q,g}(d_{A(1)}) + C_{2,Q,g} \\ \widehat{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) &= \widetilde{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) + C_{1,d_{A(1)},Q,g} \end{aligned}$$

where $C_{2,Q,g}(O)$ and $C_{1,d_{A(1)},Q,g}(O)$ do not rely on $d_{A(1)}$ or $d_{A(0)}$, respectively. It follows that $\widehat{L}_{2,Q,g}$ and $\widehat{L}_{1,d_{A(1)},Q,g}$ are valid loss functions for sequentially estimating $d_{0,A(1)}$ and $d_{0,A(0)}$ if either either $Q = Q_0$ or $g = g_0$.

Proof of Theorem 4 *(DR Version)***.** For all realizations $o \in \mathcal{O}$, define:

$$C_{2,Q,g}(o) = -\widetilde{L}_{2,Q,g}((Z \circ K_{2,Q,g}(o), 1))(o),$$

where $\widetilde{L}_{2,Q,g}((Z \circ K_{2,Q,g}(o), 1))$ represents $\widetilde{L}_{2,Q,g}$ evaluated at the static decision rule where everyone is given the treatment $Z \circ K_{2,Q,g}(o) \in \{0,1\}$ without censoring.

Checking all values of $d_{A(1)} \in \{0,1\} \times \{1\}$, $Z \circ K_{2,Q,g} \in \{0,1\}$, $a(0), a(1) \in \{0,1\}^2$ shows that:

$$|K_{2,Q,g}| I(d_{A(1)} \neq (Z \circ K_{2,Q,g}, 1)) - \widetilde{L}_{2,Q,g}(d_{A(1)}) = C_{2,Q,g}.$$

For the first time point, we define:

$$C_{1,d_{A(1)},Q,g}(o) = -\widetilde{L}^F_{1,d_{A(1)},Q,g}\left((Z \circ K_{1,d_{A(1)},Q,g}(o), 1)\right)(o).$$

Checking all values of $d_{A(0)} \in \{0,1\} \times \{1\}$, $Z \circ K_{1,d_{A(1)},Q,g} \in \{0,1\}$, and $a(0), a(1) \in \{0,1\}^2$ shows that:

$$|K_{1,d_{A(1)},Q,g}| I\left(d_{A(0)} \neq (Z \circ K_{1,d_{A(1)},Q,g}, 1)\right) - \widetilde{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) = C_{1,d_{A(1)},Q,g}.$$

The claim that $\widehat{L}_{2,Q,g}$ and $\widehat{L}_{1,d_{0,A(1)},Q,g}$ are valid loss functions for the sequential estimation of d_0 follows by the double robust version of Theorem 3. □

We close this section with a proof of Theorem 5 from the main text. A double robust extension of this result is straightforward.

Proof of Theorem 5. By the law of total expectation, for all $d_{A(1)}$ that set observations to uncensored:

$$\begin{aligned}&E_{P_0}\widehat{L}_{2,g}(d_{A(1)})\\ &\quad = E_{P_0}\left[E_{P_0}\left[|K_{2,g}| \mid A(0), V(1), Z \circ K_{2,g}(O)\right] I\left(d_{A(1),1}(A(0), V(1)) \neq Z \circ K_{2,g}\right)\right],\end{aligned}$$

where $d_{A(1),1}$ is the treatment index of the optimal rule. Let $\tilde{P}_0$ be the probability measure with:

$$\begin{aligned}&Pr_{\tilde{P}_0}\left((A(0), V(1), Z \circ K_{2,g}) \in B\right)\\ &\qquad = \frac{1}{E_{P_0}|K_{2,g}|}\int_B E_{P_0}\left[|K_{2,g}| \mid A(0), V(1), Z \circ K_{2,g}(O)\right] dP_0\end{aligned}$$

for all measurable sets B. Note that $\tilde{P}_0$ is a probability distribution over values of $A(0), V(1), Z \circ K_{2,g}$ and that $\tilde{P}_0$ is absolutely continuous with respect to P_0. Also note that

$$E_{P_0} \widehat{L}_{2,g}(d_{A(1)}) = E_{\tilde{P}_0} I\left(d_{A(1),1}(A(0), V(1)) \neq Z \circ K_{2,g}\right),$$

so we can now consider a simple $0-1$ loss under the distribution $\tilde{P}_0$.

By Theorem 4 in Bartlett et al., ϕ is classification-calibrated according to the definition in the paper. By part (c) of Theorem 3 in the same paper, it follows that:

$$\lim_{i\to\infty} \tilde{P}_0 \phi\left(f_i(A(0), V(1))(2Z \circ K_{1,d_{A(1)},g} - 1)\right) = \inf_{\tilde{f}} \tilde{P}_0 \phi\left(\tilde{f}(A(0), V(1))(2Z \circ K_{1,d_{A(1)},g} - 1)\right)$$

$$\Longrightarrow$$

$$\lim_{i\to\infty} \tilde{P}_0 I\left(I\left(f_i(A(0), V(1)) > 0\right) \neq Z \circ K_{2,g}\right) = \inf_{\tilde{f}} \tilde{P}_0 I\left(I\left(\tilde{f}(A(0), V(1)) > 0\right) \neq Z \circ K_{2,g}\right),$$

Writing the above expectations under $\tilde{P}_0$ as expectations under P_0 weighted by $d\tilde{P}_0/dP_0$ and multiplying by the constant $E_{P_0}|K_{2,g}|$ gives the desired result.

Examining the above proof shows that the conditions on ϕ can be weakened to the condition that ϕ is classification-calibrated according to the definition in Bartlett et al. (2006). □

B Example 5 proof

Proof that (13) holds in Example 5. Note that:

$$\begin{aligned}
&Var_{P_0}\left(L_{g_0}(f) - L_{g_0}(f_{20})\right)\\
&\quad \leq E_{P_0}\left(L_{g_0}(f) - L_{g_0}(f_{20})\right)^2\\
&\quad = E_{P_0}\left[I\left(I(f \geq 0) \neq I(f_{20} \geq 0)\right)(A(0), V(1)) \frac{A_2(0)}{g_{A(0)}(O)^2} \frac{A_2(1)}{g_{A(1)}(O)^2} Y^2\right]\\
&\quad \leq \delta^{-2} \sum_{a(0)} \sum_{a(1)} E_{P_0}\left[I\left(I(f \geq 0) \neq I(f_{20} \geq 0)\right)(a(0), V_{a(0)}(1)) Y^2_{a(0),a(1)}\right],
\end{aligned}$$

where the sums are over $\{0,1\} \times \{1\}$. For all $k_\delta > 0$, Theorem 3 shows that:

$$\begin{aligned}
&k_\delta Var_{P_0}\left(L_{g_0}(f) - L_{g_0}(f_{20})\right) - E_{P_0}\left[L_{g_0}(f) - L_{g_0}(f_{20})\right]\\
&\leq \max_{a(1)} \sum_{a(0)} E_{P_0}\left[I\left(I(f \geq 0) \neq I(f_{20} \geq 0)\right)\left(2k_\delta Y^2_{a(0),a(1)} - |\bar{Q}_{20}|\right)(a(0), V_{a(0)})\right]
\end{aligned}$$

where the maximum is over $a(1) \in \{0,1\} \times \{1\}$. By A4, we can choose $k_\delta > 0$ small enough so that $2k_\delta E_{P_0}[Y^2_{a(0),a(1)}|V_{a(0)}] - |\bar{Q}_{20}(a(0), V_{a(0)}(1))| \leq 0$ almost surely for all $a(0), a(1) \in \{0,1\} \times \{1\}$. The law of total expectation applied to the above then shows that, for $k_\delta > 0$ sufficiently small:

$$k_\delta Var_{P_0}\left(L_{g_0}(f) - L_{g_0}(f_{20})\right) - E_{P_0}\left[L_{g_0}(f) - L_{g_0}(f_{20})\right] \leq 0.$$

Condition (13) follows immediately, thus completing the proof. □

References

P L Bartlett, M I Jordan, and J D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

J D Cook. Basic properties of the soft maximum. *UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series, Working Paper 70*, 2011.

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.

S Dudoit and M J van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat Methodol*, 2(2):131–154, 2005.

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

R D Gill and J M Robins. Causal inference in complex longitudinal studies: continuous case. *Ann Stat*, 29(6):1785–1811, 2001.

Y Goldberg and M Kosorok. Q-learning with censored data. *Ann. Stat.*, 40: 529–560, 2012.

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

H Masnadi-Shirazi and N Vasconcelos. On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1049–1056. Curran Associates, Inc., 2009.

E Moodie, R Platt, and M Kramer. Estimating response-maximized decision rules with applications to breastfeeding. *J. Am. Stat. Assoc.*, 104:155–165, 2009.

E Moodie, B Chakraborty, and M Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.*, 40:629–645, 2012.

S Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:14551481, 2005.

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

J Neyman. Sur les applications de la thar des probabilites aux experiences Agaricales: Essay des principle (1923). Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed), Trans. *Statistical Science*, 5:463–472, 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.

D Ormoneit and S Sen. Kernel-based reinforcement learning. *Mach. Learn.*, 49:161–178, 2002.

J Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York, 2nd edition, 2009.

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):2097–2106, 2008.

E Polley and M J van der Laan. *SuperLearner: Super Learner Prediction*, 2012.

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.*, 39:1180–1210, 2011.

J Robins, L Orallana, and A Rotnitzky. Estimaton and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721, 2008a.

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. *Math Mod*, 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". *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 *Proceedings of the Biopharmaceutical Section*. 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. [Choice as an Alternative to Control in Observational Studies]: Comment. *Stat Sci*, 14(3):281–293, 1999.

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, New York, 2000.

J M Robins. Discussion of "Optimal dynamic treatment regimes" by Susan A. Murphy. *Journal of the Royal Statistical Society: Series B*, 65(2):355–366, 2003.

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

J M Robins, L Li, E Tchetgen, and A W van der Vaart. Higher order influence functions and minimax estimation of non-linear functionals. In *Essays in Honor of David A. Freedman*, IMS, Collections Probability and Statistics, pages 335–421. Springer New York, 2008b. ISBN DOI: 10.1214/193940307000000005, arXiv:0805.3040v1.

S Rosthøj, C Fullwood, R Henderson, and S Stewart. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Stat. Med.*, 88:4197–4215, 2006.

T H Rowan. Functional stability analysis of numerical algorithms. *Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin*, 1990.

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.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*, 66:688–701, 1974.

Donald B Rubin. *Matched sampling for causal effects.* Cambridge, Cambridge, MA, 2006.

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

M J van der Laan. Targeted Learning of an Optimal Dynamic Treatment and Statistical Inference for its Mean Outcome. Technical Report 317, UC Berkeley, 2013.

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 130, Division of Biostatistics, University of California, Berkeley, 2003.

M J van der Laan and A R Luedtke. Targeted Learning of the Mean Outcome Under an Optimal Dynamic Treatment Rule. Technical Report available at http://www.bepress.com/ucbbiostat/, Division of Biostatistics, University of California, Berkeley, under review at JCI, 2014.

M J van der Laan and M L Petersen. Causal Effect Models for Realistic Individualized Treatment and Intention to Treat Rules. *Int J Biostat*, 3(1): Article 3, 2007.

M J van der Laan and J M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York Berlin Heidelberg, 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, S Dudoit, and A W van der Vaart. The Cross-Validated Adaptive Epsilon-Net Estimator. *Stat Decis*, 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.

A W van der Vaart, S Dudoit, and M J van der Laan. Oracle Inequalities for Multi-Fold Cross-Validation. *Stat Decis*, 24(3):351–371, 2006.

J Ypma. The NLopt nonlinear-optimization package. 2014.

Z Yu and M J van der Laan. Measuring treatment effects using semiparametric models. Technical Report 136, Division of Biostatistics, University of California, Berkeley, 2003.

B Zhang, A A Tsiatis, M Davidian, M Zhang, and E Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Y Zhao, D Zeng, A Rush, and M Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67:1422–1433, 2011.

Y Zhao, D Zeng, A Rush, and M Kosorok. Estimating individual treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, 107:1106–1118, 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.