# Chapter 22
# Optimal Dynamic Treatment Rules

Alexander R. Luedtke and Mark J. van der Laan

Suppose we observe $n$ independent and identically distributed observations of a time-dependent random variable consisting of baseline covariates, initial treatment and censoring indicator, intermediate covariates, subsequent treatment and censoring indicator, and a final outcome. For example, this could be data generated by a sequential RCT in which one follows up a group of subjects, and treatment assignment at two time points is sequentially randomized, where the probability of receiving treatment might be determined by a baseline covariate for the first-line treatment, and time-dependent intermediate covariate (such as a biomarker of interest) for the second-line treatment. Such trials are often called sequential multiple assignment randomized trials (SMART). A dynamic treatment rule 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. 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. The optimal dynamic treatment rule is defined as the dynamic treatment rule that maximizes the mean outcome.

Previous approaches, described at the end of this chapter, rely on semiparametric models that make strong assumptions on the data generating process. We instead define the statistical model for the data distribution as nonparametric, beyond possible knowledge about the treatment mechanism (e.g., known in a RCT) and censoring mechanism. In order to not only consider the most ambitious fully optimal rule, we

A. R. Luedtke (✉)
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,
1100 Fairview Ave, Seattle, WA 98109, USA
e-mail: aluedtke@fredhutch.org

M. J. van der Laan
Division of Biostatistics and Department of Statistics, University of California, Berkeley,
101 Haviland Hall, #7358, Berkeley, CA 94720, USA
e-mail: laan@berkeley.edu

M.J. van der Laan, S. Rose, *Targeted Learning in Data Science*,
Springer Series in Statistics, https://doi.org/10.1007/978-3-319-65304-4_22

define the $V$-optimal rules as the optimal rule that only uses a user-supplied subset $V$ of the available covariates. This allows us to consider suboptimal rules that are easier to estimate and thereby allow for statistical inference for the counterfactual mean outcome under the suboptimal rule.

In this chapter, we describe how to obtain semiparametric inference about the mean outcome under the two time point $V$-optimal rule. We will show that the mean outcome under the optimal rule is a pathwise differentiable parameter of the data distribution, indicating that it is possible to develop asymptotically linear estimators of this target parameter under conditions. In fact, we obtain the surprising result that the pathwise derivative of this target parameter equals the pathwise derivative of the mean counterfactual outcome under a given dynamic treatment rule set at the optimal rule, treating the latter as known. By a reference to the earlier for double robust and efficient estimation of the mean outcome under a given rule (see Chap. 4), we then obtain a CV-TMLE for the mean outcome under the optimal rule. Subsequently, we prove asymptotic linearity and efficiency of this CV-TMLE, allowing us to construct confidence intervals for the mean outcome under the optimal dynamic treatment or its contrast with respect to a standard treatment.

In a SMART the statistical inference would only rely upon a second-order difference between the estimator of the optimal dynamic treatment and the optimal dynamic treatment itself to be asymptotically negligible. This is a reasonable condition if we restrict ourselves to rules only responding to a one-dimensional time-dependent covariate, or if we are willing to make smoothness assumptions. While this condition appears to be necessary when estimating the optimal mean outcome, it is not necessary if the parameter of interest is redefined as the average mean outcome under our cross-validated estimates of the optimal dynamic treatment. This parameter relies on the data through our estimates of the optimal dynamic treatment, and we thus refer to it as a data-adaptive parameter.

## 22.1 Optimal Dynamic Treatment Estimation Problem

For the sake of presentation, we focus on two time point treatments in this chapter. Suppose we observe $n$ i.i.d. copies $O_1, \ldots, O_n \in \mathcal{O}$ of

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

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” $j$, $j = 0, 1$. That is, $A_2(0) = 0$ implies that $(L(1), A_1(1), Y)$ is not observed, and $A_2(1) = 0$ implies that $Y$ is not observed. 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. For a time-dependent process

$X(\cdot)$, we will use the notation $\bar{X}(t) = (X(s) : s \leq t)$, where $\bar{X}(-1) = \emptyset$. 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 $g_{0A(j)}$ 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:

$$g_0(O) = \prod_{j=1}^{2} g_{01}(A_1(j) \mid \bar{A}(j-1), \bar{L}(j)) g_{02}(A_2(j) \mid A_1(j), \bar{A}(j-1), \bar{L}(j)).$$

In particular, the data might have been generated by a SMART, in which case $g_{01}$ is known.

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

$$d(a(0), v) \equiv (d_{A(0)}(v(0)), d_{A(1)}(a(0), v(1))).$$

If we assume a structural equation model (Pearl 2009a) for variables stating that

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

where the collection of functions $f = (f_{L(0)}, f_{A(0)}, f_{L(1)}, f_{A(1)})$ are unspecified or partially specified, we can define counterfactuals $Y_d$ defined by the modified system in which the equations for $A(0), A(1)$ are replaced by $A(0) = d_{A(0)}(V(0))$ and $A(1) = d_{A(1)}(A(0), V(1))$. Denote the distribution of these counterfactual quantities as $P_{0,d}$, where we note that $P_{0,d}$ is implied by the collection of functions $f$ and the joint distribution of exogenous variables $(U_{L(0)}, U_{A(0)}, U_{L(1)}, U_{A(1)}, U_Y)$. We can now define the causally optimal rule under $P_{0,d}$ as $d_0^* = \arg\max_{d \in \mathcal{D}} E_{P_{0,d}} Y_d$. If we assume a sequential randomization assumption stating that $A(0)$ is independent of $U_{L(1)}, U_Y$, given $L(0)$, and $A(1)$ is independent of $U_Y$, given $\bar{L}(1), A(0)$, then we can identify $P_{0,d}$ with observed data under the distribution $P_0$ using the $g$-computation formula:

$$\begin{aligned} &p_{0,d}(L(0), A(0), L(1), A(1), Y) \\ &\equiv I(A = d(A(0), V)) q_{0,L(0)}(L(0)) q_{0,L(1)}(L(1) \mid L(0), A(0)) q_{0,Y}(Y \mid \bar{L}(1), \bar{A}(1)), \end{aligned} \tag{22.1}$$

where $p_{0,d}$ is the density of $P_{0,d}$ and $q_{0,L(0)}$, $q_{0,L(1)}$, and $q_{0,Y}$ are the densities for $Q_{0,L(0)}$, $Q_{0,L(1)}$, and $Q_{0,Y}$, where $Q_{0,Y}$ represents the distribution of $Y$ given $\bar{L}(1), \bar{A}(1)$. We assume that all densities above are absolutely continuous with respect to some dominating measure $\mu$. We have a similar identifiability result/$g$-computation formula under the Neyman-Rubin causal model (Robins 1987). More generally, for a distribution $P \in \mathcal{M}$ we can define the $g$-computation distribution $P_d$ as the distribution with density

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

where $q_{L(0)}$, $q_{L(1)}$, and $q_Y$ are the counterparts to $q_{0,L(0)}$, $q_{0,L(1)}$, and $q_{0,Y}$ under $P$.

For the remainder of this chapter, if for a static or dynamic intervention $d$, we use notation $L_d$ (or $Y_d$, $O_d$) we mean the random variable with the probability distribution $P_d$ in (22.1) so that of all our quantities are statistical parameters. For example, the quantity $E_0(Y_{a(0)a(1)} \mid V_{a(0)}(1))$ defined in the next theorem denotes the conditional expectation of $Y_{a(0)a(1)}$, given $V_{a(0)}(1)$, under the probability distribution $P_{0,a(0)a(1)}$ (i.e., $g$-computation formula presented above for the static intervention $(a(0), a(1))$). In addition, if we write down these parameters for some $P_d$, we will automatically assume the positivity assumption at $P$ required for the $g$-computation formula to be well defined. For that it will suffice to assume the following positivity assumption at $P$:

$$\begin{aligned} Pr_P\left(0 < \min_{a_1 \in \{0,1\}} g_{0A(0)}(a_1, 1|L(0))\right) &= 1 \\ Pr_P\left(0 < \min_{a_1 \in \{0,1\}} g_{0A(1)}(a_1, 1 \mid \bar{L}(1), A(0))\right) &= 1. \end{aligned} \tag{22.2}$$

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

We now define a statistical parameter representing the mean outcome $Y_d$ under $P_d$. For any rule $d \in \mathcal{D}$, let

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

For a distribution $P$, define the $V$-optimal rule as

$$d_P = \arg\max_{d \in \mathcal{D}} E_{P_d} Y_d.$$

For simplicity, we will write $d_0$ instead of $d_{P_0}$ for the $V$-optimal rule under $P_0$. Define the parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ as $\Psi(P) = E_{P_{d_P}} Y_{d_P}$. The first part of this chapter is concerned with inference for the parameter

$$\psi_0 \equiv \Psi(P_0) = E_{P_{0,d_0}} Y_{d_0}.$$

Under our identifiability assumptions, $d_0$ is equal to the causally optimal rule $d_0^*$. Even if the sequential randomization assumption does not hold, the statistical parameter $\psi_0$ represents a statistical parameter of interest in its own right. We will not concern ourselves with the sequential randomization assumption for the remainder of this paper.

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

**Theorem 22.1.** *Suppose $V(0)$ is a function of $V(1)$. The $V$-optimal rule $d_0$ can be represented as the following explicit parameter of $P_0$:*

$$\begin{aligned}\bar{Q}_{b,20}(a(0), v(1)) &= E_0(Y_{a(0),A(1)=(1,1)} \mid V_{a(0)}(1) = v(1)) \\ &\quad -E_0(Y_{a(0),A(1)=(0,1)} \mid V_{a(0)}(1) = v(1)), \\ d_{0,A(1)}(A(0), V(1)) &= (I(\bar{Q}_{b,20}(A(0), V(1)) > 0), 1), \\ \bar{Q}_{b,10}(v(0)) &= E_0(Y_{(1,1),d_{0,A(1)}} \mid V(0)) - E_0(Y_{(0,1),d_{0,A(1)}} \mid V(0)), \\ d_{0,A(0)}(V(0)) &= (I(\bar{Q}_{b,10}(V(0)) > 0), 1),\end{aligned}$$

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

$$E_0(Y_{a(0),a(1)} \mid V(0), V_{a(0)}(1)) = E_0(Y_{a(0),a(1)} \mid V_{a(0)}(1)), \tag{22.3}$$

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

Following Robins (2004), we refer to $\bar{Q}_{b,10}$ and $\bar{Q}_{b,20}$ as the (first and second time point) blip functions.

## 22.2 Efficient Influence Curve of the Mean Outcome Under $V$-Optimal Rule

In this section, we establish the pathwise differentiability of $\Psi$ and give an explicit expression for the efficient influence curve. Before presenting this result, we give the efficient influence curve for the parameter $\Psi : \mathcal{M} \to \mathbb{R}$ where $\Psi_d(P) \equiv E_P Y_d$ and the rule $d = (d_{A(0)}, d_{A(1)}) \in \mathcal{D}$ is treated as known. This influence curve was presented in Chap. 4. The parameter mapping $\Psi_d$ has efficient influence curve

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

where

$$D_0^*(d, P) = E_P\left[Y_d \mid L(0), A(0) = d_{A(0)}(V(0))\right] - E_P Y_d,$$

$$\begin{aligned}
D_1^*(d,P) =& \frac{I(A(0)=d_{A(0)}(V(0)))}{g_{A(0)}(O)} (E_P[Y \mid \bar{A}(1)=d(A(0),V),\bar{L}(1)] \\
& - E_P[Y_d \mid L(0), A(0)=d_{A(0)}(V(0))]), \\
D_2^*(d,P) =& \frac{I(\bar{A}(1)=d(A(0),V))}{\prod_{j=0}^1 g_{A(j)}(O)} \left(Y - E_P\left[Y \mid \bar{A}(1)=d(A(0),V),\bar{L}(1)\right]\right). \qquad (22.4)
\end{aligned}$$

Above $(g_{A(0)}, g_{A(1)})$ is the intervention mechanism under the distribution $P$. We remind the reader that $Y_d$ has the $g$-computation distribution from (22.1) so that:

$$\begin{aligned}
& E_P\left[Y_d \mid L(0), A(0)=d_{A(0)}(V(0))\right] \\
& \quad = E_P\left[E_P\left[Y \mid \bar{A}(1)=d(A(0),V),\bar{L}(1)\right)\right] \mid L(0), A(0)=d_{A(0)}(V(0))\right]
\end{aligned}$$

At times it will be convenient to write $D_k^*(d,Q^d,g)$ instead of $D_k^*(d,P)$, where $Q^d$ represents both of the conditional expectations in the definitions of $D_1^*$ and the marginal distribution of $L(0)$ under $P$ and $g$ represents the intervention mechanism under $P$. We will denote these conditional expectations under $P_0$ for a given rule $d$ by $Q_0^d$. We will similarly at times denote $D^*(d,P)$ by $D^*(d,Q^d,g)$.

Whenever $D^*(P)$ does not contain an argument for a rule $d$, this $D^*(P)$ refers to the efficient influence curve of the parameter mapping $\Psi$ for which $\Psi(P)=E_P Y_{d_P}$, where the optimal rule $d_P$ under $P$ is not treated as known. Not treating $d_P$ as known means that $d_P$ depends on the input distribution $P$ in the mapping $\Psi(P)$. The following theorem presents the efficient influence curve of $\Psi$ at a distribution $P$. The main condition on this distribution $P$ is that it satisfies the nonexceptional law condition that

$$\begin{aligned}
& \max_{a_0(0)\in\{0,1\}} Pr_P\left(\bar{Q}_{b,2}((a_0(0),1), V_{a(0)=(a_0(0),1)})=0\right)=0, \\
& Pr_P\left(\bar{Q}_{b,1}(V(0))=0\right)=0, \qquad (22.5)
\end{aligned}$$

where $\bar{Q}_{b,2}$ and $\bar{Q}_{b,1}$ are defined analogously to $\bar{Q}_{b,20}$ and $\bar{Q}_{b,10}$ in Theorem 22.1 with the expectations under $P_0$ replaced by expectations under $P$. That is, we assume that each of the blip functions under $P$ is nowhere zero with probability 1. Distributions that do not satisfy this assumption have been referred to as "exceptional laws" (Robins 2004). These laws are indeed exceptional when one expects that treatment will have a beneficial or harmful effect in all $V$-strata of individuals. When one only expects that treatment will have an effect on outcome in some but not all strata of individuals then this assumption may be violated. We will make this assumption about $P_0$ for all subsequent asymptotic linearity results about $E_0Y_{d_0}$, and we will assume a weaker but still not completely trivial assumption about the consistency of the optimal rule estimate to some fixed limit for the data-adaptive target parameters in Sect. 22.3.

**Theorem 22.2.** *Suppose $P \in \mathcal{M}$ is such that $Pr_P(\mid Y \mid< M)=1$ for some $M<\infty$, $P$ satisfies the positivity assumption (22.2), and $P$ satisfies the nonexceptional law condition (22.5). Then the parameter $\Psi:\mathcal{M}\to\mathbb{R}$ is pathwise differentiable at $P$ with canonical gradient given by*

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

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

The above theorem is proved as Theorem 8 in van der Laan and Luedtke (2014) so the proof is omitted here.

We will at times denote $D^*(P)$ by $D^*(Q, g)$, where $Q$ represents $Q^{d_P}$, along with portions of the likelihood that suffice to compute the $V$-optimal rule $d_P$. We denote $d_P$ by $d_Q$ when convenient. We explore which parts of the likelihood suffice to compute the $V$-optimal rule in our companion paper, though Theorem 22.1 shows that $\bar{Q}_{b,20}$ and $\bar{Q}_{b,10}$ suffice for $d_0$ (and analogous functions suffice for a more general $d_P$). We have the following property of the efficient influence curve, which will provide a fundamental ingredient in the analysis of the CV-TMLE presented in the next section.

**Theorem 22.3.** *Let $d_Q$ be the $V$-optimal rule corresponding with $Q$. For any $Q, g$, we have*

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

*where, for all $d \in \mathcal{D}$,*

$$R_{1d}(Q^d, Q_0^d, g, g_0) \equiv P_0 D^*(d, Q^d, g) - (\Psi_d(Q_0^d) - \Psi_d(Q^d)),$$
$$R_2(Q, Q_0) \equiv \Psi_{d_Q}(Q_0^{d_Q}) - \Psi_{d_0}(Q_0^{d_0}),$$

*$\Psi_d(P) = E_P Y_d$ is the statistical target parameter that treats $d$ as known, and $D^*(d, Q_0^d, g_0)$ is the efficient influence curve of $\Psi_d$ at $P_0$ as given in Theorem 22.2.*

From the study of the statistical target parameter $\Psi_d$ in Chap. 4, we know that $P_0 D^*(d, Q^d, g) = \Psi_d(Q_0^d) - \Psi_d(Q^d) + R_{1d}(Q^d, Q_0^d, g, g_0)$, where $R_{1d}$ is a closed form second-order term involving integrals of differences $Q^d - Q_0^d$ times differences $g - g_0$.

## 22.3 Statistical Inference for the Average of Sample-Split Specific Mean Counterfactual Outcomes Under Data Adaptively Determined Dynamic Treatments

Let $\hat{d} : \mathcal{M} \to \mathcal{D}$ be an estimator that maps an empirical distribution into an individualized treatment rule. Let $B_n \in \{0, 1\}^n$ denote a random vector for a cross-validation split, and for a split $B_n$, let $P_{n,B_n}^0$ be the empirical distribution of the training sample $\{i : B_n(i) = 0\}$ and $P_{n,B_n}^1$ is the empirical distribution of the validation sample $\{i : B_n(i) = 1\}$. Consider a $J$-fold cross-validation scheme. In $J$-fold cross-validation, the data is split into $J$ mutually exclusive and exhaustive sets of size approximately $n/J$ 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. With probability $1/J$, $B_n$ has value 1 in all indices in validation set $j \in \{1, \ldots, J\}$ and 0 for all indices not corresponding to training set $j$.

In this section, we first present a method that provides an estimator and statistical inference for the data-adaptive target parameter

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

Note that this target parameter is defined as the average of data-adaptive parameters, where the data-adaptive parameters are learned from the training samples of size approximately $n(J-1)/J$. One applies the estimator $\hat{d}$ to each of the $J$ training samples, giving a target parameter value $\Psi_{\hat{d}(P^0_{n,B_n})}(P_0)$, and our target parameter $\tilde{\psi}_{0n}$ is defined as the average across these $J$ target parameters.

### 22.3.1 General Description of CV-TMLE

Here we give a general overview of the CV-TMLE procedure. In Sect. 22.6 we present a particular CV-TMLE that satisfies all of the properties described in this section. Denote the realizations of $B_n$ with $j = 1, .., J$, and let $d_{nj} = \hat{d}(P^0_{n,j})$ for some estimator of the optimal rule $\hat{d}$. Let

$$(a(0), \bar{l}(1)) \mapsto E_{nj}[Y|\bar{A}(1) = d_{nj}(a(0), v), \bar{L}(1) = \bar{l}(1)]$$

represent an initial estimate of $E_0[Y \mid \bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)]$ based on the training sample $j$. Similarly, let $l(0) \mapsto E_{nj}[Y_{d_{nj}}|L(0) = l(0)]$ represent an initial estimate of $E_0[Y_{d_{nj}}|L(0)]$ based on the training sample $j$. Finally, let $Q_{L(0),nj}$ represent the empirical distribution of $L(0)$ in validation smaple $j$. We then fluctuate these three regression functions using the following submodels:

$$\left\{E^{(\epsilon_2)}_{nj}[Y|\bar{A}(1) = d_{nj}(a(0), v), \bar{L}(1) = \bar{l}(1)] : \epsilon_2 \in \mathbb{R}\right\}$$
$$\left\{E^{(\epsilon_1)}_{nj}[Y_{d_{nj}}|L(0) = l(0)] : \epsilon_1 \in \mathbb{R}\right\}$$
$$\left\{Q^{(\epsilon_0)}_{L(0),nj} : \epsilon_0 \in \mathbb{R}\}\right\},$$

where these submodels rely on an estimate $g_{nj}$ of $g_0$ based on training sample $j$ and are such that:

$$E^{(0)}_{nj}[Y|\bar{A}(1) = d_{nj}(a(0), v), \bar{L}(1)] = E_{nj}[Y|\bar{A}(1) = d_{nj}(a(0), v), \bar{L}(1)]$$
$$E^{(0)}_{nj}[Y_{d_{nj}}|L(0)] = E_{nj}[Y_{d_{nj}}|L(0)]$$
$$Q^{(0)}_{L(0),nj} = Q_{L(0),nj}.$$

Let $Q_{nj}^{d_{nj}}(\epsilon)$ represent the parameter mapping that gives the three regression functions above fluctuated by $\epsilon \equiv (\epsilon_0, \epsilon_1, \epsilon_2)$. For a fixed $\epsilon$, $Q_{nj}^{d_{nj}}(\epsilon)$ only relies on $P_{nj}^1$ through the empirical distribution of $L(0)$ in validation sample $j$. Let $\phi$ be a valid loss function for $Q_0^d$ so that $Q_0^d = \arg\min_{Q^d} P_0\phi(Q^d)$, and let $\phi$ and the submodels above satisfy

$$D^*(d, Q^d, g) \in \left\langle \frac{d}{d\epsilon}\phi(Q^d(\epsilon))\Big|_{\epsilon=0} \right\rangle,$$

where $\langle f \rangle = \{\sum_j \beta_j f_j : \beta\}$ denotes the linear space spanned by the components of $f$. We choose $\epsilon_n$ to minimize $P_n^1\phi(Q_{nj}^{d_{nj}}(\epsilon))$ over $\epsilon \in \mathbb{R}^3$. We then define the targeted estimate $Q_{nj}^{d_{nj}*} \equiv Q_{nj}^{d_{nj}}(\epsilon_n)$ of $Q_0^{d_{nj}}$. We note that $Q_{nj}^{d_{nj}*}$ maintains the rate of convergence of $Q_{nj}$ under mild conditions that are standard to M-estimator analysis. The key property that we need from the $\epsilon_n$ and the corresponding update $Q_{nj}^{d_{nj}*}$ is that it (approximately) solves the cross-validated empirical mean of the efficient influence curve:

$$E_{B_n} P_{n,B_n}^1 D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj}) = o_{P_0}(1/\sqrt{n}). \tag{22.6}$$

The CV-TMLE implementation presented in the appendix satisfies this equation with $o_{P_0}(1/\sqrt{n})$ replaced by 0. The proposed estimator of $\tilde{\psi}_{0n}$ is given by

$$\tilde{\psi}_n^* \equiv E_{B_n}\Psi_{d_{nj}}(Q_{nj}^{d_{nj}*}).$$

We give a concrete CV-TMLE algorithm for $\tilde{\psi}_n^*$ in Sect. 22.6, but note that other CV-TMLE algorithms can be derived using the approach in this section for different choices of loss function $\phi$ and submodels.

### 22.3.2 *Statistical Inference for the Data-Adaptive Parameter $\tilde{\psi}_{0n}$*

We now proceed with the analysis of this CV-TMLE $\tilde{\psi}_n^*$ of $\tilde{\psi}_{0n}$. We first give a representation theorem for the CV-TMLE.

**Theorem 22.4.** *Let $g_{nj}$ and $d_{nj}$ represent estimates of $g_0$ and $d_0$ based on training sample $j$. Let $Q_{nj}^{d_{nj}*}$ represent a targeted estimate of $Q_0^{d_{nj}}$ as presented in Sect. 22.3.1 so that $Q_{nj}^{d_{nj}*}$ satisfies (22.6). Let $R_{1d}$ be as in Theorem 22.3. Further suppose that the supremum norm of* $\max_j D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj})$ *is bounded by some $M < \infty$ with probability tending to 1, and that*

$$\max_{j\in\{1,\ldots,J\}} P_0\{D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj}) - D^*(d_1, Q^{d_1}, g)\}^2 \to 0 \textit{ in probability}$$

*for some $d_1 \in \mathcal{D}$ and possibly misspecified $Q^{d_1}$ and $g$. Finally, suppose that*

$$\max_{j\in\{1,\dots,J\}} \left| R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}}, g_{nj}, g_0) \right| = o_{P_0}(n^{-1/2}).$$

*Then,*

$$\tilde{\psi}_n^* - \tilde{\psi}_{0n} = (P_n - P_0)D^*(d_1, Q^{d_1}, g^{d_1}) + o_{P_0}(n^{-1/2}).$$

Note that $d_1$ in the above theorem need not be the same as the optimal rule $d_0$, though later we will discuss the desirable special case where $d_1 = d_0$. The above theorem also does not require that $g_0$ is known, or even that the limit of our intervention mechanisms $g$ is equal to $g_0$.

Note in the above theorem that the condition that, if $g_0$ is known so that all $g_{nj}$ can be correctly specified, it immediately follows that

$$\max_{j\in\{1,\dots,J\}} \left| R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}}, g_{nj}, g_0) \right| = 0.$$

In practice we would recommend estimating $g_0$ according to a correctly specified model even when $g_0$ is known, because this can improve efficiency (see Section 2.3.7 of van der Laan and Robins 2003).

If the conditions of the above theorem hold, the asymptotic linearity result implies that

$$\sqrt{n}\left[\tilde{\psi}_n^* - \tilde{\psi}_{0n}\right] \to \text{Normal}(0, \sigma_0^2),$$

where $\sigma_0^2 = P_0 D^*(d_1, Q^{d_1}, g^{d_1})^2$. Under mild conditions,

$$\sigma_n^2 = \frac{1}{J}\sum_{j=1}^{J} P_{n,j}^1 \left\{D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj})\right\}^2$$

consistently estimates $\sigma_0^2$. Under the consistency of $\sigma_n^2$ and the conditions of Theorem 22.4, an asymptotically valid 95% confidence interval for $\tilde{\psi}_{0n}$ is given by

$$\left[\tilde{\psi}_n^* \pm \frac{\sigma_n}{\sqrt{n}}\right]. \tag{22.7}$$

### 22.3.3 Statistical Inference for the True Optimal Rule $\psi_0$

Suppose now that we are interested in estimating the mean outcome under the optimal rule $d_0$ rather than the data-adaptive parameter $\tilde{\psi}_{0n}$. Note that

$$\begin{aligned}\sqrt{n}\left(\tilde{\psi}_n^* - \psi_0\right) &= \sqrt{n}\left(\tilde{\psi}_n^* - \tilde{\psi}_{0n}\right) + \sqrt{n}\left(\tilde{\psi}_{0n} - \psi_0\right) \\ &= \sqrt{n}\left(\tilde{\psi}_n^* - \tilde{\psi}_{0n}\right) + \frac{\sqrt{n}}{J}\sum_{j=1}^{J}\left[\Psi_{d_{nj}}(P_0) - \psi_0\right].\end{aligned}$$

If $J^{-1}\sum_{j=1}^{J}\left[\Psi_{d_{nj}}(P_0)-\psi_0\right]$, then by Slutsky's theorem the left-hand side has the same normal limit as $\sqrt{n}\left(\tilde{\psi}_n^{*}-\tilde{\psi}_{0n}\right)$ provided the conditions of Theorem 22.4 hold. Furthermore, as $J$ is fixed as $n\to\infty$, $J^{-1}\sum_{j=1}^{J}\left[\Psi_{d_{nj}}(P_0)-\psi_0\right]=o_P(n^{-1/2})$ if

$$\Psi_{d_{nj}}(P_0)-\psi_0=o_P(n^{-1/2})\text{ for each } j. \tag{22.8}$$

To analyze $\Psi_{d_{nj}}(P_0)-\psi_0$, we will assume that the user estimates $\bar{Q}_{b,10}$ and $\bar{Q}_{b,20}$ using $\bar{Q}_{b,1nj}$ and $\bar{Q}_{b,2nj}$, and then subsequently uses the plug-in estimators of the format described in Theorem 22.1. Data-adaptive estimators of $\bar{Q}_{b,10}$ and $\bar{Q}_{b,20}$ were previously described in Luedtke and van der Laan (2016b). While we do not require that $d_{nj}$ result from a plug-in estimator, this is the estimation scheme we will focus on analyzing here. Given that the main result needed to show (22.8) for the plug-in estimator is analytic in nature, we focus on a general $Q$ with corresponding blip functions $\bar{Q}_{b,1}$, $\bar{Q}_{b,2}$ and optimal rule plug-in estimates $d_{Q,A(0)}$, $d_{Q,A(1)}$. One can then apply this result directly to our fold-specific estimator.

The following result is proved in Sect. 22.5.

**Lemma 22.1.** *Recall the definitions of $\bar{Q}_{b,20}$ and $\bar{Q}_{b,10}$ in Theorem 22.1. We can represent $\Psi(P_0)=E_0Y_{d_0}$ as follows:*

$$\begin{aligned}\Psi(P_0)=E_0Y_{(0,1),(0,1)}&+E_0\left[d_{0,A(1)}((0,1),V_{(0,1)}(1))\bar{Q}_{b,20}((0,1),V_{(0,1)}(1))\right]\\&+E_0d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0)).\end{aligned}$$

*where $V_{(0,1)}(1)$ is drawn under the* g*-computation distribution for which treatment* $(0,1)$ *is given at the first time point.*

It follows that

$$\begin{aligned}R_2(Q,Q_0)=&E_0(d_{Q,A(0)}-d_{0,A(0)})(V(0))\bar{Q}_{b,10}(V(0))\\&+E_0(d_{Q,A(1)}-d_{0,A(1)})((0,1),V_{(0,1)}(1))\bar{Q}_{b,20}((0,1),V_{(0,1)}(1))\\\equiv&R_{2,A(0)}(Q,Q_0)+R_{2,A(1)}(Q,Q_0).\end{aligned}$$

We will be able to attain a fast rate on $R_2(Q,Q_0)$ under margin assumptions. We start with the assumption that we use to bound $R_{2,A(0)}(Q,Q_0)$. Suppose there exist positive constants $C_1$,$\beta_1$ such that, for all $t>0$,

$$P_0\left\{0<\left|\bar{Q}_{b,10}(V(0))\right|\le t\right\}\le C_1t^{\beta_1}. \tag{MA1}$$

The above assumption requires that the blip function at the first time point does not concentrate too much mass near (but not at) the decision boundary (zero). The assumption is different from the exceptional law condition, since that condition requires that this blip function places no mass exactly at the decision boundary. For $\beta_1$ and $\beta_2$ small, this is a weak assumption, though it may not attain the rates of convergence needed to satisfy (22.8). These assumptions hold for $\beta_1=1$ if that the blip functions applied to the data have bounded Lebesgue density in a neighborhood of zero.

We make a similar assumption on $\bar{Q}_{b,20}$. In particular, we assume there exists some $C_2, \beta_2$ such that, for all $t > 0$,

$$P_0 \left\{ 0 < \left| \bar{Q}_{b,20}((0,1), V_{(0,1)}(1)) \right| \le t \right\} \le C_2 t^{\beta_2}. \tag{MA2}$$

We now show that (MA1) and (MA2) give a $\beta_1,\beta_2$-specific upper bound of $R_2(Q, Q_0)$ by the distance of $\bar{Q}_{b,1}$ and $\bar{Q}_{b,2}$ from $\bar{Q}_{b,10}$ and $\bar{Q}_{b,20}$.

**Theorem 22.5.** *If (MA1) holds for some $C_1, \beta_1 > 0$, then, for some constant $C > 0$,*

$$|R_{2,A(0)}(Q, Q_0)| \le C \min \left\{ \left\| \bar{Q}_{b,1} - \bar{Q}_{b,10} \right\|_{2,P_0}^{2(1+\beta_1)/(2+\beta_1)}, \left\| \bar{Q}_{b,1} - \bar{Q}_{b,10} \right\|_{\infty,P_0}^{1+\beta_1} \right\}. \tag{22.9}$$

*If (MA2) holds for some $C_2, \beta_2 > 0$, then, for some constant $C > 0$,*

$$|R_{2,A(1)}(Q, Q_0)| \le C \min \left\{ \left\| \bar{Q}_{b,2} - \bar{Q}_{b,20} \right\|_{2,P_{0,(0,1)}}^{2(1+\beta_1)/(2+\beta_1)}, \left\| \bar{Q}_{b,1} - \bar{Q}_{b,10} \right\|_{P_{0,(0,1)}}^{1+\beta_1} \right\}, \tag{22.10}$$

*where $P_{0,(0,1)}$ represents the static intervention specific g-computation distribution where treatment $(0, 1)$ is given at the first time point.*

The above analytic result is useful for evaluating the plausibility of (22.8) when $d_{nj}$ is estimated using a plug-in estimator. In a parametric model, one could typically estimate $\bar{Q}_{b,10}$ and $\bar{Q}_{b,20}$ at $n^{-1/2}$ rates for both the $L_2(P_0)$ and the supremum norms presented above. Hence, if the margin conditions hold with $\beta_1 = \beta_2 = 1$, the supremum norm result yields $n^{-1}$ rates on each $\Psi_{d_{nj}}(P_0) - \psi_0$. In practice we of course do not expect to be able to correctly specify a parametric model. Rather, we would use data-adaptive estimators for the blip functions, such as super learning, to make correct specification of the estimators more likely (Luedtke and van der Laan 2016b). Under smoothness assumptions on the blip functions, one can ensure nearly parametric rates on the $L_2(P_0)$ norm using smoothing. These rates can, under enough smoothness, achieve the $o_P(n^{-3/8}$ rate required by the above theorem at $\beta_1 = \beta_2 = 1$ to show that $\Psi_{d_{nj}}(P_0) - \psi_0 = o_P(n^{-1/2})$. Nonetheless, in general we may not expect such a fast rate to hold. If this fast rate does not hold, then one can still achieve inference for the data-adaptive parameter $\tilde{\psi}_{0n}$. If the fast rate does hold, as may be possible if $V$ is low-dimensional, then the implication is that, under the conditions of Theorem 22.4 and the consistency of $\sigma_n^2$, the confidence interval presented in (22.7) is asymptotically valid for both $\tilde{\psi}_{0n}$ and $\psi_0$.

## 22.4 Discussion

This chapter investigated semiparametric statistical inference for the mean outcome under the $V$-optimal rule and statistical inference for the data-adaptive target parameter defined as the mean outcome under a data adaptively determined $V$-optimal rule (treating the latter as given). We proved a surprising and useful result stating that the mean outcome under the $V$-optimal rule is represented by a statistical parameter whose pathwise derivative is identical to what it would have been if the unknown rule had been treated as known, under the condition that the data is generated by a nonexceptional law. As a consequence, the efficient influence curve is immediately known, and any of the efficient estimators for the mean outcome under a given rule can be applied at the estimated rule. In particular, we demonstrate a CV-TMLE, and present asymptotic linearity results. However, the dependence of the statistical target parameter on the unknown rule affects the second-order terms of the CV-TMLE, and, as a consequence, the asymptotic linearity of the CV-TMLE requires that a second-order difference between the estimated rule and the $V$-optimal rule converges to zero at a rate faster than $1/\sqrt{n}$. While this can be expected to hold for rules that are only a function of one continuous score (such as a biomarker), only strong smoothness assumptions will guarantee this when $V$ is moderate-to-high dimensional, so that, even in an RCT, we cannot expect valid statistical inference for such $V$-optimal rules.

To account for this challenge, we also described estimation of the average of sample split specific data-adaptive target parameters, as in general proposed in Hubbard et al. (2016). Specifically, our data-adaptive target parameter is defined as an average across $J$ sample splits in training and validation sample of the mean outcome under the dynamic treatment fitted on the training sample. We presented a CV-TMLE of this data-adaptive target parameter, and we established an asymptotic linearity theorem that does not require that the estimated rule be consistent for the optimal rule, let alone at a particular rate. We showed that statistical inference for this data-adaptive target parameter does not rely on the convergence rate of our estimated rule to the optimal rule, and in fact only requires that the data adaptively fitted rule converges to some (possibly suboptimal) fixed rule. As a consequence, in a sequential RCT, this method provides valid asymptotic statistical inference under very mild conditions, the primary of which is that the estimated rule converges to some (possibly suboptimal) fixed rule.

Drawing inferences concerning optimal treatment strategies is an important topic that will hopefully help guide future health policy decisions. We believe that working with a large semiparametric model is desirable because it helps to ensure that the projected health benefits from implementing an estimated treatment strategy are not due to bias from a misspecified model. The CV-TMLEs presented in this chapter have many desirable statistical properties and allow one to get estimates and make inference in this large model.

## 22.5 Proofs

*Proof (Theorem 22.1).* Let $V_d = (V(0), V_d(1))$. For a rule in $\mathcal{D}$, we have

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

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

$$\begin{aligned}&\bar{Q}_{b,20}(a(0), v(0), v(1)) =\\ &E_0(Y_{a(0),A(1)=(1,1)} - Y_{a(0),A(1)=(0,1)} \mid V_{a(0)}(1) = v(1), V(0) = v(0)).\end{aligned}$$

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

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

$$\begin{aligned}&E_{V_d}\left(E_P(Y_{a(0),d_{0,A(1)}} \mid V_{a(0)})I(a(0) = d_{A(0)}(V(0))\right)\\ &= E_0E(Y_{a(0),d_{0,A(1)}} \mid V(0))I(a(0) = d_{A(0)}(V(0)),\end{aligned}$$

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

*Proof (Theorem 22.3).* By the definition of $R_{1d}$ we have

$$\begin{aligned}&P_0D^*(Q, g) = P_0D^*(d_Q, Q, g) = \Psi_{d_Q}(Q_0^{d_Q}) - \Psi_{d_Q}(Q^{d_Q}) + R_{1d_Q}(Q^{d_Q}, Q_0^{d_Q}, g, g_0)\\ &= \Psi_{d_0}(Q_0^{d_0}) - \Psi_{d_Q}(Q^{d_Q}) + \{\Psi_{d_Q}(Q_0^{d_Q}) - \Psi_{d_0}(Q_0^{d_0})\} + R_{1d_Q}(Q^{d_Q}, Q_0^{d_Q}, g, g_0)\\ &= \Psi(Q_0) - \Psi(Q) + R_2(Q, Q_0) + R_{1d_Q}(Q^{d_Q}, Q_0^{d_Q}, g, g_0).\end{aligned}$$

*Proof (Theorem 22.4).* For all $j = 1, \ldots, J$, we have that:

$$\begin{aligned}\Psi_{d_{nj}}(Q_{nj}^{d_{nj}*}) - \Psi_{d_{nj}}(Q_0^{d_{nj}*}) = &- P_0D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj})\\ &+ R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}*}, g_{nj}, g_0)\end{aligned}$$

Summing over $j$ and using (22.6) gives:

$$\tilde{\psi}_n^* - \tilde{\psi}_{0n} = \frac{1}{J}\sum_{j=1}^{J}\left((P_{n,j}^1 - P_0)D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj}) + R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}*}, g_{nj}, g_0)\right).$$

We also have that:

$$\frac{1}{J}\sum_{j=1}^{J}(P_{n,j}^1 - P_0)\left(D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj}) - D^*(d_1, Q^{d_1}, g)\right) = o_{P_0}(n^{-1/2}).$$

The above follows from the first by applying the law of total expectation conditional on the training sample, and then noting that each $\hat{Q}^*(P_{n,B_n}^0, \epsilon_n)$ only relies on $P_{n,B_n}^0$

through the finite dimensional parameter $\epsilon_n$. Because GLM-based parametric classes easily satisfy an entropy integral condition (van der Vaart and Wellner 1996), the consistency assumption on $D^*(d_{nj}, Q_{nj}^{d_{nj}*}, g_{nj})$ shows that the above is second order. We refer the reader to Zheng and van der Laan (2010) for a detailed proof of the above result for general cross-validation schemes, including $J$-fold cross-validation.

It follows that:

$$\begin{aligned}\tilde{\psi}_n^* - \tilde{\psi}_{0n} =&(P_n - P_0)D^*(d_1, Q^{d_1}, g)\\ &+ \frac{1}{J}\sum_{j=1}^{J} R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}*}, g_{nj}, g_0) + o_{P_0}(n^{-1/2}).\end{aligned}$$

Finally, note that $\frac{1}{J}\sum_{j=1}^{J} R_{1d_{nj}}(Q_{nj}^{d_{nj}*}, Q_0^{d_{nj}*}, g_{nj}, g_0)$ is $o_P(n^{-1/2})$ by the last assumption of the theorem.

*Proof (Lemma 22.1).* For a point treatment data structure $O = (L(0), A(0), Y)$ and binary treatment $A(0)$, we have for a rule $V \to d(V)$, $E_0 Y_d = E_0 Y_0 + E_0 d(V)\bar{Q}_0(V))$ with $\bar{Q}_0(V) = E_0[Y_1 - Y_0 \mid V]$. This identity is applied twice in the following derivation:

$$\begin{aligned}\Psi(P_0) =&E_0 Y_{(0,1),d_{0,A(1)}} + E_0 d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0))\\ =&E_0 E_0[Y_{(0,1),d_{0,A(1)}} \mid V_{(0,1)}(1)] + E_0 d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0))\\ =&E_0 E_0[Y_{(0,1),(0,1)} \mid V_{(0,1)}(1)] + E_0 I(\bar{Q}_{b,20}((0,1), V_{(0,1)}(1)) > 0)\bar{Q}_{b,20}(0, V_{(0,1)}(1))\\ &+ E_0 d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0))\\ =&E_0 E_0[Y_{(0,1),(0,1)} \mid V_{(0,1)}(1)] + E_0 d_{0,A(1)}((0,1), V_{(0,1)}(1))\bar{Q}_{b,20}(0, V_{(0,1)}(1))\\ &+ E_0 d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0))\\ =&E_0 Y_{(0,1),(0,1)} + E_0 d_{0,A(1)}((0,1), V_{(0,1)}(1))\bar{Q}_{b,20}(0, V_{(0,1)}(1))\\ &+ E_0 d_{0,A(0)}(V(0))\bar{Q}_{b,10}(V(0)).\end{aligned}$$

*Proof (Theorem 22.5).* In this proof we will omit the dependence of $d_{0,A(0)}$, $d_{Q,A(0)}$, $\bar{Q}_{b,10}$, and $\bar{Q}_{b,1}$ on $V(0)$ in the notation. This part of the proof mimics the proof of Lemma 5.2 in Audibert and Tsybakov (2007). For any $t > 0$,

$$\begin{aligned}|R_{2,A(0)}(Q, Q_0)| =&E_0[|\bar{Q}_{b,10}|I(d_{0,A(0)} \neq d_{Q,A(0)})]\\ =&E_0[|\bar{Q}_{b,10}|I(d_{0,A(0)} \neq d_{Q,A(0)})I(0 < |\bar{Q}_{b,10}| \leq t)]\\ &+ E_0[|\bar{Q}_{b,10}|I(d_{0,A(0)} \neq d_{Q,A(0)})I(|\bar{Q}_{b,10}| > t)]\\ \leq&E_0[|\bar{Q}_{b,1} - \bar{Q}_{b,10}|I(0 < |\bar{Q}_{b,10}| \leq t)]\\ &+ E_0[|\bar{Q}_{b,1} - \bar{Q}_{b,10}|I(|\bar{Q}_{b,1} - \bar{Q}_{b,10}| > t)]\\ \leq&\left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{2,P_0} \Pr(0 < |\bar{Q}_{b,10}| \leq t)^{1/2} + \frac{\left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{2,P_0}^2}{t}\\ \leq&\left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{2,P_0} C_0^{1/2} t^{\beta_1/2} + \frac{\left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{2,P_0}^2}{t},\end{aligned}$$

where the first inequality holds because $d_{0,A(0)} \neq d_{Q,A(0)}$ implies that $|\bar{Q}_{b,1} - \bar{Q}_{b,10}| > |\bar{Q}_{b,10}|$, the second inequality holds by the Cauchy-Schwarz and Markov inequalities, and the third inequality holds by (MA1). The first result follows by optimizing over $t$ to find that the upper bound is minimized when $t = C' \left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{2,P_0}^{2(1+\beta_1)/(2+\beta_1)}$ for a constant $C'$ that depends on $C_0$ and $\beta_1$.

We now establish the supremum-norm result. Note that

$$\begin{aligned}|R_{2,A(0)}(Q, Q_0)| &= E_0 \left|I(d_{Q,A(0)} \neq d_{0,A(0)})\bar{Q}_{b,10}\right| \\ &\leq E_0 \left[I(0 < |\bar{Q}_{b,10}| \leq |\bar{Q}_{b,1} - \bar{Q}_{b,10}|)|\bar{Q}_{b,10}|\right] \\ &\leq E_0 \left[I\left(0 < |\bar{Q}_{b,10}| \leq \left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{\infty,P_0}\right)|\bar{Q}_{b,10}|\right] \\ &\leq \left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{\infty,P_0} \Pr\left(0 < |\bar{Q}_{b,10}| \leq \left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{\infty,P_0}\right).\end{aligned}$$

By (MA1), $|\Psi_{d_{Q,A(0)}}(P_0) - \Psi_{d_{0,A(0)}}(P_0)| \leq C_1 \left\|\bar{Q}_{b,1} - \bar{Q}_{b,10}\right\|_{\infty,P_0}^{1+\beta_1}$. Combining the two results yields (22.9). The proof of (22.10) is analogous and so is omitted.

## 22.6 CV-TMLE for the Mean Outcome Under Data-Adaptive $V$-Optimal Rule

Let $\hat{d} : \mathcal{M} \to \mathcal{D}$ be an estimator of the $V$-optimal rule $d_0$. Firstly, without loss of generality we can assume that $Y \in [0, 1]$. Denote the realizations of $B_n$ with $j = 1, \ldots, J$, and let $d_{nj} \equiv \hat{d}(P^0_{n,j})$ denote the estimated rule on training sample $j$. Let

$$(a(0), \bar{l}(1)) \mapsto E_{nj}[Y|\bar{A}(1) = d_{nj}(a(0), v), \bar{L}(1) = \bar{l}(1)] \tag{22.11}$$

represent an initial estimate of $E_0[Y \mid \bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)]$ based on the training sample $j$. Similarly, let $g_{nj}$ represent the estimated intervention mechanism based on this training sample $P^0_{n,j}$, $j = 1, \ldots, J$. Consider the fluctuation submodel

$$\begin{aligned}\operatorname{logit} E^{(\epsilon_2)}_{nj} \left[Y|\bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)\right] = &\operatorname{logit} E_{nj} \left[Y|\bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)\right] \\ &+ \epsilon_2 H_2(g_{nj})(O)\end{aligned}$$

where

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

Note that the fluctuation $\epsilon_2$ does not rely on $j$. Let

$$\epsilon_{2n} = \arg\min_{\epsilon_2} \frac{1}{J} \sum_{j=1}^{J} P^1_{n,j} \tilde{\phi}(E^{(\epsilon_2)}_{nj}),$$

where $E_{nj}^{(\epsilon_2)}$ refers to the represents the fluctuated function in (22.11) and

$$-\tilde{\phi}(f)(o) = y \log f(o) + (1 - y) \log (1 - f(o)) . \tag{22.12}$$

for all $f : \mathcal{O} \to (0, 1)$. For each $i = 1, \ldots, n$, let $j(i) \in \{1, \ldots, J\}$ represent the value of $B_n$ for which element $i$ is in the validation set. The fluctuation $\epsilon_{2n}$ can be obtained by fitting a univariate logistic regression of $(y_i : i = 1, \ldots, n)$ on $(H_2(g_{nj(i)})(o_i) : i = 1, \ldots, n)$ using

$$\left(\text{logit}\, E_{nj(i)} \left[Y|\bar{A}(1) = d_{nj(i)}(a(0)_i, v_i), \bar{L}(1) = \bar{l}(1)_i\right] : i = 1, \ldots, n\right)$$

as offset. Thus each observation $i$ is paired with nuisance parameters are fit on the training sample that does not contain observation $i$. This defines a targeted estimate

$$E_{nj}^{*} \left[Y|\bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)\right] \equiv E_{nj}^{(\epsilon_{2n})} \left[Y|\bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)\right] \tag{22.13}$$

of $E_0[Y \mid \bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)]$. We note that this targeted estimate only depends on $P_n$ through the training sample $P_{n,j}^{0}$ and the one-dimensional $\epsilon_{2n}$.

We now aim to get a targeted estimate of $E_0[Y_{d_{nj}}|L(0)]$. We can obtain an estimate

$$(a_1(0), l(0)) \mapsto E_{nj} \left[ E_{nj} \left[Y \mid \bar{A}(1) = d_{nj}(A(0), V), \bar{L}(1)\right] \middle| A(0) = (a_1(0), 1), L(0) = l(0)\right] \tag{22.14}$$

by regressing $E_{nj} \left[Y \mid \bar{A}(1) = d_{nj}(A(0)_i, V_i), \bar{L}(1)_i\right]$ against $A(0)_i, L(0)_i$ for all of the observations $i$ in training sample $j$. For an estimate $E_{nj}[Y_{d_{nj}}|L(0)]$ of $E_0[Y_{d_{nj}}|L(0)]$, we can use the regression function above but with $a(0)$ fixed to $d_{nj,A(0)}(v(0))$.

Consider the fluctuation submodel

$$\text{logit}\, E_{nj}^{(\epsilon_1)} \left[Y_{d_{nj}} \mid L(0)\right] = \text{logit}\, E_{nj} \left[Y_{d_{nj}} \mid L(0)\right] + \epsilon H_1(g_{nj})(O),$$

where

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

Again the fluctuation $\epsilon_1$ does not rely on $j$. Let

$$\epsilon_{1n} = \arg \min_{\epsilon_1} \frac{1}{J} \sum_{j=1}^{J} P_{n,j}^{1} \tilde{\phi}(E_{nj}^{(\epsilon_1)}),$$

where $\tilde{\phi}$ is defined in (22.12). For each $i = 1, \ldots, n$, again let $j(i) \in \{1, \ldots, J\}$ represent the value of $B_n$ for which element $i$ is in the validation set. The fluctuation $\epsilon_{1n}$ can be obtained by fitting a univariate logistic regression of

$$\left(E_{nj(i)}^{*} \left[Y|\bar{A}(1) = d_{nj(i)}(a(0)_i, v_i), \bar{l}(1)_i\right] : i = 1, \ldots, n\right)$$

on $(H_1(g_{nj(i)})(o_i) : i = 1, \ldots, n)$ using

$$\left(\operatorname{logit} E_{nj(i)}\left[Y_{d_{nj(i)}} | L(0) = l(0)_i\right] : i = 1, \ldots, n\right)$$

as offset. This defines a targeted estimate

$$E^*_{nj}\left[Y_{d_{nj}} | L(0)\right] \equiv E^{(\epsilon_{1n})}_{nj}\left[Y_{d_{nj}} | L(0)\right] \tag{22.15}$$

of $E_0[Y_{d_{nj}} | L(0)]$. We note that this targeted estimate only depends on $P_n$ through the training sample $P^0_{n,j}$ and the one-dimensional $\epsilon_{1n}$.

Let $Q_{L(0),nj}$ be the empirical distribution of $L(0)_i$ for the validation sample $P^1_{n,j}$. For all $j = 1, \ldots, J$, let $Q^{d_{nj}*}_{nj}$ be the parameter mapping representing the collection containing $Q_{L(0),nj}$ and the targeted regressions in (22.13) and (22.15). This defines an estimator $\psi^*_{nj} = P^1_{n,j}\bar{Q}^*_{b,1nj}$ of $\psi_{d_{nj}0} = \Psi_{d_{nj}}(P_0)$ for each $j = 1, \ldots, J$. The cross-validated TMLE is now defined as $\psi^*_n = \frac{1}{J}\sum_{j=1}^J \psi^*_{nj}$. This CV-TMLE solves the cross-validated efficient influence curve equation:

$$\frac{1}{J}\sum_{j=1}^{J} P^1_{n,j} D^*(d_{nj}, Q^{d_{nj}*}_{nj}, g_{nj}) = 0.$$

Further, each $Q^{d_{nj}*}_{nj}$ only relies on $P^1_{n,j}$ through the univariate parameters $\epsilon_{1n}$ and $\epsilon_{2n}$. This will allow us to use the entropy integral arguments presented in Zheng and van der Laan (2010) that show that no restrictive empirical process conditions are needed on the initial estimates in (22.11) and (22.14).

The only modification relative to the original CV-TMLE presented in Zheng and van der Laan (2010) is that in the above description we change our target on each training sample into the training sample specific target parameter implied by the fit $\hat{d}(P^0_{n,B_n})$ on the training sample, while in the original CV-TMLE formulation, the target would still be $\Psi_{d_0}(P_0)$. With this minor twist, the (same) CV-TMLE is now used to target the average of training sample specific target parameters averaged across the $J$ training samples.

## 22.7 Notes and Further Reading

Examples of multiple time-point dynamic treatment regimes are given in Lavori and Dawson (2000, 2008); Murphy (2005); Rosthø j et al. (2006); Thall et al. (2002); Wagner et al. (2001) ranging from rules that change the dose of a drug, change or augment the treatment, to making a decision on when to start a new treatment, in response to the history of the subject. For an excellent overview on dynamic treatments we refer to Chakraborty and Moodie (2013).

We define the optimal dynamic multiple time-point treatment regime 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 shows that we can describe the optimal dynamic treatment among *all* dynamic treatments in a sequential manner (Murphy 2003; Robins 2004; Murphy 2005). The optimal rule can be learned through fitting the likelihood and then calculating the optimal rule under this fit of the likelihood. This approach can be implemented with maximum likelihood estimation based on parametric models. It has been noted (e.g., Robins 2004) that the estimator of the parameters of one of the regressions (except the first one) when using parametric regression models is a nonsmooth function of the estimator of the parameters of the previous regression, and that this results in nonregularity of the estimators of the parameter vector. This raises challenges for obtaining statistical inference, even when assuming that these parametric regression models are correctly specified. Chakraborty and Moodie (2013) discuss various approaches and advances that aim to resolve this delicate issue such as inverting hypothesis testing (Robins 2004), establishing nonnormal limit distributions of the estimators (Laber et al. 2014a), or using the $m$ out of $n$ bootstrap (Chakraborty et al. 2014). The proof of the fast rate for the estimate of the optimal rule provided in Theorem 22.5 is similar to the proofs of the fast classification rates obtained in Audibert and Tsybakov (2007). It was presented for single time point optimal treatment rules in van der Laan and Luedtke (2015).

Murphy (2003) and Robins (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. Statistical inference for the parameters of the blip function proceeds accordingly, but Robins (2004) points out the irregularity of the estimator, resulting in some serious challenges for statistical inference as referenced above. Structural nested mean models have also been generalized to blip functions that condition on a (counterfactual) subset of the past, thereby allowing the learning of optimal rules that are restricted to only using this subset of the past (Robins 2004 and Section 6.5 in van der Laan and Robins 2003).

Each of the above referenced approaches for learning an optimal dynamic treatment that also aims to provide statistical inference relies on parametric assumptions: obviously, $Q$-learning based on parametric models, but also the structural nested mean model rely on parametric models for the blip function. As a consequence, even in a SMART, the statistical inference for the optimal dynamic treatment heavily relies on assumptions that are generally believed to be false, and will thus be expected to be biased. To avoid these biases, in this chapter we defined our model as nonparametric, beyond possible restrictions on the treatment/censoring mechanism.