6. Optimal Multiple Decision Treatment Regimes

Introduction

Consider: The class of Ψ-specific regimes $\mathcal{D}$ for a given specification $\Psi = (\Psi_1, \ldots, \Psi_K)$ of feasible sets

Recall: An optimal regime $d^{opt} \in \mathcal{D}$ is one that satisfies

$$E\{Y^*(d^{opt})\} \geq E\{Y^*(d)\} \text{ for all } d \in \mathcal{D}$$

- We first characterize an optimal regime d^{opt} in terms of potential outcomes using the principle of *backward induction* and show that it satisfies this condition
- Under SUTVA, SRA, and positivity, we show that a optimal regime can be expressed equivalently in terms of observed data
- We then present several methods for estimation of an optimal regime

Recap: Ψ-specific regimes

Reminder: We summarize the definitions for given Ψ for convenience

- Define

$$\Gamma_1 = \{x_1 \in \mathcal{X}_1 \text{ satisfying } P(X_1 = x_1) > 0\}$$

and for $k = 2, \ldots, K$

$$\Lambda_k = \{(\overline{x}_k, \overline{a}_k) \;\text{ such that }\; (\overline{x}_k, \overline{a}_{k-1}) = h_k \in \Gamma_k,\; a_k \in \Psi_k(h_k)\}$$

$$\Gamma_k = \Big[h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \overline{\mathcal{X}}_k \times \overline{\mathcal{A}}_{k-1} \text{ satisfying } (\overline{x}_{k-1}, \overline{a}_{k-1}) \in \Lambda_{k-1}$$
$$\text{and } P\{X_k^*(\overline{a}_{k-1}) = x_k \mid \overline{X}_{k-1}^*(\overline{a}_{k-2}) = \overline{x}_{k-1}\} > 0\Big]$$

- $\Gamma_k \subseteq \mathcal{H}_k$ contains all possible histories h_k consistent with having followed a Ψ-specific regime through Decision $k-1$
- Λ_k is the set of all possible histories $h_k \in \Gamma_k$ and associated treatment options in $\Psi_k(h_k)$ at Decision k

Recap: Ψ-specific regimes

- $Y^*(\overline{a})$ and observed outcome Y take values $y \in \mathcal{Y}$

$$\Gamma_{K+1} = \Big[(\overline{x},\overline{a},y) \in \overline{\mathcal{X}} \times \overline{\mathcal{A}} \times \mathcal{Y} \text{ satisfying } (\overline{x},\overline{a}) \in \Lambda_K \text{ and } P\{Y^*(\overline{a}) = y \mid \overline{X}_K^*(\overline{a}_{K-1}) = \overline{x}_K\} > 0\Big]$$

Ψ-specific regime: d comprises rules $d_k(h_k)$ such that

$$d_k : \ \Gamma_k \rightarrow \mathcal{A}_k$$

satisfying $d_k(h_k) \in \Psi_k(h_k)$ for every $h_k \in \Gamma_k$, $k = 1, \ldots, K$

- With ℓ_k distinct subsets $\mathcal{A}_{k,l} \subseteq \mathcal{A}_k$, $l = 1, \ldots, \ell_k$, that are feasible sets at Decision k

$$d_k(h_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\}\, d_{k,l}(h_k)$$

  where $d_{k,l}$ maps from $\Gamma_{k,l} \subseteq \Gamma_k$ to $\mathcal{A}_{k,l}$ and thus $\mathcal{A}_k$, and $d_{k,l}(h_k) \in \Psi_k(h_k)$

Recap: Ψ-specific regimes

Observed data: $(X_1, A_1, X_2, A_2, \ldots, X_K, A_K, Y)$

Under: SUTVA (5.10), SRA (5.11), and positivity (5.15)

$$
\begin{aligned}
P(A_k = a_k | H_k = h_k) &= P(A_k = a_k | \overline{X}_k = \overline{x}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \\
\text{for } h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \Gamma_k, &\quad \text{and} \quad a_k \in \Psi_k(h_k) = \Psi_k(\overline{x}_k, \overline{a}_{k-1}), \\
k = 1, \ldots, K &
\end{aligned} \tag{6.1}
$$

- Γ_k, $k = 2, \ldots, K$, and Γ_{K+1} can be written equivalently

$$
\Gamma_k = \Big[h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \overline{\mathcal{X}}_k \times \overline{\mathcal{A}}_{k-1} \text{ satisfying } (\overline{x}_{k-1}, \overline{a}_{k-1}) \in \Lambda_{k-1} \\
\text{and } P(X_k = x_k \mid \overline{X}_{k-1} = \overline{x}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \Big]
$$

$$
\Gamma_{K+1} = \Big[(\overline{x}, \overline{a}, y) \in \overline{\mathcal{X}} \times \overline{\mathcal{A}} \times \mathcal{Y} \text{ satisfying } (\overline{x}, \overline{a}) \in \Lambda_K \text{ and} \\
P(Y = y \mid \overline{X} = \overline{x}, \overline{A}_{K-1} = \overline{a}_{K-1}) > 0 \Big]
$$

Characterization in terms of potential outcomes

Potential outcomes: Recall

$$W^* = \Big\{ X_2^*(a_1), X_3^*(\bar{a}_2), \ldots, X_K^*(\bar{a}_{K-1}), Y^*(\bar{a}),$$
$$\text{for } a_1 \in \mathcal{A}_1, \bar{a}_2 \in \overline{\mathcal{A}}_2, \ldots, \bar{a}_{K-1} \in \overline{\mathcal{A}}_{K-1}, \bar{a} \in \overline{\mathcal{A}} \Big\}.$$

- We first characterize an optimal regime d^{opt} in terms of baseline information X_1 and W^*
- To understand the *backward inductive* reasoning, it first suffices to consider $K = 2$

$K = 2$ **decisions:** Consider a randomly chosen individual

Characterization in terms of potential outcomes

At Decision 2 (final decision point):

- If she started with realized baseline info $X_1 = x_1 = h_1 \in \Gamma_1$ and received option $a_1 \in \Psi_1(x_1) = \Psi_1(h_1)$ at Decision 1, she *already will have achieved* intervening info $X_2^*(a_1)$
- Thus, with Decision 1 option a_1 and $\overline{X}_2^*(a_1) = \{X_1, X_2^*(a_1)\}$ *already determined* and with realized value $\overline{x}_2 = (x_1, x_2)$, $h_2 = (\overline{x}_2, a_1) \in \Gamma_2$, the *optimal decision* at Decision 2 is to choose the option $a_2 \in \Psi_2(h_2)$ that would result in the largest expected outcome *given that she is already at this point*
- The outcome that she *would achieve* under $a_2 \in \Psi_2(h_2)$, having already received a_1 at Decision 1, is $Y^*(a_1, a_2)$
- Her *expected outcome given where she is now* is thus

$$E\{Y^*(a_1, a_2) | \overline{X}_2^*(a_1) = \overline{x}_2\} \tag{6.2}$$

well defined because $(\overline{x}_2, \overline{a}_2) \in \Lambda_2$, $(\overline{x}, \overline{a}, y) \in \Gamma_3$

Characterization in terms of potential outcomes

Optimal Ψ-specific rule at Decision 2: Should select $a_2 \in \Psi_2(h_2)$ to *maximize* (6.2); i.e.,

$$d_2^{opt}(h_2) = d_2^{opt}(\overline{x}_2, a_1) = \underset{a_2 \in \Psi_2(h_2)}{\arg\max}\; E\{Y^*(a_1, a_2) | \overline{X}_2^*(a_1) = \overline{x}_2\} \quad (6.3)$$

- Takes $h_2 = (\overline{x}_2, a_1) = (x_1, x_2, a_1) \in \Gamma_2$ as input and chooses the option in $\Psi_2(h_2)$ maximizing expected outcome given this history
- The resulting maximum expected outcome is

$$V_2(h_2) = V_2(\overline{x}_2, a_1) = \max_{a_2 \in \Psi_2(h_2)} E\{Y^*(a_1, a_2) | \overline{X}_2^*(a_1) = \overline{x}_2\} \quad (6.4)$$

Characterization in terms of potential outcomes

At Decision 1:

- If an individual presents with baseline information $X_1 = x_1 = h_1 \in \Gamma_1$, the optimal decision *now* is to choose $a_1 \in \Psi_1(h_1)$ that maximizes her expected outcome given $X_1 = x_1$, *taking into account that she will receive treatment at Decision 2 by following the optimal rule* d_2^{opt} in (6.3)
- If $a_1 \in \Psi_1(h_1)$ is selected now, she *will present* at Decision 2 with

$$\overline{X}_2^*(a_1) = \{x_1, X_2^*(a_1)\}$$

  (with X_1 at its realized value x_1)
- Upon receiving an option in $\Psi_2\{x_1, X_2^*(a_1), a_1\}$ according to d_2^{opt}, her expected outcome *given this info* is, from (6.4)

$$V_2\{x_1, X_2^*(a_1), a_1\} = \max_{a_2 \in \Psi_2\{x_1, X_2^*(a_1), a_1\}} E\{Y^*(a_1, a_2) | X_2^*(a_1), X_1 = x_1\}$$

Characterization in terms of potential outcomes

Optimal Ψ-specific rule at Decision 1: Should select $a_1 \in \Psi_1(h_1)$ to make the expected value of $V_2\{x_1, X_2^*(a_1), a_1\}$ given X_1 evaluated at $X_1 = x_1$ *as large as possible*, i.e.

$$d_1^{opt}(h_1) = d_1^{opt}(x_1) = \arg\max_{a_1 \in \Psi_1(h_1)} E[V_2\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1] \tag{6.5}$$

well defined because $(x_1, a_1) \in \Lambda_1$ and $X_2^*(a_1)$ takes values in Γ_2

- (6.5) selects $a_1 \in \Psi_1(h_1)$ to *maximize the maximum expected outcome* that would result from choosing the option at Decision 2 optimally given the history available at that point
- Resulting maximum of the maximum expected outcome an individual with realized baseline info x_1 *would achieve under rules d_1^{opt} and d_2^{opt}* is

$$V_1(h_1) = V_1(x_1) = \max_{a_1 \in \Psi_1(h_1)} E[V_2\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1]$$

$$= \max_{a_1 \in \Psi_1(h_1)} E\left[\max_{a_2 \in \Psi_2\{x_1, X_2^*(a_1), a_1\}} E\{Y^*(a_1, a_2) | X_2^*(a_1), X_1 = x_1\} \middle| X_1 = x_1 \right]$$

Characterization in terms of potential outcomes

Result:

- Clearly, $d^{opt} = (d_1^{opt}, d_2^{opt})$ defined by (6.3) and (6.5) is a regime in $\mathcal{D}$ (a set of decision rules, each mapping history to feasible treatment options)
- Intuition suggests that it is an *optimal regime* satisfying

$$E\{Y^*(d^{opt})\} \geq E\{Y^*(d)\} \text{ for all } d \in \mathcal{D}$$

but this must be shown formally (sketch coming up)

K **Decisions:** This backward inductive reasoning extends

- Consider a randomly chosen individual presenting at baseline with info $X_1 = x_1 = h_1 \in \Gamma_1$

Characterization in terms of potential outcomes

At Decision K: He has *already received* $a_k \in \Psi_k(h_k)$, $k = 1, \ldots, K-1$, based on $\overline{X}^*_{K-1}(\overline{a}_{K-2}) = \{X_1, X_2^*(a_1), \ldots, X^*_{K-1}(\overline{a}_{K-2})\}$ and has info $X^*_K(\overline{a}_{K-1})$ that has accrued since Decision $K-1$

- The optimal decision, with $\overline{a}_{K-1}$ and $\overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K$ *already determined*, so that $h_K = (\overline{x}_K, \overline{a}_{K-1}) \in \Gamma_K$, is to choose $a_K \in \Psi_K(h_K)$ such that his expected outcome *given he is at this point* is largest, i.e., maximize in a_K

$$E\{Y^*(\overline{a}_{K-1}, a_K) \mid \overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K\}$$

- Thus define for $h_K = (\overline{x}_K, \overline{a}_{K-1}) \in \Gamma_K$

$$d_K^{(1)opt}(h_K) = \underset{a_K \in \Psi_K(h_K)}{\arg\max}\, E\{Y^*(\overline{a}_{K-1}, a_K) | \overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K\} \qquad (6.6)$$

- Maximum expected outcome achieved

$$V_K^{(1)}(h_K) = \max_{a_K \in \Psi_K(h_K)} E\{Y^*(\overline{a}_{K-1}, a_K) | \overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K\} \qquad (6.7)$$

Characterization in terms of potential outcomes

At Decision $K-1$: He presents with $\overline{X}^*_{K-1}(\overline{a}_{K-2})$ *already determined* with realized value $\overline{x}_{K-1}$ following options $a_k \in \Psi_k(h_k)$, $k = 1, \ldots, K-2$, such that $h_{K-1} = (\overline{x}_{K-1}, \overline{a}_{K-2}) \in \Gamma_{K-1}$

- Now select the option in $\Psi_{K-1}(h_{K-1})$ to maximize his expected outcome given this history and *acknowledging that he will receive treatment at Decision K by following* $d_K^{(1)opt}$ in (6.6)

- If $a_{K-1} \in \Psi_{K-1}(h_{K-1})$ is selected *now*, he *will present* at Decision K with information

$$\overline{X}^*_K(\overline{a}_{K-1}) = \{\overline{x}_{K-1}, X^*_K(\overline{a}_{K-1})\}$$

and, upon receiving an option in $\Psi_K\{\overline{x}_{K-1}, X^*_K(\overline{a}_{K-1}), \overline{a}_{K-1}\}$ dictated by $d_K^{(1)opt}$, will have expected outcome *given this information* equal to

$$V_K^{(1)}\{\overline{x}_{K-1}, X^*_K(\overline{a}_{K-1}), \overline{a}_{K-1}\}$$
$$= \max_{a_K \in \Psi_K\{\overline{x}_{K-1}, X^*_K(\overline{a}_{K-1}), \overline{a}_{K-1}\}} E\{Y^*(\overline{a}_{K-1}, a_K) | X^*_K(\overline{a}_{K-1}), \overline{X}^*_{K-1}(\overline{a}_{K-2}) = \overline{x}_{K-1}\}$$

from (6.7)

Characterization in terms of potential outcomes

- Thus, $a_{K-1} \in \Psi_{K-1}(h_{K-1})$ should be chosen to make the expected value of $V_K^{(1)}\{\overline{x}_{K-1}, X_K^*(\overline{a}_{K-1}), \overline{a}_{K-1}\}$ given $\overline{X}_{K-1}^*(\overline{a}_{K-2}) = \overline{x}_{K-1}$ *as large as possible*, leading to

$$
\begin{aligned}
& d_{K-1}^{(1)opt}(h_{K-1}) \\
& = \underset{a_{K-1} \in \Psi_{K-1}(h_{K-1})}{\arg\max} E[V_K^{(1)}\{\overline{x}_{K-1}, X_K^*(\overline{a}_{K-2}, a_{K-1}), \overline{a}_{K-2}, a_{K-1}\} | \\
& \qquad \overline{X}_{K-1}^*(\overline{a}_{K-2}) = \overline{x}_{K-1}]
\end{aligned} \tag{6.8}
$$

- And maximum expected outcome achieved

$$
\begin{aligned}
& V_{K-1}^{(1)}(h_{K-1}) \\
& = \max_{a_{K-1} \in \Psi_{K-1}(h_{K-1})} E[V_K^{(1)}\{\overline{x}_{K-1}, X_K^*(\overline{a}_{K-2}, a_{K-1}), \overline{a}_{K-2}, a_{K-1}\} | \\
& \qquad \overline{X}_{K-1}^*(\overline{a}_{K-2}) = \overline{x}_{K-1}]
\end{aligned}
$$

Characterization in terms of potential outcomes

At Decisions $k = K-1,\ldots,2$: Continuing this reasoning, for $\overline{x}_k \in \overline{\mathcal{X}}_k$, $\overline{a}_{k-1} \in \overline{\mathcal{A}}_{k-1}$ for which $h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \Gamma_k$

$$d_k^{(1)opt}(h_k) \tag{6.9}$$
$$= \underset{a_k \in \Psi_k(h_k)}{\arg\max}\, E[V_{k+1}^{(1)}\{\overline{x}_k, X_{k+1}^*(\overline{a}_{k-1}, a_k), \overline{a}_{k-1}, a_k\} | \overline{X}_k^*(\overline{a}_{k-1}) = \overline{x}_k]$$

$$V_k^{(1)}(h_k) = \max_{a_k \in \Psi_k(h_k)} E[V_{k+1}^{(1)}\{\overline{x}_k, X_{k+1}^*(\overline{a}_{k-1}, a_k), \overline{a}_{k-1}, a_k\} | \overline{X}_k^*(\overline{a}_{k-1}) = \overline{x}_k] \tag{6.10}$$

At Decision 1:

$$d_1^{(1)opt}(x_1) = \underset{a_1 \in \Psi_1(h_1)}{\arg\max}\, E[V_2^{(1)}\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1] \tag{6.11}$$

$$V_1^{(1)}(x_1) = \max_{a_1 \in \Psi_1(h_1)} E[V_2^{(1)}\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1] \tag{6.12}$$

Note: All conditional expectations in (6.6)–(6.12) are well defined because $h_k \in \Gamma_k$, $k = 1,\ldots,K$

Characterization in terms of potential outcomes

Equivalent representation: By the definitions of $d_k^{(1)opt}(h_k)$, $k = K, \ldots, 1$ in (6.6)– (6.11), can write

$$V_K^{(1)}(h_K) = E[Y^*\{\overline{a}_{K-1}, d_K^{(1)opt}(h_K)\}|\overline{X}_K^*(\overline{a}_{K-1}) = \overline{x}_K]$$

$$V_k^{(1)}(h_k) =$$

$$E\left(V_{k+1}^{(1)}\left[\overline{x}_k, X_{k+1}^*\{\overline{a}_{k-1}, d_k^{(1)opt}(h_k)\}, \overline{a}_{k-1}, d_k^{(1)opt}(h_k)\right]\Big| \overline{X}_k^*(\overline{a}_{k-1}) = \overline{x}_k\right)$$

$$k = K-1, \ldots, 2$$

$$V_1^{(1)}(h_1) = E\left(V_2^{(1)}\left[x_1, X_2^*\{d_1^{(1)opt}(h_1)\}, d_1^{(1)opt}(h_1)\right]\Big| X_1 = x_1\right)$$

Characterization in terms of potential outcomes

Result: $d^{(1)opt} = (d_1^{(1)opt}, \ldots, d_K^{(1)opt})$ defined above is clearly a treatment regime in $\mathcal{D}$

- A set of rules, each using individual history to select treatment from among the feasible options
- Intuition suggests that it is an optimal regime, but this must *shown formally* (sketch coming up)
- *Uniqueness:* At any decision point $k = 1, \ldots, K$, for some $h_k \in \Gamma_k$, there may be *more than one option* in $\Psi_k(h_k)$ achieving the maximum in (6.7), (6.10), or (6.12)
- A unique representation of $d_k^{(1)opt}$ is defined by choosing one of the options in $\Psi_k(h_k)$ as the default

Note: The superscript "(1)" emphasizes that the rules pertain to an individual presenting at Decision 1 (rather than "midstream" at a later decision point)

Justification

Formally: Confirm that $d^{(1)opt} \in \mathcal{D}$ defined in (6.6), (6.8), (6.9) for $k = K-2, \ldots, 2$, and (6.11) is an optimal regime in $\mathcal{D}$; i.e., show that

$$E\{Y^*(d^{(1)opt})\} \geq E\{Y^*(d)\} \quad \text{for all} \;\; d \in \mathcal{D} \tag{6.13}$$

- For any $d \in \mathcal{D}$, from (5.7)

$$Y^*(d) = \sum_{\overline{a} \in \overline{\mathcal{A}}} Y^*(\overline{a}) \prod_{j=1}^{K} \mathrm{I}\left[d_j\{\overline{X}_j^*(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j\right]$$
$$= \sum_{\overline{a}_{K-1} \in \overline{\mathcal{A}}_{K-1}} \left(\prod_{j=1}^{K-1} \mathrm{I}\left[d_j\{\overline{X}_j^*(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j\right] Y^*\left[\overline{a}_{K-1}, d_K\{\overline{X}_K^*(\overline{a}_{K-1}), \overline{a}_{K-1}\}\right] \right) \tag{6.14}$$

- $Y^*\left[\overline{a}_{K-1}, d_K\{\overline{X}_K^*(\overline{a}_{K-1}), \overline{a}_{K-1}\}\right]$ in (6.14) is the potential outcome if an individual were to receive $\overline{a}_{K-1}$ at Decisions 1 to $K-1$ and then receive the option at Decision K dictated by d_K for this treatment history and the associated intervening information $\overline{X}_K^*(\overline{a}_{K-1})$

Justification

- It follows from (6.14) that

$$E\{Y^*(d)\} = \sum_{\overline{a}_{K-1} \in \overline{\mathcal{A}}_{K-1}} E\left\{ \prod_{j=1}^{K-1} \mathrm{I}\left[d_j\{\overline{X}_j^*(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j \right] \right.$$
$$\left. \times\, E\left(Y^*\left[\overline{a}_{K-1}, d_K\{\overline{X}_K^*(\overline{a}_{K-1}), \overline{a}_{K-1}\}\right] \middle| \overline{X}_K^*(\overline{a}_{K-1}) \right) \right\} \qquad (6.15)$$

- Because d and $d^{(1)opt}$ are regimes in $\mathcal{D}$, the set of rules $(\overline{d}_{K-1}, d_K^{(1)opt})$ is also a regime in $\mathcal{D}$
- Thus, as in (6.14) and (6.15)

$$E\{Y^*(\overline{d}_{K-1}, d_K^{(1)opt})\} = \sum_{\overline{a}_{K-1} \in \overline{\mathcal{A}}_{K-1}} E\left\{ \prod_{j=1}^{K-1} \mathrm{I}\left[d_j\{\overline{X}_j^*(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j \right] \right.$$
$$\left. \times\, E\left(Y^*\left[\overline{a}_{K-1}, d_K^{(1)opt}\{\overline{X}_K^*(\overline{a}_{K-1}), \overline{a}_{K-1}\}\right] \middle| \overline{X}_K^*(\overline{a}_{K-1}) \right) \right\} \qquad (6.16)$$

Justification

- Conditional on $\overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K$ for $h_K = (\overline{x}_K, \overline{a}_{K-1}) \in \Gamma_K$, $d_K(h_K) \in \Psi_K(h_K)$; thus, from the definition of $d_K^{(1)opt}(h_K)$ in (6.6)

$$E\left[Y^* \{\overline{a}_{K-1}, d_K(\overline{x}_K, \overline{a}_{K-1})\} \middle| \overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K \right]$$
$$\leq E\left[Y^* \left\{\overline{a}_{K-1}, d_K^{(1)opt}(\overline{x}_K, \overline{a}_{K-1})\right\} \middle| \overline{X}^*_K(\overline{a}_{K-1}) = \overline{x}_K \right]$$

- Comparing (6.16) to (6.15), it follows that

$$E\{Y^*(d)\} \leq E\{Y^*(\overline{d}_{K-1}, d_K^{(1)opt})\} \tag{6.17}$$

- By the definition of $V_K^{(1)}(h_K)$ in (6.7), (6.16) can be written as

$$E\{Y^*(\overline{d}_{K-1}, d_K^{(1)opt})\} \tag{6.18}$$
$$= \sum_{\overline{a}_{K-1} \in \overline{\mathcal{A}}_{K-1}} E\left(\prod_{j=1}^{K-1} \mathrm{I}\left[d_j\{\overline{X}^*_j(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j\right] V_K^{(1)}\{\overline{X}^*_K(\overline{a}_{K-1}), \overline{a}_{K-1}\} \right)$$

Justification

- For brevity define $\mathcal{J}_k^d = \prod_{j=1}^{k} \mathrm{I}\left[d_j\{\overline{X}_j^*(\overline{a}_{j-1}), \overline{a}_{j-1}\} = a_j \right], \quad k = 1, \ldots, K$

$\mathcal{K}_k(d_{k-1}) = \left[\overline{a}_{k-2}, d_{k-1}\{\overline{X}_{k-1}^*(\overline{a}_{k-2}), \overline{a}_{k-2}\} \right], \quad k = K, \ldots, 3 \quad \mathcal{K}_2(d_1) = d_1(X_1)$

- Using this notation, (6.18) can be written

$$E\{Y^*(\overline{d}_{K-1}, d_K^{(1)opt})\} = \sum_{\overline{a}_{K-2} \in \overline{\mathcal{A}}_{K-2}} E\left\{ \mathcal{J}_{K-2}^d \right.$$
$$\left. \times\ E\left(V_K^{(1)} \left[\overline{X}_K^*\{\mathcal{K}_K(d_{K-1})\}, \mathcal{K}_K(d_{K-1}) \right] \Big| \overline{X}_{K-1}^*(\overline{a}_{K-2}) \right) \right\} \qquad (6.19)$$

- $(\overline{d}_{K-2}, d_{K-1}^{(1)opt}, d_K^{(1)opt})$ is also a regime in $\mathcal{D}$, so from (6.19)

$$E\{Y^*(\overline{d}_{K-2}, d_{K-1}^{(1)opt}, d_K^{(1)opt})\} = \sum_{\overline{a}_{K-2} \in \overline{\mathcal{A}}_{K-2}} E\left\{ \mathcal{J}_{K-2}^d \right.$$
$$\left. \times\ E\left(V_K^{(1)} \left[\overline{X}_K^*\{\mathcal{K}_K(d_{K-1}^{(1)opt})\}, \mathcal{K}_K(d_{K-1}^{(1)opt}) \right] \Big| \overline{X}_{K-1}^*(\overline{a}_{K-2}) \right) \right\}. \qquad (6.20)$$

Justification

- Conditional on $\overline{X}^*_{K-1}(\overline{a}_{K-2}) = \overline{x}_{K-1}$ for $h_{K-1} = (\overline{x}_{K-1}, \overline{a}_{K-2}) \in \Gamma_{K-1}$, $d_{K-1}(h_{K-1}) \in \Psi_{K-1}(h_{K-1})$. Thus, from definition of $d^{(1)opt}_{K-1}(h_{K-1})$ in (6.8) and by reasoning like that above, comparing (6.20) to (6.19) and using (6.17), yields

$$E\{Y^*(d)\} \leq E\{Y^*(\overline{d}_{K-1}, d^{(1)opt}_K)\} \leq E\{Y^*(\overline{d}_{K-2}, d^{(1)opt}_{K-1}, d^{(1)opt}_K)\} \tag{6.21}$$

- Continuing, from (6.10) with $k = K - 1$, (6.20) can be written as

$$\begin{aligned}
&E\{Y^*(\overline{d}_{K-2}, d^{(1)opt}_{K-1}, d^{(1)opt}_K)\} \\
&\quad = \sum_{\overline{a}_{K-2} \in \overline{\mathcal{A}}_{K-2}} E\left[\mathcal{J}^d_{K-2} V^{(1)}_{K-1}\{\overline{X}^*_{K-1}(\overline{a}_{K-2}), \overline{a}_{K-2}\}\right] \\
&\quad = \sum_{\overline{a}_{K-3} \in \overline{\mathcal{A}}_{K-3}} E\Big\{\mathcal{J}^d_{K-3} \\
&\quad \times E\left(V^{(1)}_{K-1}\left[\overline{X}^*_{K-1}\{\mathcal{K}_{K-1}(d_{K-2})\}, \mathcal{K}_{K-1}(d_{K-2})\right] \Big| \overline{X}^*_{K-2}(\overline{a}_{K-3})\right)\Big\}
\end{aligned} \tag{6.22}$$

Justification

- Because $(\overline{d}_{K-3}, d_{K-2}^{(1)opt}, d_{K-1}^{(1)opt}, d_K^{(1)opt}) \in \mathcal{D}$

$$E\{Y^*(\overline{d}_{K-3}, d_{K-2}^{(1)opt}, d_{K-1}^{(1)opt}, d_K^{(1)opt})\} = \sum_{\overline{a}_{K-3} \in \overline{\mathcal{A}}_{K-3}} E\left\{\mathcal{J}_{K-3}^d\right. \tag{6.23}$$

$$\left. \times E\left(V_{K-1}^{(1)}\left[\overline{X}_{K-1}^*\{\mathcal{K}_{K-1}(d_{K-2}^{(1)opt})\}, \mathcal{K}_{K-1}(d_{K-2}^{(1)opt})\right] \middle| \overline{X}_{K-2}^*(\overline{a}_{K-3})\right)\right\}$$

- From (6.9) with $k = K-2$, comparing (6.23) to (6.22) and using (6.21)

$$E\{Y^*(d)\} \leq E\{Y^*(\overline{d}_{K-1}, d_K^{(1)opt})\} \leq E\{Y^*(\overline{d}_{K-2}, d_{K-1}^{(1)opt}, d_K^{(1)opt})\}$$
$$\leq E\{Y^*(\overline{d}_{K-3}, d_{K-2}^{(1)opt}, d_{K-1}^{(1)opt}, d_K^{(1)opt})\}$$

- Continuing backward, the final step yields

$$E\{Y^*(d^{(1)opt})\} = E\left\{E\left(V_2^{(1)}\left[\overline{X}_2^*\{d_1^{(1)opt}(X_1)\}, d_1^{(1)opt}(X_1)\right] \middle| X_1\right)\right\}$$
$$= E\{V_1^{(1)}(X_1)\} \tag{6.24}$$

from (6.12)

Justification

Result: Defining for any $d \in \mathcal{D}$

$$\underline{d}_k = (d_k, d_{k+1}, \ldots, d_K), \ \ k = 1, \ldots, K$$

- Putting together yields that for any $d \in \mathcal{D}$

$$E\{Y^*(d)\} \leq \cdots \leq E\{Y^*(\overline{d}_{k-1}, \underline{d}_k^{opt(1)})\} \leq \cdots \leq E\{Y^*(d^{(1)opt})\}$$

- We have shown that $d^{(1)opt}$ satisfies (6.13) and is thus an optimal regime
- Moreover, (6.24) gives an expression for the value of an optimal regime

$$E\{Y^*(d^{(1)opt})\} = E\{V_1^{(1)}(X_1)\}$$

which we will see again next when we express an optimal regime in terms of the observed data

Characterization in terms of observed data

Practically speaking: The foregoing developments characterize an optimal regime in terms of potential outcomes

- An equivalent characterization in terms of observed data

$$(X_1, A_1, X_2, A_2, \dots, X_K, A_K, Y)$$

  is possible under SUTVA (5.10), SRA (5.11), and positivity (6.1)
- Motivates methods for estimation of an optimal regime

Demonstration: Is immediate from the equivalent representation of the sets Γ_k, $k = 1, \dots, K+1$, in terms of potential outcomes as in (5.13) and in terms of the observed data as in (5.14)

Characterization in terms of observed data

Define: For $h_K = (\overline{x}_K, \overline{a}_{K-1}) \in \Gamma_K$

$$\Gamma_K = \Big[h_K = (\overline{x}_K, \overline{a}_{K-1}) \in \overline{\mathcal{X}}_K \times \overline{\mathcal{A}}_{K-1} \text{ satisfying } (\overline{x}_{K-1}, \overline{a}_{K-1}) \in \Lambda_{K-1}$$
$$\text{and } P(X_K = x_K \mid \overline{X}_{K-1} = \overline{x}_{K-1}, \overline{A}_{K-1} = \overline{a}_{K-1}) > 0 \Big]$$

$$Q_K(h_K, a_K) = Q_K(\overline{x}_K, \overline{a}_K) = E(Y|\overline{X} = \overline{x}, \overline{A} = \overline{a})$$
$$d_K^{opt}(h_K) = d_K^{opt}(\overline{x}_K, \overline{a}_{K-1}) = \underset{a_K \in \Psi_K(h_K)}{\arg\max}\; Q_K(h_K, a_K) \tag{6.25}$$
$$V_K(h_K) = V_K(\overline{x}_K, \overline{a}_{K-1}) = \max_{a_K \in \Psi_K(h_K)} Q_K(h_K, a_K) \tag{6.26}$$

- If Y takes values in $\mathcal{Y}$ so that $(\overline{x}, \overline{a}, y) \in \Gamma_{K+1}$,

$$\Gamma_{K+1} = \Big[(\overline{x}, \overline{a}, y) \in \overline{\mathcal{X}} \times \overline{\mathcal{A}} \times \mathcal{Y} \text{ satisfying } (\overline{x}, \overline{a}) \in \Lambda_K \text{ and}$$
$$P(Y = y \mid \overline{X} = \overline{x}, \overline{A}_{K-1} = \overline{a}_{K-1}) > 0 \Big]$$

$Q_K(h_K, \overline{a}_K)$ is well defined

Characterization in terms of observed data

Define: Similarly for $k = K-1, \ldots, 2$, $h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \Gamma_k$

$$\Gamma_k = \Big[h_k = (\overline{x}_k, \overline{a}_{k-1}) \in \overline{\mathcal{X}}_k \times \overline{\mathcal{A}}_{k-1} \text{ satisfying } (\overline{x}_{k-1}, \overline{a}_{k-1}) \in \Lambda_{k-1}$$
$$\text{and } P(X_k = x_k \mid \overline{X}_{k-1} = \overline{x}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \Big]$$

$$\begin{aligned} Q_k(h_k, a_k) &= Q_k(\overline{x}_k, \overline{a}_k) \\ &= E\{V_{k+1}(\overline{x}_k, X_{k+1}, \overline{a}_k) | \overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{a}_k\} \end{aligned}$$

$$d_k^{opt}(h_k) = d_k^{opt}(\overline{x}_k, \overline{a}_{k-1}) = \arg\max_{a_k \in \Psi_k(h_k)} Q_k(h_k, a_k) \quad (6.27)$$

$$V_k(h_k) = V_k(\overline{x}_k, \overline{a}_{k-1}) = \max_{a_k \in \Psi_k(h_k)} Q_k(h_k, a_k) \quad (6.28)$$

Characterization in terms of observed data

Define: And for $h_1 = x_1 \in \Gamma_1 = \{x_1 \in \mathcal{X}_1 \text{ satisfying } P(X_1 = x_1) > 0\}$

$$Q_1(h_1, a_1) = Q_1(x_1, a_1) = E\{V_2(x_1, X_2, a_1) | X_1 = x_1, A_1 = a_1\}$$

$$d_1^{opt}(h_1) = d_1^{opt}(x_1) = \underset{a_1 \in \Psi_1(h_1)}{\arg\max}\, Q_1(h_1, a_1) \tag{6.29}$$

$$V_1(h_1) = V_1(x_1) = \max_{a_1 \in \Psi_1(h_1)} Q_1(h_1, a_1) \tag{6.30}$$

- Because $h_k \in \Gamma_k$ for $k = 1, \ldots, K$, all conditional expectations in all of these expressions are well defined
- $Q_k(h_k, a_k)$, $k = 1, \ldots, K$, are referred to as *Q-functions*, arising from usage in the literature on the reinforcement learning method known as *Q-learning* (coming up next)

Characterization in terms of observed data

Comparison:

- $d_K^{(1)opt}(h_K)$ and $V_K^{(1)}(h_K)$ in (6.6) and (6.7), defined in terms of potential outcomes, are the same as $d_K^{opt}(h_K)$ and $V_K(h_K)$ in (6.25) and (6.26), defined in terms of observed data, i.e.,

$$d_K^{(1)opt}(h_K) = d_K^{opt}(h_K), \quad V_K^{(1)}(h_K) = V_K(h_K)$$

$$\text{if} \quad E\{Y^*(\overline{a}) \mid \overline{X}_K^*(\overline{a}_{K-1}) = \overline{x}_K\} = E(Y \mid \overline{X} = \overline{x}, \overline{A} = \overline{a}) \tag{6.31}$$

- $d_k^{(1)opt}(h_k)$ and $V_k^{(1)}(h_k)$ in (6.9) and (6.10), $k = K-1, \ldots, 2$, are the same as $d_k^{opt}(h_k)$ and $V_k(h_k)$ in (6.27) and (6.28), i.e.,

$$d_k^{(1)opt}(h_k) = d_k^{opt}(h_k), \quad V_k^{(1)}(h_k) = V_k(h_k)$$

$$\begin{aligned} \text{if} \quad & E[V_{k+1}^{(1)}\{\overline{x}_k, X_{k+1}^*(\overline{a}_k), \overline{a}_k\} \mid \overline{X}_k^*(\overline{a}_{k-1}) = \overline{x}_k\} \\ & = E\{V_{k+1}(\overline{x}_k, X_{k+1}, \overline{a}_k \mid \overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{a}_k\} \end{aligned} \tag{6.32}$$

Characterization in terms of observed data

- $d_1^{(1)opt}(h_1)$ and $V_1^{(1)}(h_1)$ in (6.11) and (6.12) are the same as $d_1^{opt}(h_1)$ and $V_1(h_1)$ in (6.29) and (6.30), i.e.,

$$d_1^{(1)opt}(h_1) = d_1^{opt}(h_1), \quad V_1^{(1)}(h_k) = V_1(h_k)$$

if $E[V_2^{(1)}\{x_1, X_2^*(a_1), a_1\} \mid X_1 = x_1] = E\{V_2(x_1, X_2, a_1) \mid X_1 = x_1, \}$ (6.33)

Result: We showed in the demonstration of the equivalence of (5.13) and (5.14) that, under SUTVA, SRA, and positivity, the conditional distributions of

- $Y^*(\overline{a})$ given $\overline{X}_K^*(\overline{a}_{K-1})$ and Y given $(\overline{X}, \overline{A})$
- $X_{k+1}^*(\overline{a}_k)$ given $\overline{X}_k^*(\overline{a}_{k-1})$ and X_{k+1} given $(\overline{X}_k, \overline{A}_k)$
- $X_2^*(a_1)$ given X_1 and X_2 given X_1

are *the same*

Characterization in terms of observed data

Thus: The equalities in (6.31), (6.32), and (6.33) all hold, and it follows immediately that

$$d_k^{(1)opt}(h_k) = d_k^{opt}(h_k), \quad V_k^{(1)}(h_k) = V_k(h_k), \quad k = 1, \dots, K$$

- Confirms that an optimal regime can be expressed equivalently in terms of the observed data
- Moreover, from (6.24) and (6.33)

$$\mathcal{V}(d^{opt}) = E\{Y^*(d^{opt})\} = E\{V_1(H_1)\} = E\{V_1(X_1)\} \qquad (6.34)$$

- (6.34) is the basis for estimation of the value of an optimal regime in some approaches

6. Optimal Multiple Decision Treatment Regimes

Q-learning

Immediate: The characterization of a Ψ-specific optimal regime $d^{opt} \in \mathcal{D}$ in terms of the observed data leads to the method of *Q-learning*

- From (6.25), (6.27), and (6.29), d^{opt} can be represented in terms of the *Q-functions*

$$Q_K(h_K, a_K) = Q_K(\overline{x}_K, \overline{a}_K) = E(Y|\overline{X}_K = \overline{x}_K, \overline{A}_K = \overline{a}_K)$$

and for $K - 1, \ldots, 1$

$$Q_k(h_k, a_k) = Q_k(\overline{x}_k, \overline{a}_k) = E\{V_{k+1}(\overline{x}_k, X_{k+1}, \overline{a}_k)|\overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{a}_k\}$$

$$V_k(h_k) = V_k(\overline{x}_k, \overline{a}_{k-1}) = \max_{a_k \in \Psi_k(h_k)} Q_k(h_k, a_k), \quad k = 1, \ldots, K$$

Q-learning

Obvious approach: This suggests

- *Posit models* for the Q-functions

$$Q_k(h_k, a_k; \beta_k) = Q_k(\overline{x}_k, \overline{a}_k; \beta_k), \quad k = K, K-1, \ldots, 1$$

  depending on finite-dimensional parameters β_k, $k = 1, \ldots, K$
- Relevant models depend on the outcome (continuous, discrete)
- E.g., linear or nonlinear in β_k, can include main effects of and interactions among the elements of $\overline{x}_k$ and $\overline{a}_k$
- β_k are usually taken to be distinct/variationally independent across $k = 1, \ldots, K$
- *Linear models* are popular in practice
- Fit the models based on the observed data

$$(X_{1i}, A_{1i}, \ldots, X_{Ki}, A_{Ki}, Y_i), \quad i = 1, \ldots, n$$

  via a *backward iterative algorithm* and substitute in the definitions of $d_k^{opt}(h_k)$, $k = 1, \ldots, K$

Q-learning

At Decision K: Posit model $Q_K(h_K, a_K; \beta_K) = Q_K(\overline{x}_K, \overline{a}_{K-1}; \beta_K)$ for $E(Y|X = x, A = a)$

- Obtain $\widehat{\beta}_K$ by solving in β_K the WLS estimating equation

$$\sum_{i=1}^{n} \frac{\partial Q_K(H_{Ki}, A_{Ki}; \beta_K)}{\partial \beta_K} \Sigma_K^{-1}(H_{Ki}, A_{Ki})\{Y_i - Q_K(H_{Ki}, A_{Ki}; \beta_K)\} = 0$$

  $\Sigma_K(\overline{x}_K, \overline{a}_K)$ is a working variance model; OLS if $\Sigma_K(\overline{x}_K, \overline{a}_K) \equiv 1$
- This is a *standard* regression modeling/fitting problem
- Substitute the fitted model to obtain

$$\widehat{d}^{opt}_{Q,K}(h_K) = \underset{a_K \in \Psi_K(h_K)}{\arg\max}\ Q_K(h_K, a_K; \widehat{\beta}_K)$$

- Based on (6.26), form the *pseudo outcomes*

$$\widetilde{V}_{Ki} = \max_{a_K \in \Psi_K(H_{Ki})} Q_K(H_{Ki}, a_K; \widehat{\beta}_K)$$

Q-learning

At Decision $K-1$: Posit model $Q_{K-1}(h_{K-1}, a_{K-1}; \beta_{K-1}) = Q_{K-1}(\overline{x}_{K-1}, \overline{a}_{K-1}; \beta_{K-1})$ for

$$E\{V_K(\overline{x}_{K-1}, X_K, \overline{a}_{K-1}) | \overline{X}_{K-1} = \overline{x}_{K-1}, \overline{A}_{K-1} = \overline{a}_{K-1}\}$$

- Obtain $\widehat{\beta}_{K-1}$ by solving in β_{K-1}

$$\sum_{i=1}^{n} \frac{\partial Q_{K-1}(H_{K-1,i}, A_{K-1,i}; \beta_{K-1})}{\partial \beta_{K-1}} \Sigma_{K-1}^{-1}(H_{K-1,i}, A_{K-1,i}) \times \{\widetilde{V}_{Ki} - Q_{K-1}(H_{K-1,i}, A_{K-1,i}; \beta_{K-1})\} = 0$$

- $\Sigma_{K-1}(h_{K-1}, a_{K-1})$ is a working variance model
- This is a *nonstandard* regression problem, as the pseudo outcomes $\widetilde{V}_{Ki}$ are treated as genuine observed outcomes
- Substitute the fitted model to obtain

$$\widehat{d}^{opt}_{Q,K-1}(h_{K-1}) = \underset{a_{K-1} \in \Psi_{K-1}(h_{K-1})}{\arg\max} \; Q_{-1}(h_{K-1}, a_{K-1}; \widehat{\beta}_{K-1})$$

Q-learning

At Decisions $k = K-1, \dots, 1$: Posit a model $Q_k(h_k, a_k; \beta_k)$ $= Q_k(\overline{x}_k, \overline{a}_k; \beta_k)$ for

$$E\{V_{k+1}(\overline{x}_k, X_{k+1}, \overline{a}_k) | \overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{a}_k\}$$

- Form pseudo outcomes

$$\widetilde{V}_{k+1,i} = \max_{a_{k+1} \in \mathcal{A}_{k+1}} Q_{k+1}(H_{k+1,i}, a_{k+1}; \widehat{\beta}_{k+1})$$

- Obtain $\widehat{\beta}_k$ by solving in β_k

$$\sum_{i=1}^{n} \frac{\partial Q_k(H_{ki}, A_{ki}; \beta_k)}{\partial \beta_k} \Sigma_k^{-1}(H_{ki}, A_{ki}) \{\widetilde{V}_{k+1,i} - Q_k(H_{ki}, A_{ki}; \beta_k)\} = 0$$

- Substitute the fitted model to obtain

$$\widehat{d}^{opt}_{Q,k}(h_k) = d_k^{opt}(h_k; \widehat{\beta}_k) = \arg\max_{a_k \in \Psi_k(h_k)} Q_k(h_k, a_k; \widehat{\beta}_k)$$

Q-learning

Result: An estimated optimal Ψ-specific regime

$$\widehat{d}_Q^{opt} = \{\widehat{d}_{Q,1}^{opt}(h_1), \ldots, \widehat{d}_{Q,K}^{opt}(h_K)\}$$

and, with

$$\widetilde{V}_{1i} = \max_{a \in \Psi_1(H_{1i})} Q_1(H_{1i}, a_1; \widehat{\beta}_1), \quad i = 1, \ldots, n$$

from (6.34), an estimator for $\mathcal{V}(d^{opt})$

$$\widehat{\mathcal{V}}_Q(d^{opt}) = n^{-1} \sum_{i=1}^{n} \widetilde{V}_{1i} = n^{-1} \sum_{i=1}^{n} \max_{a \in \Psi_1(H_{1i})} Q_1(H_{1i}, a_1; \widehat{\beta}_1)$$

- As in the single decision case, $\widehat{\mathcal{V}}_Q(d^{opt})$ is a *nonregular* estimator for $\mathcal{V}(d^{opt})$

Q-learning

Estimating equations:

- We have presented conventional WLS estimating equations, $k = 1, \ldots, K$, with leading term in the summand

$$\frac{\partial Q_k(H_{ki}, A_{ki}; \beta_k)}{\partial \beta_k} \Sigma_k^{-1}(H_{ki}, A_{ki}) \tag{6.35}$$

- For Decision K, with actual observed outcomes Y_i, by standard estimating equation theory, (6.35) is the optimal leading term if

$$\text{var}(Y|\overline{X}_K = \overline{x}_K, \overline{A}_K = \overline{a}_K) = \Sigma_K(h_K, a_K)$$

  yielding efficient estimator $\widehat{\beta}_K$
- However, for $k < K$, based on pseudo outcomes, this theory may no longer apply
- Derivation of the optimal leading term would be very challenging

Q-learning

Simple demonstration: $K = 2$, $\Psi_k(h_k) = \mathcal{A}_k = \{0, 1\}$, $k = 1, 2$, continuous Y, *linear models*

$$\widetilde{h}_1 = (1, h_1^T)^T = (1, x_1^T)^T, \qquad \widetilde{h}_2 = (1, h_2^T)^T = (1, x_1^T, a_1, x_2^T)^T$$

- *Decision 2:* Model for $Q_2(h_2, a_2) = E(Y|\overline{X}_2 = \overline{x}_2, \overline{A}_2 = \overline{a}_2)$

$$Q_2(h_2, a_2; \beta_2) = \widetilde{h}_2^T \beta_{21} + a_2(\widetilde{h}_2^T \beta_{22}), \qquad \beta_2 = (\beta_{21}^T, \beta_{22}^T)^T \quad (6.36)$$

- Fit by OLS to obtain $\widehat{\beta}_2 = (\widehat{\beta}_{21}^T, \widehat{\beta}_{22}^T)^T$
- Under model (6.36)

$$V_2(h_2; \beta_2) = \max_{a_2 \in \{0,1\}} Q_2(h_2, a_2; \beta_2) = \widetilde{h}_2^T \beta_{21} + (\widetilde{h}_2^T \beta_{22}) \mathrm{I}(\widetilde{h}_2^T \beta_{22} > 0)$$

$$d_2^{opt}(h_2) = d_2^{opt}(\overline{x}_2, a_1) = \mathrm{I}(\widetilde{h}_2^T \beta_{22} > 0)$$

- Estimator

$$\widehat{d}_{Q,2}^{opt}(h_2) = d_{Q,2}^{opt}(h_2; \widehat{\beta}_2) = \mathrm{I}(\widetilde{h}_2^T \widehat{\beta}_{22} > 0)$$

Q-learning

Simple demonstration, continued: Form pseudo outcomes

$$\widetilde{V}_{2i} = \widetilde{H}_{2i}^T \widehat{\beta}_{21} + (\widetilde{H}_{2i}^T \widehat{\beta}_{22}) \mathrm{I}(\widetilde{H}_{2i}^T \widehat{\beta}_{22} > 0), \quad i = 1, \ldots, n$$

- *Decision 1:* Model for

$$Q_1(h_1, a_1) = E\{V_2(x_1, X_2, a_1) | X_1 = x_1, A_1 = a_1\}$$

$$Q_1(h_1, a_1; \beta_1) = \widetilde{h}_1^T \beta_{11} + a_1(\widetilde{h}_1^T \beta_{12}), \qquad \beta_1 = (\beta_{11}^T, \beta_{12}^T)^T \quad (6.37)$$

- Fit by OLS with "outcomes" $\widetilde{V}_{2i}$ to obtain $\widehat{\beta}_1 = (\widehat{\beta}_{11}^T, \widehat{\beta}_{12}^T)^T$
- Under model (6.37)

$$V_1(h_1; \beta_1) = \max_{a_1 \in \{0,1\}} Q_1(h_1, a_1; \beta_1) = \widetilde{h}_1^T \beta_{11} + (\widetilde{h}_1^T \beta_{12}) I(\widetilde{h}_1^T \beta_{12} > 0)$$

$$d_{Q,1}^{opt}(h_1) = d_{Q,1}^{opt}(x_1; \beta_1) = I(\widetilde{h}_1^T \beta_{12} > 0)$$

- Estimator

$$\widehat{d}_{Q,1}^{opt}(h_1) = d_{Q,1}^{opt}(x_1; \widehat{\beta}_1) = I(\widetilde{h}_1^T \widehat{\beta}_{12} > 0)$$

Q-learning

Simple demonstration, continued: Form pseudo outcomes

$$\widetilde{V}_{1i} = \widetilde{H}_{1i}^T \widehat{\beta}_{11} + (\widetilde{H}_{1i}^T \widehat{\beta}_{12}) I(\widetilde{H}_{1i}^T \widehat{\beta}_{12} > 0), \quad i = 1, \ldots, n$$

- Estimator for $\mathcal{V}(d^{opt})$ By (6.34), the value of d^{opt} can be estimated by

$$\widehat{\mathcal{V}}_Q(d^{opt}) = n^{-1} \sum_{i=1}^{n} \widetilde{V}_{1i}$$

Key issue: This simple example illustrates the potential for almost certain misspecification of the Q-function models for $k = K - 1, \ldots, 1$

- Here, with $K - 1 = 1$, the linear model (6.37) is a model for

$$Q_1(h_1, a_1) = E\{V_2(x_1, X_2, a_1) | X_1 = x_1, A_1 = a_1\}$$

$$V_2(h_2) = \max_{a_2 \in \{0,1\}} E(Y | \overline{X}_2 = \overline{x}_2, A_1 = a_1, A_2 = a_2)$$

Q-learning

Specific example: Suppose that, *in truth*

$$Q_2(h_2, a_2) = E(Y|\overline{X}_2 = \overline{x}_2, \overline{A}_2 = \overline{a}_2) = -x_1 + a_2(0.5a_1 + x_2)$$

- The linear model (6.36) is correctly specified
- Suppose further that $X_2|(X_1 = x_1, A_1 = a_1) \sim \mathcal{N}(\delta x_1, \sigma^2)$
- It can be shown that the *true* conditional expectation

$$\begin{aligned} Q_1(h_1, a_1) &= E\{V_2(x_1, X_2, a_1) \mid X_1 = x_1, A_1 = a_1\} \\ &= -x_1 + \frac{\sigma}{\sqrt{2\pi}} \exp\{-(0.5a_1 + \delta x_1)^2/(2\sigma^2)\} \\ &\quad + (0.5a_1 + \delta x_1)\Phi\{(0.5a_1 + \delta x_1)/\sigma\} \end{aligned}$$

- *Clearly:* This complex relationship is unlikely to be well approximated by the linear model (6.37)
- *Thus:* Even though the true $Q_2(h_2, a_2)$ is linear, $Q_1(h_1, a_1)$ cannot be
- Intuitively, for $K > 2$, such misspecification would propagate through all Q-function models at Decisions $K - 1, \ldots, 1$

Q-learning

Implementation: Straightforward in principle

- Fitting of the models at each step can be carried out using established methods and software
- But for $k = K - 1, \ldots, 1$ is not a standard regression exercise; almost certain model misspecification
- Can use nonparametric regression models/methods; leads to "black box" rules
- Nonetheless popular in practice
- $\widehat{d}_Q^{opt}$ can achieve reasonable performance with possibly misspecified parametric models in that $\mathcal{V}(\widehat{d}_Q^{opt})$ can approach the true value $\mathcal{V}(d^{opt})$

Q-learning

Feasible sets: ℓ_k distinct subsets $\mathcal{A}_{k,l} \subseteq \mathcal{A}_k$, $l = 1, \ldots, \ell_k$ feasible sets at Decision k; $s_k(h_k) = l$ means $\Psi_k(h_k)$ corresponds to $\mathcal{A}_{k,l}$

- ℓ_k separate models $Q_{k,l}(h_k, a_k; \beta_{kl})$, $l = 1, \ldots, \ell_k$, $k = 1, \ldots, K$

$$Q_k(h_k, a_k; \beta_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\} Q_{k,l}(h_k, a_k; \beta_{kl}), \quad \beta_k = (\beta_{k1}^T, \ldots, \beta_{k\ell_k}^T)^T$$

- Estimate β_K by $\widehat{\beta}_K$ solving

$$\sum_{i=1}^{n} \left[\sum_{l=1}^{\ell_K} \mathrm{I}\{s_K(H_{Ki}) = l\} \frac{\partial Q_{K,l}(H_{Ki}, A_{Ki}; \beta_{Kl})}{\partial \beta_{Kl}} \Sigma_{K,l}^{-1}(H_{Ki}, A_{Ki}) \right. \\ \left. \times \{Y_i - Q_{K,l}(H_{Ki}, A_{Ki}; \beta_{Kl})\} \right] = 0$$

- For h_K such that $s_K(h_K) = l$

$$\widehat{d}_{Q,K,l}^{opt}(h_K) = \arg\max_{a_K \in \mathcal{A}_{K,l}} Q_{K,l}(h_K, a_K; \widehat{\beta}_{Kl}), \quad \widehat{d}_{Q,K}^{opt}(h_K) = \sum_{l=1}^{\ell_K} \mathrm{I}\{s_K(h_K) = l\} \widehat{d}_{Q,K,l}^{opt}(h_K)$$

Q-learning

Feasible sets: For $k = K-1, \ldots, 1$

- Pseudo outcomes for i with $s_{k+1}(H_{k+1,i}) = l$

$$\widetilde{V}_{k+1,i} = \max_{a_{k+1} \in \mathcal{A}_{k+1,l}} Q_{k+1,l}(H_{k+1,i}, a_{k+1}; \widehat{\beta}_{k+1,l})$$

- Estimate β_k by $\widehat{\beta}_k$ solving

$$\sum_{i=1}^{n} \Bigg[\sum_{l=1}^{\ell_K} \mathrm{I}\{s_k(H_{ki}) = l\} \frac{\partial Q_{k,l}(H_{ki}, A_{ki}; \beta_{kl})}{\partial \beta_{kl}} \Sigma_{k,l}^{-1}(H_{ki}, A_{ki}) \\ \times \{\widetilde{V}_{k+1,i} - Q_{k,l}(H_{ki}, A_{ki}; \beta_{kl})\} \Bigg] = 0$$

- For h_k such that $s_k(h_k) = l$

$$\widehat{d}^{opt}_{Q,k,l}(h_k) = \arg\max_{a_k \in \mathcal{A}_{k,l}} Q_{k,l}(h_k, a_k; \widehat{\beta}_{kl}), \quad \widehat{d}^{opt}_{Q,k}(h_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\} \widehat{d}^{opt}_{Q,k,l}(h_k)$$

Q-learning

Remarks:

- If $\mathcal{A}_{k,l}$ contain overlapping options, could try to exploit that (although probably not worth the trouble)
- If $\mathcal{A}_{k,l}$ contains a single option, $\widehat{d}^{opt}_{Q,k,l}(h_k)$ and thus $\widehat{d}^{opt}_{Q,k}(h_k)$ must return this option for h_k such that $s_k(h_k) = l$; no model needed
- If $M_k(h_k)$ denotes number of options in $\Psi_k(h_k)$ solve

$$\sum_{i:M_k(H_{ki})>1} \left[\frac{\partial Q_k(H_{ki}, A_{ki}; \beta_k)}{\partial \beta_k} \Sigma_k^{-1}(H_{ki}, A_{ki}) \right.$$
$$\left. \times \{ \widetilde{V}_{k+1,i} - Q_k(H_{ki}, A_{ki}; \beta_k) \} \right] = 0$$

- Can show: For i with $M_k(H_{ki}) = 1$, can take $\widetilde{V}_{ki} = \widetilde{V}_{k+1,i}$ and thus "carry backward" the pseudo outcome

Restricted class of regimes

As for $K = 1$: Restrict deliberately to a class $\mathcal{D}_\eta$

- E.g., with rules involving thresholds for components of h_k
- $K = 3$, $\mathcal{A}_k = \{0, 1\}$, $k = 1, 2, 3$, $h_1 = x_1 = (x_{11}, x_{12})^T$, $x_2 = (x_{21}, x_{22})^T$, $x_3 = (x_{31}, x_{32})^T$

$$h_2 = (x_{11}, x_{12}, x_{21}, x_{22}, a_1), \quad h_3 = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32}, \overline{a}_2)$$

- Regimes in $\mathcal{D}_\eta$ have rules

$$d_1(h_1; \eta_1) = \mathrm{I}(x_{11} < \eta_{11}, x_{12} < \eta_{12}), \quad \eta_1 = (\eta_{11}, \eta_{12})^T$$
$$d_2(h_2; \eta_2) = \mathrm{I}(x_{11} < \eta_{21}, x_{21} < \eta_{22}, x_{22} < \eta_{23}), \quad \eta_2 = (\eta_{21}, \eta_{22}, \eta_{23})^T$$
$$d_3(h_3; \eta_3) = \mathrm{I}(x_{31} < \eta_{31})\mathrm{I}(a_2 = 0) + \mathrm{I}(x_{32} > \eta_{32})\mathrm{I}(a_2 = 1),$$
$$\eta_3 = (\eta_{31}, \eta_{32})^T$$

- $\mathcal{D}_\eta$ has elements

$$d_\eta = \{d_1(h_1; \eta_1), d_2(h_2; \eta_2), d_3(h_3; \eta_3)\}, \quad \eta = (\eta_1^T, \eta_2^T, \eta_3^T)^T$$

Restricted class of regimes

In general: Based on interpretability, cost, implementation

- $\mathcal{D}_\eta$ has elements

$$d_\eta = \{d_1(h_1;\eta_1),\ldots,d_K(h_K;\eta_K)\}, \quad \eta = (\eta_1^T,\ldots,\eta_K^T)^T$$

- For brevity write

$$d_{\eta,k}(h_k) = d_k(h_k;\eta_k), \quad k = 1,\ldots,K$$

- With ℓ_k distinct subsets $\mathcal{A}_{k,l} \subseteq \mathcal{A}_k$, $l = 1,\ldots,\ell_k$, as in (5.5)

$$d_k(h_k;\eta_k) = \sum_{l=1}^{\ell_k} \text{I}\{s_k(h_k) = l\}\, d_{k,l}(h_k;\eta_{kl}), \quad \eta_k = (\eta_{k1}^T,\ldots,\eta_{k\ell_k}^T)^T$$

- For brevity write

$$d_{\eta,k}(h_k) = \sum_{l=1}^{\ell_k} \text{I}\{s_k(h_k) = l\}\, d_{\eta,k,l}(h_k), \quad k = 1,\ldots,K$$

Restricted class of regimes

Optimal restricted regime d_η^{opt} in $\mathcal{D}_\eta$:

$$d_\eta^{opt} = \{d_1(h_1; \eta_1^{opt}), \ldots, d_K(h_K; \eta_K^{opt})\},$$
$$\eta^{opt} = (\eta_1^{opt\,T}, \ldots, \eta_K^{opt\,T})^T = \arg\max_\eta \mathcal{V}(d_\eta).$$

Approach: Given an estimator $\widehat{\mathcal{V}}(d)$ for the value of a fixed d

- Estimate $\mathcal{V}(d_\eta)$ by $\widehat{\mathcal{V}}(d_\eta)$ for fixed $\eta = (\eta_1^T, \ldots, \eta_K^T)^T$
- Regard $\widehat{\mathcal{V}}(d_\eta)$ as a function of η and obtain

$$\widehat{\eta}^{opt} = (\widehat{\eta}_1^{opt\,T}, \ldots, \widehat{\eta}_K^{opt\,T})^T = \arg\max_\eta \widehat{\mathcal{V}}(d_\eta)$$

  and estimate d_η^{opt} by

$$\widehat{d}_\eta^{opt} = \{d_1(h_1, \widehat{\eta}_1^{opt}), \ldots, d_K(h_K, \widehat{\eta}_K^{opt})\}$$

- *Value search* or *policy or direct search* estimation

Value search estimation

Natural choices for $\widehat{\mathcal{V}}(d_\eta)$: IPW or AIPW estimators

- As on Slides 291-297, propensities at Decision k; e.g. for the lth subset $\mathcal{A}_{k,l} \subseteq \mathcal{A}_k$, $\mathcal{A}_{k,l} = \{1,\ldots,m_{kl}\}$, $l = 1,\ldots,\ell_k$, $a_k \in \mathcal{A}_{k,l}$, $k = 1,\ldots,K$

$$\omega_{k,l}(h_k,a_k) = P(A_k = a_k | H_k = h_k), \quad \omega_{k,l}(h_k, m_{kl}) = 1 - \sum_{a_k=1}^{m_{kl}-1} \omega_{k,l}(h_k,a_k)$$

  with $\omega_{k,l}(h_k,a_k) \equiv 1$ if $m_{kl} = 1$

- Models, maximum likelihood estimators $\widehat{\gamma}_k = (\widehat{\gamma}_{k1}^T,\ldots,\widehat{\gamma}_{k\ell_k}^T)^T$

$$\omega_{k,l}(h_k,a_k;\gamma_{kl}), \quad l = 1,\ldots,\ell_k, \quad k = 1,\ldots,K$$

- Recursive representation as in (5.2), with $d_{\eta,1}(h_1) = d_{\eta,1}(x_1)$,

  $\overline{d}_{\eta,k}(\overline{x}_k) = [d_{\eta,1}(x_1), d_{\eta,2}\{\overline{x}_2, d_{\eta,1}(x_1)\},\ldots,d_{\eta,k}\{\overline{x}_k, \overline{d}_{\eta,k-1}(\overline{x}_{k-1})\}]$

  with $\overline{d}_\eta(\overline{x}) = \overline{d}_{\eta,K}(\overline{x}_K)$

Value search estimation

Natural choices for $\widehat{\mathcal{V}}(d_\eta)$: Define

$$\mathcal{C}_{d_\eta} = \mathrm{I}\{\overline{A} = \overline{d}_\eta(\overline{X})\}$$

$$\pi_{d_\eta,1}(X_1) = p_{A_1|X_1}\{d_{\eta,1}(X_1)|X_1\}$$
$$\pi_{d_\eta,k}(\overline{X}_k) = p_{A_k|\overline{X}_k,\overline{A}_{k-1}}[d_{\eta,k}\{\overline{X}_k, \overline{d}_{\eta,k-1}(\overline{X}_{k-1})\}|\overline{X}_k, \overline{d}_{\eta,k-1}(\overline{X}_{k-1})],$$
$$k = 2, \ldots, K$$

$$\pi_{d_\eta,1}(X_1;\gamma_1) = \sum_{l=1}^{\ell_1} \mathrm{I}\{s_1(h_1) = l\} \prod_{a_1=1}^{m_{1l}} \omega_{1,l}(X_1, a_1; \gamma_{1l})^{\mathrm{I}\{d_{\eta,1}(X_1)=a_1\}}$$

$$\pi_{d_\eta,k}(\overline{X}_k;\gamma_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\}$$
$$\times \prod_{a_k=1}^{m_{kl}} \omega_{k,l}\{\overline{X}_k, \overline{d}_{\eta,k-1}(\overline{X}_{k-1}), a_k; \gamma_{kl}\}^{\mathrm{I}[d_{\eta,k}\{\overline{X}_k,\overline{d}_{\eta,k-1}(\overline{X}_{k-1})\}=a_k]},$$
$$k = 2, \ldots, K$$

Value search estimation

IPW estimator: From (5.32), for fixed η and thus d_η

$$\widehat{\mathcal{V}}_{IPW}(d_\eta) = n^{-1} \sum_{i=1}^{n} \frac{\mathcal{C}_{d_\eta,i} Y_i}{\left\{\prod_{k=2}^{K} \pi_{d_\eta,k}(\overline{X}_{ki}; \widehat{\gamma}_k)\right\} \pi_{d_\eta,1}(X_{1i}; \widehat{\gamma}_1)} \qquad (6.38)$$

- Consistent for $\mathcal{V}(d_\eta)$ under SUTVA, SRA, positivity, correct propensity models
- Maximize (6.38) in η

$$\widehat{\eta}^{opt}_{IPW} = (\widehat{\eta}^{opt\,T}_{1,IPW}, \ldots, \widehat{\eta}^{opt\,T}_{K,IPW})^T$$

- Estimator for d^{opt}_η

$$\widehat{d}^{opt}_{\eta,IPW} = \{d_1(h_1, \widehat{\eta}^{opt}_{1,IPW}), \ldots, d_K(h_K, \widehat{\eta}^{opt}_{K,IPW})\}$$

- $\widehat{\mathcal{V}}_{IPW}(d^{opt}_\eta)$ by substituting $\widehat{\eta}^{opt}_{IPW}$ in $\widehat{\mathcal{V}}_{IPW}(d_\eta)$

Value search estimation

AIPW estimator: Define

$$\overline{\pi}_{d_\eta,1}(X_1;\widehat{\gamma}_1) = \pi_{d_\eta,1}(X_1;\widehat{\gamma}_1), \quad \widehat{\overline{\gamma}}_k = (\widehat{\gamma}_1^T, \ldots, \widehat{\gamma}_k^T)^T$$

$$\overline{\pi}_{d_\eta,k}(\overline{X}_k;\widehat{\overline{\gamma}}_k) = \left\{ \prod_{j=2}^{k} \pi_{d_\eta,j}(\overline{X}_j;\widehat{\gamma}_j) \right\} \pi_{d_\eta,1}(X_1;\widehat{\gamma}_1), \quad k = 2, \ldots, K$$

where $\overline{\pi}_{d_\eta,0} \equiv 1$

$$\mathcal{C}_{\overline{d}_{\eta,k}} = \mathrm{I}\{\overline{A}_k = \overline{d}_{\eta,k}(\overline{X}_k)\}, \quad k = 1, \ldots, K$$

where $\mathcal{C}_{\overline{d}_{\eta,K}} = \mathcal{C}_{d_\eta}$, $\mathcal{C}_{\overline{d}_{\eta,0}} \equiv 1$

Value search estimation

AIPW estimator: From (5.34)

$$\widehat{\mathcal{V}}_{AIPW}(d_\eta) = n^{-1} \sum_{i=1}^{n} \left[\frac{\mathcal{C}_{d_\eta,i} Y_i}{\left\{ \prod_{k=2}^{K} \pi_{d_\eta,k}(\overline{X}_{ki}; \widehat{\gamma}_k) \right\} \pi_{d_\eta,1}(X_{1i}; \widehat{\gamma}_1)} \right.$$

$$\left. + \sum_{k=1}^{K} \left\{ \frac{\mathcal{C}_{\overline{d}_{\eta,k-1},i}}{\overline{\pi}_{d_\eta,k-1}(\overline{X}_{k-1,i}; \widehat{\overline{\gamma}}_{k-1})} - \frac{\mathcal{C}_{\overline{d}_{\eta,k},i}}{\overline{\pi}_{d_\eta,k}(\overline{X}_{k,i}, \widehat{\overline{\gamma}}_k)} \right\} L_k(\overline{X}_{ki}) \right]$$

- $L_k(\overline{x}_k)$ are arbitrary functions of $\overline{x}_k$, $k = 1, \ldots, K$
- Optimal choice of $L_k(\overline{x}_k)$

$$L_k(\overline{x}_k) = E\{ Y^*(d_\eta) \mid \overline{X}_k^*(\overline{d}_{\eta,k-1}) = \overline{x}_k \}, \quad k = 1, \ldots, K$$

- Models for $E\{ Y^*(d_\eta) | \overline{X}_k^*(\overline{d}_{\eta,k-1}) = \overline{x}_k \}$ (coming up)

$$\mathcal{Q}_{d_\eta,k}(\overline{x}_k; \beta_k), \quad k = 1, \ldots, K$$

estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$

Value search estimation

AIPW estimator: Substituting these models

$$\widehat{\mathcal{V}}_{AIPW}(d_\eta) = n^{-1}\sum_{i=1}^{n}\Bigg[\frac{\mathcal{C}_{d_\eta,i}Y_i}{\left\{\prod_{k=2}^{K}\pi_{d_\eta,k}(\overline{X}_{ki};\widehat{\gamma}_k)\right\}\pi_{d_\eta,1}(X_{1i};\widehat{\gamma}_1)} \tag{6.39}$$

$$+\sum_{k=1}^{K}\left\{\frac{\mathcal{C}_{\overline{d}_{\eta,k-1},i}}{\overline{\pi}_{d_\eta,k-1}(\overline{X}_{k-1,i};\widehat{\overline{\gamma}}_{k-1})}-\frac{\mathcal{C}_{\overline{d}_{\eta,k},i}}{\overline{\pi}_{d_\eta,k}(\overline{X}_{k,i},\widehat{\overline{\gamma}}_k)}\right\}\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki};\widehat{\beta}_k)\Bigg].$$

- (6.39) is consistent for $\mathcal{V}(d_\eta)$ if either propensity models or $\mathcal{Q}_{d_\eta,k}(\overline{x}_k;\beta_k)$ are correctly specified (SUTVA, SRA, positivity)
- Doubly robust
- Maximize (6.39) in η to obtain $\widehat{\eta}^{opt}_{AIPW} = (\widehat{\eta}^{opt\,T}_{1,AIPW},\ldots,\widehat{\eta}^{opt\,T}_{K,AIPW})^T$
- Estimator for d^{opt}_η

$$\widehat{d}^{opt}_{\eta,AIPW} = \{d_1(h_1,\widehat{\eta}^{opt}_{1,AIPW}),\ldots,d_K(h_K,\widehat{\eta}^{opt}_{K,AIPW})\}$$

- $\widehat{\mathcal{V}}_{AIPW}(d^{opt}_\eta)$ by substituting $\widehat{\eta}^{opt}_{AIPW}$ in $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$

Value search estimation

Implementation of AIPW: Key challenge

- Positing and fitting models $\mathcal{Q}_{d_\eta,k}(\overline{x}_k;\beta_k)$ for

$$L_k(\overline{x}_k) = E\{Y^*(d_\eta) \mid \overline{X}_k^*(\overline{d}_{\eta,k-1}) = \overline{x}_k\}, \quad k = 1,\ldots,K$$

- One strategy: Use the *g-computation algorithm* to draw samples from (5.24)

$$p_{X_1,X_2^*(d_{\eta,1}),X_3^*(d_{\eta,2}),\ldots,X_K^*(d_{\eta,K-1})}(x_1,\ldots,x_K,y)$$
$$= p_{Y|\overline{X},\overline{A}}\{y|\overline{x},\overline{d}_{\eta,K}(\overline{x})\}\left[\prod_{k=2}^{K} p_{X_k|\overline{X}_{k-1},\overline{A}_{k-1}}\{x_k|\overline{x}_{k-1},\overline{d}_{\eta,k-1}(\overline{x}_{k-1})\}\right] p_{X_1}(x_1)$$

  for each subject i based on models for the densities on the RHS and use to estimate $L_k(\overline{X}_{ki})$ empirically to obtain $\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki};\widehat{\beta}_k)$, $i = 1,\ldots,n$, $k = 1,\ldots,K$

- Problem: This depends on d_η and thus η, so must be repeated for each i at each internal iteration of optimization of $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$

Value search estimation

Implementation of AIPW: With

$$\mu_k^{d_\eta}(\overline{x}_k) = E\{Y^*(d_\eta) \mid \overline{X}_k^*(\overline{d}_{\eta,k-1}) = \overline{x}_k\}, \quad k = 1, \ldots, K+1$$

- By standard properties of conditional expectation

$$\mu_k^{d_\eta}\{\overline{X}_k^*(\overline{d}_{\eta,k-1})\} = E\left[\mu_{k+1}^{d_\eta}\{\overline{X}_{k+1}^*(\overline{d}_{\eta,k})\}\middle| \overline{X}_k^*(\overline{d}_{\eta,k-1})\right]$$

so by SUTVA, SRA, positivity can show

$$\mu_K^d(\overline{x}_K) = E\{Y^*(d) \mid \overline{X}_K^*(\overline{d}_{K-1}) = \overline{x}_K\} = E\{Y|\overline{X}_K = \overline{x}_K, \overline{A}_K = \overline{d}_K(\overline{x}_K)\}$$

$$\mu_k^d(\overline{x}_k) = E\left[\mu_{k+1}^d\{\overline{X}_{k+1}^*(\overline{d}_k)\}\middle| \overline{X}_k^*(\overline{d}_{k-1}) = \overline{x}_k\right]$$
$$= E\{\mu_{k+1}^d(\overline{X}_{k+1}) \mid \overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{d}_k(\overline{x}_k)\}$$

- Posit models $\mu_k^{d_\eta}(\overline{x}_k;\beta_k)$; obtain $\widehat{\beta}_k$, $k = 1, \ldots, K$, by backward recursive algorithm involving only individuals with $\overline{A}_k = \overline{d}_k(\overline{X}_k)$
- Take $\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki};\widehat{\beta}_k) = \mu_k^{d_\eta}(\overline{X}_{ki};\widehat{\beta}_k)$, $i = 1, \ldots, n$, $k = 1, \ldots, K$
- Must be repeated at each internal iteration of optimization

Value search estimation

Implementation of AIPW: Q-learning-like backward algorithm

- Define $Q_K^{d_\eta}(h_K, a_K) = E(Y|H_K = h_K, A_K = a_K)$

$$Q_k^{d_\eta}(h_k, a_k) = E\{V_{k+1}^{d_\eta}(H_{k+1}) \mid H_k = h_k, A_k = a_k\}, \;\; k = 1, \ldots, K-1$$

$$V_k^{d_\eta}(h_k) = Q_k^{d_\eta}\{h_k, d_{\eta,k}(h_k)\}, \;\; k = 1, \ldots, K$$

- By SUTVA, SRA, positivity can show

$$V_k^{d_\eta}\{\overline{x}_k, \overline{d}_{\eta,k-1}(\overline{x}_{k-1})\} = E\{Y^*(d_\eta) \mid \overline{X}_k^*(\overline{d}_{\eta,k-1}) = \overline{x}_k\}$$

- Posit models $Q_k^{d_\eta}(h_k, a_k; \beta_k)$, $k = 1, \ldots, K$, by backward algorithm (next); let

$$V_k^{d_\eta}\{\overline{x}_k, \overline{d}_{\eta,k-1}(\overline{x}_{k-1}); \beta_k\} = Q_k^{d_\eta}\{\overline{x}_k, \overline{d}_{\eta,k}(\overline{x}_k); \beta_k\}$$

- Take for $i = 1, \ldots, n$, $k = 1, \ldots, K$

$$\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki}; \widehat{\beta}_k) = V_k^{d_\eta}\{\overline{X}_{ki}, \overline{d}_{\eta,k-1}(\overline{X}_{k-1,i}); \widehat{\beta}_k\} = Q_k^{d_\eta}\{\overline{X}_{ki}, \overline{d}_{\eta,k}(\overline{X}_{k,i}); \widehat{\beta}_k\}$$

Value search estimation

Implementation of AIPW: For fixed η

- Estimate β_K by $\widehat{\beta}_K$ solving M-estimating equation; e.g., OLS

$$\sum_{i=1}^{n} \frac{\partial Q_K^{d_\eta}(H_{Ki}, A_{Ki}; \beta_K)}{\partial \beta_K} \{Y_i - Q_K^{d_\eta}(H_{Ki}, A_{Ki}; \beta_K)\} = 0$$

$$\widetilde{V}_{Ki}^{d_\eta} = Q_K^{d_\eta}\{H_{Ki}, d_{\eta,K}(H_{Ki}); \widehat{\beta}_K\}, \quad i = 1, \ldots, n$$

- For $k = K-1, \ldots, 1$ estimate β_k by $\widehat{\beta}_k$ solving (e.g., OLS)

$$\sum_{i=1}^{n} \frac{\partial Q_k^{d_\eta}(H_{ki}, A_{ki}; \beta_k)}{\partial \beta_k} \{\widetilde{V}_{k+1,i}^{d_\eta} - Q_k^{d_\eta}(H_{ki}, A_{ki}; \beta_k)\} = 0$$

$$\widetilde{V}_{ki}^{d_\eta} = Q_k^{d_\eta}\{H_{ki}, d_{\eta,k}(H_{ki}); \widehat{\beta}_k\}, \quad i = 1, \ldots, n$$

- Must be repeated at each internal iteration of optimization

Value search estimation

Implementation of IPW or AIPW: Another challenge

- Dimension of η
- For given feasible sets, at Decision k, as before

$$d_k(h_k;\eta_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\}\, d_{k,l}(h_k;\eta_{kl}), \quad \eta_k = (\eta_{k1}^T, \ldots, \eta_{k\ell_k}^T)^T$$

- Thus, even if $d_{k,l}(h_k;\eta_{kl})$ have relatively simple forms, for general K and ℓ_k, $k = 1, \ldots, K$, the overall dimension of η can be high
- Making maximization of $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ or $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$, which are nonsmooth in η, extremely computationally challenging if not impossible

Value search estimation

Result: Although in principle straightforward, directly maximizing $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ and $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$ in η involves formidable obstacles

- Avoiding repeated model fitting by using $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ is inefficient
- Suggests ad hoc strategies for implementing $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$
- Zhang et al. (2013) suggest carrying out Q-learning with models $Q_k(h_k, a_k; \beta_k)$ for the Q-functions $Q_k(h_k, a_k)$, $k = 1, \ldots, K$, fitting of which does not depend on d_η or η, to obtain

$$Q_k(h_k, a_k; \widehat{\beta}_k) = Q_k(\overline{x}_k, \overline{a}_k; \widehat{\beta}_k), \quad k = 1, \ldots, K$$

- Take

$$\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki}; \widehat{\beta}_k) = Q_k\{\overline{X}_{ki}, \overline{d}_{\eta,k}(\overline{X}_{ki}); \widehat{\beta}_k), \quad k = 1, \ldots, K$$

  depend on η only through substitution of $\overline{d}_{\eta,k}(\overline{X}_{ki})$, so no need to refit models at each internal iteration of optimization
- Zhang et al. (2013) advocate this when functional forms of $d_k(h_k; \eta_k)$ are the same as those induced by Q- function models
- Hope: is "close enough"

Value search estimation

Bottom line: Implementation by direct global maximization of $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ or $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$ in η may be infeasible in practice, even with ad hoc approaches for the latter to reduce computational burden

Alternative approach: Reduce global maximization to a recursive series of lower dimensional optimizations

- First proposed for IPW by Zhao et al. (2015) and for AIPW by Zhang and Zhang (2018)

Backward iterative implementation

For simplicity: Describe first in the case of $\widehat{\mathcal{V}}_{IPW}(d_\eta)$

- The same principle applies to $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$, but details/notation are messier

Propensity models: As before, for $k = 1, \ldots, K$

$$\omega_k(h_k, a_k; \gamma_k) = \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(h_k) = l\}\, \omega_{k,l}(h_k, a_k; \gamma_{kl})$$

- Define $\underline{\gamma}_k = (\gamma_k^T, \ldots, \gamma_K^T)^T$, $k = 1, \ldots, K$, and

$$\underline{\omega}_{k,K}(h_K, a_K; \underline{\gamma}_k) = \prod_{j=k}^{K} \omega_j(h_j, a_j; \gamma_j), \quad k = 1, \ldots, K \qquad (6.40)$$

- Define estimators $\underline{\widehat{\gamma}}_k$ similarly

Backward iterative implementation

Equivalent expression: Because when $\mathcal{C}_{d_\eta} = 1$, $A_1 = d_{\eta,1}(X_1)$ and $A_k = d_{\eta,k}\{\overline{X}_k, \overline{d}_{\eta,k-1}(\overline{X}_{k-1})\}$, using (6.40)

$$n^{-1} \sum_{i=1}^{n} \frac{\mathcal{C}_{d_\eta,i} Y_i}{\left\{\prod_{k=2}^{K} \pi_{d_\eta,k}(\overline{X}_{ki}; \widehat{\gamma}_k)\right\} \pi_{d_\eta,1}(X_{1i}; \widehat{\gamma}_1)} = n^{-1} \sum_{i=1}^{n} \frac{\mathcal{C}_{d_\eta,i} Y_i}{\underline{\omega}_{1,K}(H_{Ki}, A_{Ki}; \underline{\widehat{\gamma}}_1)}$$

Define: For $k = 1, \ldots, K$

$$\underline{d}_{\eta,k} = (d_{\eta,k}, d_{\eta,k+1}, \ldots, d_{\eta,K}), \quad \text{so } \underline{d}_{\eta,1} = d_\eta$$

$$\mathfrak{C}_{d_\eta,k,K} = \mathrm{I}\{A_k = d_{\eta,k}(H_k), \ldots, A_K = d_{\eta,K}(H_K)\}, \quad \text{so } \mathcal{C}_{d_\eta} = \mathfrak{C}_{d_\eta,1,K}$$

$$\mathcal{G}_{IPW,k}(\underline{d}_{\eta,k}; \underline{\gamma}_k) = \frac{\mathfrak{C}_{d_\eta,k,K} Y}{\underline{\omega}_{k,K}(H_K, A_K; \underline{\gamma}_k)} \tag{6.41}$$

- Dependence of (6.41) on (H_K, A_K, Y) is implicit

Backward iterative implementation

Using (6.41): Can write

$$\widehat{\mathcal{V}}_{IPW}(d_\eta) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,1i}(\underline{d}_{\eta,1}; \underline{\widehat{\gamma}}_1), \tag{6.42}$$

- $\mathcal{G}_{IPW,ki}(\underline{d}_{\eta,1}; \underline{\gamma}_1)$ denotes evaluation at (H_{Ki}, A_{Ki}, Y_i), $k = 1, \ldots, K$

Backward iterative strategy: Based on *backward induction* idea

Backward iterative implementation

Decision K: Selection of treatment option is analogous to a single decision problem with "baseline" information H_K

- Define

$$\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,Ki}(d_{\eta,K}; \widehat{\gamma}_K) = n^{-1} \sum_{i=1}^{n} \frac{\mathrm{I}\{A_{Ki} = d_{\eta,K}(H_{Ki})\} Y_i}{\omega_K(H_{Ki}, A_{Ki}; \widehat{\gamma}_K)} \tag{6.43}$$

- Have used

$$\underline{\omega}_{K,K}(h_K, a_K; \underline{\gamma}_{K,K}) = \omega_K(h_K, a_K; \gamma_K), \quad \mathfrak{C}_{d_\eta,K,K} = \mathrm{I}\{A_K = d_{\eta,K}(H_K)\}$$

- (6.43) has form of an IPW estimator for a single decision problem, with Decision K and $d_{\eta,K}$ playing the roles of the single decision point and the corresponding decision rule, and H_K the role of baseline history

Backward iterative implementation

Interpretation of (6.43): Let

$$Y^*(\overline{a}_{k-1}, d_{\eta,k}, \ldots, d_{\eta,K}) = Y^*(\overline{a}_{k-1}, \underline{d}_{\eta,k}), \quad k = 1, \ldots, K$$

be the potential outcome an individual would achieve if she were to receive options $a_1, \ldots, a_{k-1}$ at Decisions 1 to $k-1$ and then be treated according to the rules $d_{\eta,k}, \ldots, d_{\eta,K}$ at Decisions k to K

- With propensity model $\omega_K(h_K, a_K; \gamma_K)$ correctly specified, can view $\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K})$ as a consistent estimator for
$$E\{Y^*(\overline{A}_{K-1}, d_{\eta,K})\} \tag{6.44}$$
- $Y^*(\overline{A}_{K-1}, d_{\eta,K})$ is the potential outcome for an individual observed to receive $\overline{A}_{K-1}$ at Decisions 1 to $K-1$
- I.e., can show $E\{\mathcal{G}_{IPW,K}(d_{\eta,K}; \gamma_{K,0}) \mid H_K\} = E\{Y^*(\overline{A}_{K-1}, d_{\eta,K}) \mid H_K\}$ so that
$$E\{\mathcal{G}_{IPW,K}(d_{\eta,K}; \gamma_{K,0})\} = E\{Y^*(\overline{A}_{K-1}, d_{\eta,K})\}$$

Backward iterative implementation

Intuitively: (6.44) is the "value" of the "single decision regime" with rule $d_{\eta,K}$, and $\overline{A}_{K-1}$ is part of the "baseline" information H_K

- $E\{Y^*(\overline{A}_{K-1}, d_{\eta,K})\}$ depends on η only through η_K
- Thus, under this analogy, define

$$d^{opt}_{\eta,K,B}(h_K) = d_K(h_K; \eta^{opt}_{K,B}), \qquad \eta^{opt}_{K,B} = \arg\max_{\eta_K} E\{Y^*(\overline{A}_{K-1}, d_{\eta,K})\}. \tag{6.45}$$

- $\eta^{opt}_{K,B}$ defined in (6.45) *need not be* the same as the component η^{opt}_K of η^{opt} globally maximizing $\mathcal{V}(d_\eta)$ in all of η
- We discuss conditions under which $\eta^{opt}_{K,B} = \eta^{opt}_K$ shortly

Backward iterative implementation

At Decision K: Maximize (6.43) $\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K})$ in η_K to obtain $\widehat{\eta}^{opt}_{K,B,IPW}$, an estimator for $\eta^{opt}_{K,B}$, and the estimator

$$\widehat{d}^{opt}_{\eta,K,B}(h_K) = d_K(h_K; \widehat{\eta}^{opt}_{K,B,IPW}) \tag{6.46}$$

- Clearly, $\widehat{\eta}^{opt}_{K,B,IPW}$ maximizing (6.43) is *not likely to be the same* as $\widehat{\eta}^{opt}_{K,IPW}$ globally maximizing $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ in all of $\eta_1, \dots, \eta_K$ (so jointly with $\widehat{\eta}^{opt}_{1,IPW}, \dots, \widehat{\eta}^{opt}_{K-1,IPW}$)
- We discuss this shortly

Backward iterative implementation

Decision $K-1$: Define

$$
\begin{aligned}
\widehat{\mathcal{V}}^{(K-1)}_{IPW}(\underline{d}_{\eta,K-1}) &= n^{-1}\sum_{i=1}^{n} \mathcal{G}_{IPW,K-1,i}(\underline{d}_{\eta,K-1};\underline{\widehat{\gamma}}_{K-1}) \\
&= n^{-1}\sum_{i=1}^{n} \mathcal{G}_{IPW,K-1,i}(d_{\eta,K-1}, d_{\eta,K};\underline{\widehat{\gamma}}_{K-1}) \qquad (6.47) \\
&= n^{-1}\sum_{i=1}^{n} \frac{\mathrm{I}\{A_{K-1,i} = d_{\eta,K-1}(H_{K-1,i}), A_{Ki} = d_{\eta,K}(H_{Ki}) Y_i}{\underline{\omega}_{K-1,K}(H_{Ki}, A_{ki};\underline{\widehat{\gamma}}_{K-1})}
\end{aligned}
$$

- (6.47) has form of a value estimator for a two decision problem
- Decisions $K-1$ and K play the roles of Decisions 1 and 2, $d_{\eta,K-1}$ and $d_{\eta,K}$ play the roles of the corresponding decision rules, and H_{K-1} plays the role of "baseline" information

Backward iterative implementation

By analogy to Decision K:

- With $\omega_{K-1}(h_{K-1}, a_{K-1}; \gamma_{K-1})$, $\omega_K(h_K, a_K; \gamma_K)$ correctly specified, can view $\widehat{\mathcal{V}}^{(K-1)}_{IPW}(\underline{d}_{\eta,K-1})$ in (6.47) as a consistent estimator for the "value"

$$E\{Y^*(\overline{A}_{K-2}, d_{\eta,K-1}, d_{\eta,K})\} \tag{6.48}$$

- Similar to (6.45), define

$$\begin{gathered} d^{opt}_{\eta,K-1,B}(h_{K-1}) = d_{K-1}(h_{K-1}; \eta^{opt}_{K-1,B}), \\ \eta^{opt}_{K-1,B} = \arg\max_{\eta_{K-1}} E\{Y^*(\overline{A}_{K-2}, d_{\eta,K-1}, d^{opt}_{\eta,K,B})\} \end{gathered} \tag{6.49}$$

- In (6.49), $d_{\eta,K}$ is fixed at $d^{opt}_{\eta,K,B}$, so $\eta^{opt}_{K-1,B}$ is *not necessarily* the global maximizer of (6.48) in $(\eta^T_{K-1}, \eta^T_K)^T$
- Nor is $\eta^{opt}_{K-1,B}$ necessarily equal to η^{opt}_{K-1} globally maximizing $\mathcal{V}(d_\eta)$

Backward iterative implementation

Thus: Can view $\widehat{\mathcal{V}}^{(K-1)}_{IPW}(d_{\eta,K-1}, d^{opt}_{\eta,K,B})$ as an estimator for

$$E\{Y^*(\overline{A}_{K-1}, d_{\eta,K-1}, d^{opt}_{\eta,K,B})\}$$

At Decision $K-1$: Maximize

$$\widehat{\mathcal{V}}^{(K-1)}_{IPW}(d_{\eta,K-1}, \widehat{d}^{opt}_{\eta,K,B})$$

in η_{K-1} to obtain $\widehat{\eta}^{opt}_{K-1,B,IPW}$

- $d_{\eta,K}$ is held fixed at $\widehat{d}^{opt}_{\eta,K,B}$ in the maximization
- Obtain

$$\widehat{d}^{opt}_{\eta,K-1,B}(h_{K-1}) = d_{K-1}(h_{K-1}; \widehat{\eta}^{opt}_{K-1,B,IPW}). \tag{6.50}$$

- $\widehat{\eta}^{opt}_{K-1,B,IPW}$ is almost certainly not the same as $\widehat{\eta}^{opt}_{K-1,IPW}$ globally maximizing $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ in all of $\eta_1, \ldots, \eta_K$

Backward iterative implementation

Continuing for $k = K-2, \ldots, 1$: At Decision k, define

$$\widehat{\mathcal{V}}^{(k)}_{IPW}(\underline{d}_{\eta,k}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,ki}(d_{\eta,k}, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k).$$

- Can be viewed as an estimator for

$$E\{Y^*(\overline{A}_{k-1}, \underline{d}_{\eta,k})\} = E\{Y^*(\overline{A}_{k-1}, d_{\eta,k}, \underline{d}_{\eta,k+1})\}. \tag{6.51}$$

- With $\underline{d}^{opt}_{\eta,k+1,B} = (d^{opt}_{\eta,k+1,B}, \ldots, d^{opt}_{\eta,K,B})$, define

$$\begin{aligned} d^{opt}_{\eta,k,B}(h_k) &= d_k(h_k; \eta^{opt}_{k,B}), \\ \eta^{opt}_{k,B} &= \arg\max_{\eta_k} E\{Y^*(\overline{A}_{k-1}, d_{\eta,k}, \underline{d}^{opt}_{\eta,k+1,B})\}, \end{aligned} \tag{6.52}$$

- With $\underline{d}^{opt}_{\eta,k+1}$ fixed at $\underline{d}^{opt}_{\eta,k+1,B}$, $\eta^{opt}_{k,B}$ is not necessarily the global maximizer of (6.51) nor equal to η^{opt}_k globally maximizing $\mathcal{V}(d_\eta)$

Backward iterative implementation

Thus: Can view $\widehat{\mathcal{V}}^{(k)}_{IPW}(d_{\eta,k}, \underline{d}^{opt}_{\eta,k+1,B})$ as an estimator for

$$E\{Y^*(\overline{A}_{k-1}, d_{\eta,k}, \underline{d}_{\eta,k+1}, B)\}$$

At Decision k: Maximize

$$\widehat{\mathcal{V}}^{(k)}_{IPW}(d_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B})$$

in η_k to obtain $\widehat{\eta}^{opt}_{k,B,IPW}$ and

$$\widehat{d}^{opt}_{\eta,k,B}(h_k) = d_k(h_k; \widehat{\eta}^{opt}_{k,B,IPW}). \tag{6.53}$$

- $\widehat{\eta}^{opt}_{k,B,IPW}$ need not be the same as $\widehat{\eta}^{opt}_{k,IPW}$ globally maximizing $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ in all of $\eta_1, \ldots, \eta_K$

Backward iterative implementation

At Decision 1: At conclusion of the algorithm

- Estimator for $d^{opt}_{\eta,B} = \{d^{opt}_{\eta,1,B}(h_1), \ldots, d^{opt}_{\eta,K,B}(h_K)\}$ is, from (6.46), (6.50), and (6.53)

$$\widehat{d}^{opt}_{\eta,B,IPW} = \{d_1(h_1; \widehat{\eta}^{opt}_{1,B,IPW}), \ldots, d_K(h_K; \widehat{\eta}^{opt}_{K,B,IPW})\} \tag{6.54}$$

- Estimator $\widehat{\mathcal{V}}_{B,IPW}(d^{opt}_{\eta,B})$ for $\mathcal{V}(d^{opt}_{\eta,B})$ is obtained by substitution of (6.54) in $\widehat{\mathcal{V}}_{IPW}(d_\eta)$

However: $\widehat{d}^{opt}_{\eta,B,IPW}$ is clearly *not the same* as the estimator $\widehat{d}^{opt}_{\eta,IPW}$ found by maximizing $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ jointly in all of $\eta = (\eta_1^T, \ldots, \eta_K^T)^T$ to obtain $\widehat{\eta}^{opt}_{IPW} = (\widehat{\eta}^{opt\,T}_{1,IPW}, \ldots, \widehat{\eta}^{opt\,T}_{K,IPW})^T$

- Thus, $\widehat{d}^{opt}_{\eta,B,IPW}$ is not necessarily a valid estimator for $d^{opt}_\eta \in \mathcal{D}_\eta$

Backward iterative implementation

Question: When is $\widehat{d}^{opt}_{\eta,B,IPW}$ a valid estimator for d^{opt}_{η}?

- Under SUTVA, SRA, positivity, can show for any $d \in \mathcal{D}$ that $d^{opt}_k, \ldots, d^{opt}_K$ maximize

$$E\{Y^*(\overline{A}_{k-1}, d_k, \ldots, d_K) \mid H_k\}, \quad k = 1, \ldots, K$$

and thus

$$\underline{d}^{opt}_k = (d^{opt}_k, \ldots, d^{opt}_K) = \underset{d_k, \ldots, d_K}{\arg\max}\, E\{Y^*(\overline{A}_{k-1}, d_k, \ldots, d_K)\}, \quad k = 1, \ldots, K$$

- Thus, if $d^{opt} \in \mathcal{D}_\eta$, $\mathcal{V}(d^{opt}_\eta) = \mathcal{V}(d^{opt})$, and d^{opt}_η and d^{opt} are *equivalent*
- Then from (6.45), (6.49), and (6.52)

$$\mathcal{V}(d^{opt}_{\eta,B}) = \mathcal{V}(d^{opt}_\eta) = \mathcal{V}(d^{opt})$$

so $d^{opt}_{\eta,B}$ is *equivalent* to d^{opt} and thus d^{opt}_η

Backward iterative implementation

Question: When is $\widehat{d}^{opt}_{\eta,B,IPW}$ a valid estimator for d^{opt}_{η}?

- Thus: $\widehat{\eta}^{opt}_{B,IPW}$ globally maximizes $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ in η, so is a valid estimator for η^{opt}
- And $\widehat{\mathcal{V}}_{B,IPW}(d^{opt}_{\eta})$ found by substituting $\widehat{\eta}^{opt}_{B,IPW}$ in $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ is a valid estimator for $\mathcal{V}(d^{opt}_{\eta})$
- *However:* When $d^{opt} \notin \mathcal{D}_\eta$, it is not necessarily the case that $\mathcal{V}(d^{opt}_{\eta,B}) = \mathcal{V}(d^{opt}_{\eta})$, so that $d^{opt}_{\eta,B}$ is not necessarily equivalent to an optimal restricted regime d^{opt}_{η}
- And thus $\widehat{\eta}^{opt}_{B,IPW}$ need not not maximize $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ in η, and $\widehat{d}^{opt}_{\eta,B,IPW}$ need not estimate $d^{opt}_{\eta} \in \mathcal{D}_\eta$
- Nonetheless, simulation evidence suggests that estimators from the backward strategy can perform well in practice

Backward iterative implementation

Sketch for AIPW: Same idea, complicated by augmentation term

- Take in $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$ in (6.39), $i = 1, \ldots, n$, $k = 1, \ldots, K$

$$\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki}; \widehat{\beta}_k) = V_k^{d_\eta}\{\overline{X}_{ki}, \overline{d}_{\eta,k-1}(\overline{X}_{k-1,i}); \widehat{\beta}_k\} = Q_k^{d_\eta}\{\overline{X}_{ki}, \overline{d}_{\eta,k}(\overline{X}_{k,i}); \widehat{\beta}_k\}$$
$$= Q_k^{d_\eta}\Big[\overline{X}_{ki}, \overline{d}_{\eta,k-1}(\overline{X}_{k-1,i}), d_{\eta,k}\{\overline{X}_{ki}, \overline{d}_{\eta,k-1}(\overline{X}_{k-1,i})\}; \widehat{\beta}_k\Big]$$

- Straightforward that $\mathcal{Q}_{d_\eta,k}(\overline{X}_{ki}; \widehat{\beta}_k)$ can be replaced by

$$V_k^{d_\eta}(\overline{X}_{ki}, \overline{A}_{k-1,i}; \widehat{\beta}_k) = Q_k^{d_\eta}\{\overline{X}_{ki}, \overline{A}_{k-1,i}, d_{\eta,k}(\overline{X}_{ki}, \overline{A}_{k-1,i}); \widehat{\beta}_k\}$$
$$= Q_k^{d_\eta}\{H_{ki}, d_{\eta,k}(H_{ki}); \widehat{\beta}_k\} = Q_k^{d_\eta}\{H_{ki}, d_k(H_{ki}; \eta_k); \widehat{\beta}_k\}$$

  for which dependence on η is only through η_k

- Thus, at each Decision $k = K, K-1, \ldots, 1$, maximization is still only in η_k

Backward iterative implementation

Alternative representation: For $k = 1, \ldots, K$, let

$$
\begin{aligned}
\mathcal{G}_{AIPW,k}(\underline{d}_{\eta,k}; \underline{\gamma}_k, \underline{\beta}_k) &= \frac{\mathfrak{C}_{d_\eta,k,K} Y}{\underline{\omega}_{k,K}(H_K, A_K; \underline{\gamma}_{k,K})} \\
&- \left[\frac{\mathrm{I}\{A_k = d_{\eta,k}(H_k)\} - \omega_k(H_k, A_k; \gamma_k)}{\omega_k(H_k, A_k; \gamma_k)} \right] Q_k^{d_\eta}\{H_k, d_{\eta,k}(H_k); \beta_k\} \\
&- \mathrm{I}(k < K) \sum_{r=k+1}^{K} \left(\frac{\mathfrak{C}_{d_\eta,k,r-1}}{\underline{\omega}_{k,r-1}(H_{r-1}, A_{r-1}; \underline{\gamma}_{k,r-1})} \right. \\
&\quad \left. \times \left[\frac{\mathrm{I}\{A_r = d_{\eta,r}(H_r)\} - \omega_r(H_r, A_r; \gamma_r)}{\omega_r(H_r, A_r; \gamma_r)} \right] Q_r^{d_\eta}\{H_r, d_{\eta,r}(H_r); \beta_r\} \right)
\end{aligned}
$$

$$
\underline{\gamma}_{\ell,r} = (\gamma_\ell^T, \ldots, \gamma_r^T)^T, \quad \underline{\omega}_{\ell,r}(h_r, a_r; \underline{\gamma}_{\ell,r}) = \prod_{j=\ell}^{r} \omega_j(h_j, a_j; \gamma_j)
$$

$$
\mathfrak{C}_{d_\eta,\ell,r} = \mathrm{I}\{A_\ell = d_{\eta,\ell}(H_\ell), \ldots, A_r = d_{\eta,r}(H_r)\}, \quad r \geq \ell = 1, \ldots, K
$$

Backward iterative implementation

Can show: $\widehat{\mathcal{V}}_{AIPW}(d_\eta) = n^{-1} \sum_{i=1}^n \mathcal{G}_{AIPW,1i}(\underline{d}_{\eta,1}; \widehat{\underline{\gamma}}_1, \widehat{\underline{\beta}}_1)$

Decision K: Maximize in η_K

$$
\begin{aligned}
\widehat{\mathcal{V}}^{(K)}_{AIPW}(d_{\eta,K}) &= n^{-1} \sum_{i=1}^n \mathcal{G}_{AIPW,Ki}(d_{\eta,K}; \widehat{\gamma}_K, \widehat{\beta}_K) \\
&= n^{-1} \sum_{i=1}^n \left(\frac{\mathrm{I}\{A_{Ki} = d_{\eta,K}(H_{Ki})\} Y_i}{\omega_K(H_{Ki}, A_{Ki}; \widehat{\gamma}_K)} \right. \\
&\quad \left. - \left[\frac{\mathrm{I}\{A_{Ki} = d_{\eta,K}(H_{Ki})\} - \omega_K(H_{Ki}, A_{Ki}; \widehat{\gamma}_{Ki})}{\omega_K(H_{Ki}, A_{Ki}; \widehat{\gamma}_{Ki})} \right] Q_K^{d_\eta}\{H_{Ki}, d_{\eta,K}(H_{Ki}); \widehat{\beta}_K\} \right)
\end{aligned}
$$

to obtain $\widehat{\eta}^{opt}_{K,B,AIPW}$ and $\widehat{d}^{opt}_{\eta,K,B}(h_K) = d_K(h_K; \widehat{\eta}^{opt}_{K,B,AIPW})$

- $Q_K^{d_\eta}(h_K, a_k; \beta_K)$ depends on η_K only through substitution of $d_{\eta,K}(h_K)$ so need not be refitted at each internal iteration

Backward iterative implementation

Decision $K-1$: Maximize $\widehat{\mathcal{V}}^{(K-1)}_{AIPW}(d_{\eta,K-1}, \widehat{d}^{opt}_{\eta,K,B})$ in η_{K-1}

$$\widehat{\mathcal{V}}^{(K-1)}_{AIPW}(\underline{d}_{\eta,K-1}) = n^{-1}\sum_{i=1}^{n} \mathcal{G}_{AIPW,K-1,i}(d_{\eta,K-1}, d_{\eta,K}; \underline{\widehat{\gamma}}_{K-1}, \underline{\widehat{\beta}}_{K-1})$$

$$\mathcal{G}_{AIPW,K-1}(\underline{d}_{\eta,K-1}; \underline{\gamma}_{K-1}, \underline{\beta}_{K-1}) = \frac{\mathfrak{C}_{d_\eta,K-1,K}\,Y}{\underline{\omega}_{K-1,K}(H_K, A_k; \underline{\gamma}_{K-1,K})}$$

$$- \left[\frac{\mathrm{I}\{A_{K-1} = d_{\eta,K-1}(H_{K-1})\} - \omega_{K-1}(H_{K-1}, A_{K-1}; \gamma_{K-1})}{\omega_{K-1}(H_{K-1}, A_{K-1}; \gamma_{K-1})}\right]$$

$$\times\, Q^{d_\eta}_{K-1}\{H_{K-1}, d_{\eta,K-1}(H_{K-1}); \beta_{K-1}\}$$

$$- \frac{\mathfrak{C}_{d_\eta,K-1,K}}{\underline{\omega}_{K-1,K}(H_K, A_K; \underline{\gamma}_{K-1,K})}\left[\frac{\mathrm{I}\{A_K = d_{\eta,K}(H_K)\} - \omega_K(H_K, A_K; \gamma_K)}{\omega_K(H_K, A_K; \gamma_K)}\right]$$

$$\times\, Q^{d_\eta}_{K}\{H_K, d_{\eta,K}(H_K); \beta_K\}$$

- Fit model $Q^{d_\eta}_{K-1}(h_{K-1}, a_{K-1}; \beta_{K-1})$ using pseudo-outcomes

$$\widetilde{V}^{d_\eta}_{Ki} = Q^{d_\eta}_{K}\{H_{Ki}, \widehat{d}^{opt}_{\eta,K,B}(H_{Ki}); \widehat{\beta}_K\}, \quad i = 1, \ldots, n$$

Backward iterative implementation

Decision k: Maximize $\widehat{\mathcal{V}}^{(k)}_{AIPW}(d^{opt}_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B})$ in η_k

$$\widehat{\mathcal{V}}^{(k)}_{AIPW}(\underline{d}_{\eta,k}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{AIPW,ki}(d_{\eta,k}, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k, \underline{\widehat{\beta}}_k)$$

- $Q^{d_\eta}_k(h_k, a_k; \beta_k)$ is a model for $E\{V^{d^{opt}_\eta}_{k+1}(H_{k+1}) \mid H_k = h_k, A_k = a_k\}$, where
$$V^{d^{opt}_\eta}_{k+1}(h_k) = Q^{d_\eta}_{k+1}\{h_{k+1}, d^{opt}_{\eta,k+1,B}(h_{k+1})\}$$
- Fit model $Q^{d_\eta}_k(h_k, a_k; \beta_k)$ using pseudo-outcomes

$$\widetilde{V}^{d_\eta}_{k+1,i} = Q^{d_\eta}_{k+1}\{H_{k+1,i}, \widehat{d}^{opt}_{\eta,k+1,B}(H_{k+1,i}); \widehat{\beta}_{k+1}\}, \quad i = 1, \ldots, n$$

At conclusion:

$$\widehat{d}^{opt}_{\eta,B,AIPW} = \{d_1(h_1; \widehat{\eta}^{opt}_{1,B,AIPW}), \ldots, d_K(h_K; \widehat{\eta}^{opt}_{K,B,AIPW})\}$$

Backward iterative implementation

Remarks:

- For either IPW or AIPW, potentially high-dimensional global maximization of $\widehat{\mathcal{V}}_{IPW}(d_\eta)$ or $\widehat{\mathcal{V}}_{AIPW}(d_\eta)$ is replaced by a series of lower dimensional maximizations in each of $\eta_K, \ldots, \eta_1$
- However, the successive maximizations of $\widehat{\mathcal{V}}^{(K)}_{AIPW}(d_{\eta,K})$ or $\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K})$ in η_K and $\widehat{\mathcal{V}}^{(k)}_{AIPW}(d^{opt}_{\eta,k}, \underline{\widehat{d}}^{opt}_{k+1,B})$ or $\widehat{\mathcal{V}}^{(k)}_{IPW}(d^{opt}_{\eta,k}, \underline{\widehat{d}}^{opt}_{k+1,B})$ in η_k, $k = K-1, \ldots, 1$, although of lower dimension, are still challenging optimization tasks
- Because these are nonsmooth objective functions

Classification analogy

Motivation: As in our review of the single decision case, optimization of nonsmooth objective functions is well-studied in the classification literature

- With two options at each Decision, $\mathcal{A}_k = \{0, 1\}$, $k = 1, \ldots, K$, feasible for all individuals, cast the optimization at each decision point as minimization of a *weighted classification error*
- Can be extended to general $\mathcal{A}_k$, with ℓ_k distinct subsets of $\mathcal{A}_k$, $\mathcal{A}_{k,l}$, $l = 1, \ldots, \ell_k$, that are feasible sets, $k = 1, \ldots, K$, where each $\mathcal{A}_{k,l}$ comprises one or two options
- We demonstrate with the IPW estimator

Classification analogy

Decision K: From (6.43), maximize in η_K

$$\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,Ki}(d_{\eta,K}; \widehat{\gamma}_K)$$

- Can be expressed equivalently as

$$\begin{aligned}
\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K}) &= n^{-1} \sum_{i=1}^{n} \big[\mathcal{G}_{IPW,Ki}(1; \widehat{\gamma}_K) \mathrm{I}\{d_{\eta,K}(H_{Ki}) = 1\} \\
&\qquad + \mathcal{G}_{IPW,Ki}(0; \widehat{\gamma}_K) \mathrm{I}\{d_{\eta,K}(H_{Ki}) = 0\}\big] \\
&= n^{-1} \sum_{i=1}^{n} \Big\{ d_K(H_{Ki}; \eta_K) \widehat{C}_{Ki} + \mathcal{G}_{IPW,Ki}(0; \widehat{\gamma}_K) \Big\}
\end{aligned}$$

$$\widehat{C}_{Ki} = \mathcal{G}_{IPW,Ki}(1; \widehat{\gamma}_K) - \mathcal{G}_{IPW,Ki}(0; \widehat{\gamma}_K) \tag{6.55}$$

Classification analogy

Thus: Maximizing $\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K})$ in η_K is equivalent to maximizing

$$n^{-1}\sum_{i=1}^{n} d_K(H_{Ki};\eta_K)\widehat{C}_{Ki}$$

- By manipulations identical to those on Slide 185, can show that maximizing $\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K})$ is equivalent to minimizing in η_K the weighted classification error

$$n^{-1}\sum_{i=1}^{n} |\widehat{C}_{Ki}|\, \mathrm{I}\Big\{\mathrm{I}(\widehat{C}_{Ki} > 0) \neq d_K(H_{Ki};\eta_K)\Big\} \qquad (6.56)$$

- Thus: Take $\mathcal{D}_\eta$ to comprise regimes whose rules are induced by a classifier (SVM, CART, etc), and use classification software to obtain $\widehat{\eta}^{opt}_{K,B,IPW}$ and $\widehat{d}^{opt}_{\eta,K,B}(h_K) = d_K(h_K;\widehat{\eta}^{opt}_{K,B,IPW})$

Classification analogy

As in the single decision case: With decision function $f_K(h_K; \eta_K)$

$$d_K(h_K; \eta_K) = \mathrm{I}\{f_K(h_K; \eta_K) > 0\}$$

can write (6.56) as

$$n^{-1} \sum_{i=1}^{n} |\widehat{C}_{Ki}|\, \ell_{0\text{-}1}\Big[\Big\{2\mathrm{I}(\widehat{C}_{Ki} > 0) - 1\Big\} f_K(H_{Ki}; \eta_K)\Big]$$

- In terms of the non convex 0-1 loss function $\ell_{0\text{-}1}(x) = I(x \leq 0)$

Classification analogy

Decisions $k = K-1, \dots, 1$: Similar argument; maximizing in η_k

$$\widehat{\mathcal{V}}^{(k)}_{IPW}(d_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,ki}(d_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B}; \underline{\widehat{\gamma}}_k)$$

is equivalent to maximizing

$$n^{-1} \sum_{i=1}^{n} d_k(H_{ki}; \eta_k) \widehat{C}_{ki}(\underline{\widehat{d}}^{opt}_{\eta,k+1,B})$$

$$\widehat{C}_{ki}(\underline{d}_{\eta,k+1}) = \mathcal{G}_{IPW,ki}(1, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k) - \mathcal{G}_{IPW,ki}(0, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k) \qquad (6.57)$$

- And by the same manipulations is equivalent to minimizing in η_k

$$n^{-1} \sum_{i=1}^{n} |\widehat{C}_{ki}(\underline{\widehat{d}}^{opt}_{\eta,k+1,B})| \, \mathrm{I}\Big[\mathrm{I}\{\widehat{C}_{ki}(\underline{\widehat{d}}^{opt}_{\eta,k+1,B}) > 0\} \neq d_k(H_{ki}; \eta_k) \Big]$$

Classification analogy

AIPW: Entirely similar formulation with

$$\widehat{C}_{Ki} = \mathcal{G}_{AIPW,Ki}(1; \widehat{\gamma}_K, \widehat{\beta}_K) - \mathcal{G}_{AIPW,Ki}(0; \widehat{\gamma}_K, \widehat{\beta}_K) \tag{6.58}$$

and for $k = K-1, \ldots, 1$

$$\begin{aligned} &\widehat{C}_{ki}(\underline{d}_{\eta,k+1}) \qquad (6.59) \\ &\quad = \mathcal{G}_{AIPW,ki}(1, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k, \underline{\widehat{\beta}}_k) - \mathcal{G}_{AIPW,ki}(0, \underline{d}_{\eta,k+1}; \underline{\widehat{\gamma}}_k, \underline{\widehat{\beta}}_k), \end{aligned}$$

Classification analogy

Predictors:

- Decision K: $\widehat{C}_{Ki}$ in (6.55) or (6.58) can be viewed as a predictor for the "contrast function"

$$C_K(H_K) = E\{Y^*(\overline{A}_{K-1}, 1) - Y^*(\overline{A}_{K-1}, 0) \mid H_K\}$$

corresponding to the difference in expected outcome for an individual with history H_K were he to receive option 1 versus option 0 at Decision K

- Decisions $k = K-1, \ldots, 1$: $\widehat{C}_{ki}(\underline{d}_{\eta,k+1})$ in (6.57) or (6.59) can be viewed as a predictor for the "contrast function"

$$C_k(H_k, \underline{d}_{\eta,k+1}) = E\{Y^*(\overline{A}_{k-1}, 1, \underline{d}_{\eta,k+1}) - Y^*(\overline{A}_{k-1}, 0, \underline{d}_{\eta,k+1}) \mid H_k\}$$

corresponding to the difference in expected outcomes for an individual with history H_k were he to receive option 1 versus option 0 at Decision k and then follow the rules $(d_{\eta,k+1}, \ldots, d_{\eta,K})$ at Decisions $k+1, \ldots, K$ thereafter

Backward outcome weighted learning (BOWL)

Extension of OWL to $K > 1$ decisions: Code $\mathcal{A}_k = \{-1, 1\}$

- Assume $\omega_k(h_k, a_k) = P(A_k = a_k \mid H_k = h_k)$ is defined for $a_k = 1, -1$ $\omega_k(h_k, a_k)$
- Proposed by Zhao et al. (2015) for known but extension to fitted models is immediate
- Again, from (6.42), wish to maximize

$$\widehat{\mathcal{V}}_{IPW}(d_\eta) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,1i}(d_\eta; \underline{\widehat{\gamma}}_1)$$

$$= n^{-1} \sum_{i=1}^{n} \left[\prod_{j=1}^{K} \frac{Y_i \text{I}\{A_{ji} = d_j(H_{ji}; \eta_j)\}}{\omega_j(H_{ji}, A_{ji}; \widehat{\gamma}_j)} \right]$$

- Represent rules in terms of decision function $f_k(h_k; \eta_k)$ as

$$d_k(h_k; \eta_k) = \text{sign}\{f_k(h_k; \eta_k)\}, \quad k = 1, \ldots, K$$

Backward outcome weighted learning (BOWL)

Decision K: Maximize in η_K

$$\widehat{\mathcal{V}}^{(K)}_{IPW}(d_{\eta,K}) = n^{-1}\sum_{i=1}^{n}\mathcal{G}_{IPW,Ki}(d_{\eta,K};\widehat{\gamma}_K) = n^{-1}\sum_{i=1}^{n}\frac{\mathrm{I}\{A_{Ki} = d_{\eta,K}(H_{Ki})\}\,Y_i}{\omega_K(H_{Ki},A_{Ki};\widehat{\gamma}_K)}$$

which is equivalent to minimizing in η_K

$$n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathrm{I}\{A_{Ki} \neq d_{\eta,K}(H_{Ki})\}}{\omega_K(H_{Ki},A_{Ki};\widehat{\gamma}_K)} = n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathrm{I}[A_{Ki} \neq \mathrm{sign}\{f_K(h_K;\eta_K)\}]}{\omega_K(H_{Ki},A_{Ki};\widehat{\gamma}_K)}.$$

- As for OWL

$$\mathrm{I}[A_{Ki} \neq \mathrm{sign}\{f_K(h_K;\eta_K)\}] = \mathrm{I}\{A_{Ki}f_K(H_{Ki};\eta_K) \leq 0\} = \ell_{0\text{-}1}\{A_{Ki}f_K(H_{Ki};\eta_K)\}$$

- Replace non convex 0-1 loss by convex surrogate hinge loss

$$\ell_{hinge}(x) = (1-x)^+, \quad x^+ = \max(0,x)$$

Backward outcome weighted learning (BOWL)

Decision K: Minimize the penalized objective

$$n^{-1} \sum_{i=1}^{n} \frac{Y_i}{\omega_K(H_{Ki}, A_{Ki}; \widehat{\gamma}_K)} \{1 - A_{Ki} f_1(H_{Ki}; \eta_K)\}^+ + \lambda_{K,n} \|f_K\|^2 \qquad (6.60)$$

- $\| \cdot \|$ is a norm for f_K, $\lambda_{K,n}$ is a scalar tuning parameter controlling complexity (penalty for overfitting)
- With $\widehat{\eta}^{opt}_{K,B,BOWL}$ the minimizer of (6.60), estimated Decision K rule

$$\widehat{d}^{opt}_{\eta,K,B}(h_K) = d_K(h_K; \widehat{\eta}^{opt}_{K,B,BOWL}) = \text{sign}\{f_K(h_K; \widehat{\eta}^{opt}_{K,B,BOWL})\}$$

Backward outcome weighted learning (BOWL)

Decisions $k = K-1, \ldots, 1$: Maximize in η_k

$$\widehat{\mathcal{V}}^{(k)}_{IPW}(d_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B}) = n^{-1} \sum_{i=1}^{n} \mathcal{G}_{IPW,ki}(d_{\eta,k}, \underline{\widehat{d}}^{opt}_{\eta,k+1,B}; \underline{\widehat{\gamma}}_k)$$

$$= n^{-1} \sum_{i=1}^{n} \frac{\prod_{j=k+1}^{K} \mathrm{I}\{A_{ji} = d_j(H_{ji}; \widehat{\eta}^{opt}_{j,B,BOWL})\} Y_i}{\prod_{j=k}^{K} \omega_j(H_{ji}, A_{ji}; \widehat{\gamma}_j)} \mathrm{I}\{A_{ki} = d_k(H_{ki}; \eta_k)\}$$

- Equivalent to minimizing in η_k

$$n^{-1} \sum_{i=1}^{n} \frac{\prod_{j=k+1}^{K} Y_i \mathrm{I}\{A_{ji} = d_j(H_{ji}; \widehat{\eta}^{opt}_{j,B,BOWL})\}}{\prod_{j=k}^{K} \omega_j(H_{ji}, A_{ji}; \widehat{\gamma}_j)} \mathrm{I}[A_{ki} \neq \mathrm{sign}\{f_k(h_k; \eta_k)\}]$$

Backward outcome weighted learning (BOWL)

Decisions $k = K-1, \ldots, 1$: Replace 0-1 loss by hinge loss and minimize in η_k the penalized objective

$$n^{-1} \sum_{i=1}^{n} \frac{\prod_{j=k+1}^{K} Y_i \mathrm{I}\{A_{ji} = d_j(H_{ji}; \widehat{\eta}^{opt}_{j,B,BOWL})\}}{\prod_{j=k}^{K} \omega_j(H_{ji}, A_{ji}; \widehat{\gamma}_j)} \{1 - A_{ki} f_k(H_{ki}; \eta_k)\}^{+} + \lambda_{k,n} \|f_k\|^2 \tag{6.61}$$

- $\lambda_{k,n}$ is a scalar tuning parameter
- With $\widehat{\eta}^{opt}_{k,B,BOWL}$ the minimizer of (6.61),estimated Decision k rule $k = K-1, \ldots, 1$, is

$$\widehat{d}^{opt}_{\eta,k,B}(h_k) = d_k(h_k; \widehat{\eta}^{opt}_{k,B,BOWL}) = \mathrm{sign}\{f_k(h_k; \widehat{\eta}^{opt}_{k,B,BOWL})\}$$

Backward outcome weighted learning (BOWL)

Remarks:

- As for OWL, replacing 0-1 loss by hinge loss means $d_k(h_k; \widehat{\eta}^{opt}_{k,B,BOWL})$ are not necessarily the same as estimated rules minimizing the original objectives
- Zhao et al. (2015) propose using a very flexible class of decision functions and thus classification method, inducing restricted class $\mathcal{D}_\eta$ with richly parameterized rules (high-dimensional η_k)
- Hope: $d^{opt} \in \mathcal{D}_\eta$
- Could take this same approach using AIPW
- Could also use simpler decision functions (e.g., SVM)
- For all IPW-based methods, the number of individuals with treatments received consistent with the first k rules decreases as k increases, so estimators can be unstable. Zhao et al. (2015) propose modifications to address this

A-learning

Consider for simplicity: $\Psi_k(h_k) = \mathcal{A}_k$ for all h_k, $k = 1, \ldots, K$

$$\mathcal{A}_k = \{0, 1, \ldots, m_k - 1\}$$

- m_k options at Decision k; option 0 is control or reference option
- Q-functions

$$Q_K(h_K, a_K) = Q_K(\overline{x}_K, \overline{a}_K) = E(Y|\overline{X}_K = \overline{x}_K, \overline{A}_K = \overline{a}_K)$$

  and for $k = K-1, \ldots, 1$

$$Q_k(h_k, a_k) = Q_k(\overline{x}_k, \overline{a}_k) = E\{V_{k+1}(\overline{x}_k, X_{k+1}, \overline{a}_k)|\overline{X}_k = \overline{x}_k, \overline{A}_k = \overline{a}_k\}$$

- Value functions

$$V_k(h_k) = \max_{j \in \{0,1,\ldots,m_k-1\}} Q_k(h_k, j) = Q_k\{h_k, d_k^{opt}(h_k)\}, \quad k = 1, \ldots, K$$

- Optimal regime comprises rules of form

$$d_k^{opt}(h_k) = d_k^{opt}(\overline{x}_k, \overline{a}_{k-1}) = \underset{j \in \{0,1,\ldots,m_k-1\}}{\arg\max}\, Q_k(h_k, j), \quad k = 1, \ldots, K$$

A-learning

Contrast functions: It suffices to know

$$C_{kj}(h_k) = Q_k(h_k, j) - Q_k(h_k, 0), \quad j = 0, \ldots, m_k - 1$$

$C_{k0}(h_k) \equiv 0$, to deduce $d_k^{opt}(h_k)$

$$d_k^{opt}(h_k) = \underset{j \in \{0,1,\ldots,m_k-1\}}{\arg\max} \; C_{kj}(h_k), \quad k = 1, \ldots, K \tag{6.62}$$

- $C_{kj}(h_k)$, $k = 1, \ldots, K$, can be regarded as *optimal blip to zero functions* as in Robins (2004), Moodie et al. (2007)
- E.g., for $K = 2$, using SUTVA and SRA

$$\begin{aligned} C_{2j}(h_2) &= E(Y \mid H_2 = h_2, A_2 = j) - E(Y \mid H_2 = h_2, A_2 = 0) \\ &= E\{Y^*(a_1, j) \mid H_2 = h_2\} - E\{Y^*(a_1, 0) \mid H_2 = h_2\} \end{aligned}$$

so is the difference in expected outcome if an individual with realized history h_2 were to receive a “blip” of treatment via option j versus control

A-learning

- Similarly, using SUTVA and SRA

$$
\begin{aligned}
Q_1(h_1, a_1) &= E\left[E\{Y \mid H_2, A_2 = d_2^{opt}(H_2)\} \Big| H_1 = h_1, A_1 = a_1 \right] \\
&= E\left\{ E\left(Y^*[a_1, d_2^{opt}\{h_1, X_2^*(a_1), a_1\}] \Big| H_1 = h_1, X_2^*(a_1), A_1 = a_1, \right.\right. \\
&\qquad\qquad \left.\left. A_2 = d_2^{opt}\{h_1, X_2^*(a_1), a_1\} \right) \Big| H_1 = h_1, A_1 = a_1 \right\} \\
&= E\left(Y^*[a_1, d_2^{opt}\{h_1, X_2^*(a_1), a_1\}] \Big| H_1 = h_1 \right)
\end{aligned}
$$

- And thus

$$
\begin{aligned}
C_{1j}(h_1) = E&\left(Y^*[j, d_2^{opt}\{h_1, X_2^*(j), j\}] \Big| H_1 = h_1 \right) \\
&- E\left(Y^*[0, d_2^{opt}\{h_1, X_2^*(0), 0\}] \Big| H_1 = h_1 \right)
\end{aligned}
$$

so is the difference in expected outcome an an individual with realized history h_1 would have if he were to receive a “blip” of treatment via option j versus control at Decision 1 and then follow an optimal regime at the final decision point

A-learning

Can write: For $k = 1, \dots, K$

$$Q_k(h_k, a_k) = \nu_k(h_k) + \sum_{j=1}^{m_k-1} \mathrm{I}(a_k = j) C_{kj}(h_k), \quad \nu_k(h_k) = Q_k(h_k, 0)$$

$$V_k(h_k) = \nu_k(h_k) + \max_{j \in \{0,1,\dots,m_k-1\}} C_{kj}(h_k)$$

Posit models:

$$C_{kj}(h_k; \psi_{kj}), \quad j = 1, \dots, m_k - 1; \;\; k = 1, \dots, K \tag{6.63}$$

- (6.63) imply models for the Q-functions

$$\nu_k(h_k) + \sum_{j=1}^{m_k-1} \mathrm{I}(a_k = j) C_{kj}(h_k; \psi_{kj}), \quad k = 1, \dots, K,$$

$$\psi_k = (\psi_{k1}^T, \dots, \psi_{k,m_k-1}^T)^T$$

A-learning

Also: Models $\omega_k(h_k, a_k; \gamma_k)$, $k = 1, \ldots, K$, for

$$\omega_k(h_k, a_k) = P(A_k = a_k | H_k = h_k), \quad k = 1, \ldots, K,$$
$$\omega_k(h_k, m_k - 1) = 1 - \sum_{a_k=0}^{m_k-2} \omega_k(h_k, a_k);$$

- Multinomial (polytomous) logistic regression models
- Fit via maximum likelihood

A-learning: Backward recursive scheme similar to Q-learning

A-learning

Decision K: Robins (1997, 2004) showed all consistent, asymptotically normal estimators for ψ_K solve

$$\sum_{i=1}^{n} \left(\left[\sum_{j=1}^{m_K-1} \lambda_{Kj}(H_{Ki}) \left\{ \mathrm{I}(A_{Ki} = j) - \omega_K(H_{Ki}, j) \right\} \right] \times \left\{ Y_i - \sum_{j=1}^{m_K-1} \mathrm{I}(A_{Ki} = j) C_{Kj}(H_{Ki}; \psi_K) + \theta_K(H_{Ki}) \right\} \right) = 0 \tag{6.64}$$

- Arbitrary vector-valued functions $\lambda_{Kj}(h_1)$, $j = 1, \ldots, m_K - 1$, of dimension of ψ_K
- Arbitrary real-valued function $\theta_K(h_K)$

A-learning

Decision K: If ψ_{Kj}, $j = 1, \dots, m_K - 1$, are nonoverlapping, (6.64) reduces to $j = 1, \dots, m_K - 1$ separate equations

$$\sum_{i=1}^{n} \Bigg[\lambda_{Kj}(H_{Ki}) \{ \mathrm{I}(A_{Ki} = j) - \omega_K(H_{Ki}, j) \} \qquad (6.65)$$
$$\times \left\{ Y_i - \sum_{j'=1}^{m_K - 1} \mathrm{I}(A_{Ki} = j') C_{Kj'}(H_{Ki}; \psi_{Kj'}) + \theta_K(H_{Ki}) \right\} \Bigg] = 0$$

- $\lambda_{Kj}(h_1)$, $j = 1, \dots, m_K - 1$, arbitrary of dimension of ψ_{Kj}
- If $C_{Kj}(h_K; \psi_{Kj})$, $j = 1, \dots, m_K - 1$, are correctly specified, optimal choice

$$\theta_K(h_K) = -\nu_K(h_K)$$

- And if $\mathrm{var}(Y \mid H_K = h_K, A_K = a_K)$ is constant, optimal $\lambda_{Kj}(h_K)$

$$\frac{\partial C_{Kj}(h_K; \psi_K)}{\partial \psi_K} \quad \text{and} \quad \frac{\partial C_{Kj}(h_K; \psi_{Kj})}{\partial \psi_{Kj}} \qquad (6.66)$$

(otherwise very complicated; use (6.66) in practice)

A-learning

Decision K: In practice, taking (6.65) as an example, posit a model $\nu_K(h_K;\phi_K)$ and estimate ψ_{Kj}, $j = 1,\dots,m_K-1$; ϕ_K; and γ_K by solving jointly

$$\sum_{i=1}^{n}\left[\frac{\partial C_{Kj}(H_{Ki};\psi_{Kj})}{\partial \psi_{Kj}}\{\mathrm{I}(A_{Ki}=j)-\omega_K(H_{Ki},j;\gamma_K)\}\right.$$
$$\left.\times\left\{Y_i-\sum_{j'=1}^{m_K-1}\mathrm{I}(A_{Ki}=j')C_{Kj'}(H_{Ki};\psi_{Kj'})-\nu_K(H_{Ki};\phi_K)\right\}\right]=0$$
$$j=1,\dots,m_K-1, \qquad (6.67)$$
$$\sum_{i=1}^{n}\left[\frac{\partial \nu_K(H_{Ki};\phi_K)}{\partial \phi_K}\left\{Y_i-\sum_{j'=1}^{m_K-1}\mathrm{I}(A_{Ki}=j')C_{Kj'}(H_{Ki};\psi_{Kj'})-\nu_K(H_{Ki};\phi_K)\right\}\right]=0$$

with the maximum likelihood score equations for γ_K

A-learning

Double robustness: As in the case $K = 1$, under SUTVA, SRA, positivity,

- If $C_{Kj}(h_K; \psi_{Kj})$, $j = 1, \ldots, m_K - 1$, are correctly specified, resulting estimator $\widehat{\psi}_K$ for ψ_K is consistent if either or both of $\omega_K(h_K, a_K; \gamma_K)$ or $\nu_K(h_K; \phi_K)$ are correctly specified
- That is, $\widehat{\psi}_K$ is doubly robust

Optimal Decision *K* rule: Estimator

$$\widehat{d}_{A,K}^{opt}(h_K) = \underset{j \in \{0,1,\ldots,m_K-1\}}{\arg\max} C_{Kj}(h_K; \widehat{\psi}_K)$$

A-learning

Similar to Slide 156: For $k = K-1, \ldots, 1$

$$E\left\{ V_{k+1}(H_{k+1}) + \max_{j \in \{0,1,\ldots,m_K-1\}} C_{kj}(H_k) - \sum_{j=1}^{m_k-1} \mathrm{I}(A_k = j) C_{kj}(H_k) \middle| H_k \right\}$$

$$= E\left[E\{V_{k+1}(H_{k+1}) \mid H_k, A_k\} + \max_{j \in \{0,1,\ldots,m_K-1\}} C_{kj}(H_k) - \right.$$

$$\left. - \sum_{j=1}^{m_k-1} \mathrm{I}(A_k = j) C_{kj}(H_k) \middle| H_k \right]$$

$$= E\left\{ Q_k(H_k, A_k) + \max_{j \in \{0,1,\ldots,m_K-1\}} C_{kj}(H_k) - \sum_{j=1}^{m_k-1} \mathrm{I}(A_k = j) C_{kj}(H_k) \middle| H_k \right\}$$

$$= E\left\{ \nu_k(H_k) + \max_{j \in \{0,1,\ldots,m_k-1\}} C_{kj}(H_k) \middle| H_k \right\} = V_k(H_k)$$

A-learning

Similar to Slide 156: For $k = K$, replacing $V_{k+1}(H_{k+1})$ by Y yields

$$E\left\{ Y + \max_{j \in \{0,1,\ldots,m_K-1\}} C_{Kj}(H_K) - \sum_{j=1}^{m_K-1} C_{Kj}(H_K) \mathrm{I}(A_K = j) \middle| H_K \right\} = V_K(H_K)$$

Suggests: Define pseudo outcomes $\widetilde{V}_{K+1,i} = Y_i$,

$$\widetilde{V}_{ki} = \widetilde{V}_{k+1,i} + \max_{j \in \{0,1,\ldots,m_k-1\}} C_{kj}(H_{ki}; \widehat{\psi}_{kj}) - \sum_{j=1}^{m_k-1} \mathrm{I}(A_{ki} = j) C_{kj}(H_{ki}; \widehat{\psi}_{kj}), \quad k = K, K-1, \ldots, 1$$

- Form estimating equations analogous to (6.67) for Decisions $k = K-1, \ldots, 1$

A-learning

Decisions $k = K-1,\ldots,1$: With ψ_{kj} nonoverlapping, obtain $\widehat{\psi}_k$, $k = K-1,\ldots,1$, by solving stacked estimating equations

$$\sum_{i=1}^{n}\left[\frac{\partial C_{kj}(H_{ki};\psi_{kj})}{\partial \psi_{kj}}\{\mathrm{I}(A_{ki}=j)-\omega_k(H_{ki},j;\gamma_k)\}\right.$$

$$\left.\times\left\{\widetilde{V}_{k+1,i}-\sum_{j'=1}^{m_k-1}\mathrm{I}(A_{ki}=j')C_{kj'}(H_{ki};\psi_{kj'})-\nu_k(H_{ki};\phi_k)\right\}\right]=0$$

$$j = 1,\ldots,m_k-1,$$

$$\sum_{i=1}^{n}\left[\frac{\partial \nu_k(H_{ki};\phi_k)}{\partial \phi_k}\left\{\widetilde{V}_{k+1,i}-\sum_{j'=1}^{m_k-1}\mathrm{I}(A_{ki}=j')C_{kj'}(H_{ki};\psi_{kj'})-\nu_k(H_{ki};\phi_k)\right\}\right]=0$$

with the maximum likelihood score equation for γ_k

- In practice, take

$$\lambda_k(h_k;\psi_k)=\frac{\partial C_k(h_k;\psi_k)}{\partial \psi_k}$$

A-learning

Optimal Decision k rule: For $k = K-1, \ldots, 1$, estimator

$$\widehat{d}^{opt}_{A,k}(h_k) = \underset{j \in \{0,1,\ldots,m_k-1\}}{\arg\max}\; C_{kj}(h_k; \widehat{\psi}_k)$$

A-learning estimator for optimal regime d^{opt}:

$$\widehat{d}^{opt}_{A} = \{\widehat{d}^{opt}_{A,1}(h_1), \ldots, \widehat{d}^{opt}_{A,K}(h_K)\}$$

A-learning

Feasible sets: With ℓ_k distinct subsets of $\mathcal{A}_k$ that are feasible sets at Decision k, $\mathcal{A}_{k,l} \subseteq A_k$, $l = 1, \ldots, \ell_k$

- Posit separate models $C_{kj,l}(h_k; \psi_{kj,l})$, $\nu_{k,l}(h_k; \phi_{kl})$, $\omega_{k,l}(h_k, a_k; \gamma_{kl})$, $l = 1, \ldots, \ell_k$
- Pseudo outcomes $\widetilde{V}_{K+1,i} = Y_i$

$$\widetilde{V}_{ki} = \widetilde{V}_{k+1,i} + \sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(H_{ki}) = l\} \Bigg[\max_{j \in \{0,1,\ldots,m_{kl}-1\}} C_{kj,l}(H_{ki}; \widehat{\psi}_{kj,l}) - \sum_{j=1}^{m_{kl}-1} \mathrm{I}(A_{ki} = j) C_{kj,l}(H_{ki}; \widehat{\psi}_{kj,l}) \Bigg], \quad k = K, K-1, \ldots, 1,$$

- With all parameters nonoverlapping, solve estimating equations on next slide to obtain

$$\widehat{d}^{opt}_{A,k,l}(h_k) = \operatorname*{arg\,max}_{j \in \{0,1,\ldots,m_{kl}-1\}} C_{kj,l}(h_k; \widehat{\psi}_{kl})$$

A-learning

$$
\sum_{i=1}^{n} \left[\sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(H_{ki}) = l\} \frac{\partial C_{kj,l}(H_{ki};\psi_{kj,l})}{\partial \psi_{kj,l}} \{\mathrm{I}(A_{ki} = j) - \omega_{k,l}(H_{ki}, j; \gamma_{kl})\} \right.
$$

$$
\left. \times \left\{ \widetilde{V}_{k+1,i} - \sum_{j'=1}^{m_{kl}-1} \mathrm{I}(A_{ki} = j') C_{kj',l}(H_{ki};\psi_{kj',l}) - \nu_{k,l}(H_{ki};\phi_{kl}) \right\} \right] = 0
$$

$$
j = 1, \dots, m_{kl} - 1,
$$

$$
\sum_{i=1}^{n} \left[\sum_{l=1}^{\ell_k} \mathrm{I}\{s_k(H_{ki}) = l\} \frac{\partial \nu_{k,l}(H_{ki};\phi_{kl})}{\partial \phi_{kl}} \right.
$$

$$
\left. \times \left\{ \widetilde{V}_{k+1,i} - \sum_{j'=1}^{m_{kl}-1} \mathrm{I}(A_{ki} = j') C_{kj',l}(H_{ki};\psi_{kj',l}) - \nu_{k,l}(H_{ki};\phi_{kl}) \right\} \right] = 0
$$

plus maximum likelihood score equations for γ_{kl}, $l = 1, \dots, \ell_k$

A-learning

Remarks:

- Murphy (2003) proposes an A-learning approach based on direct modeling of the advantage or regret function

$$\max_{j \in \{0,1,\ldots,m_k-1\}} C_{kj}(H_k) - \sum_{j=1}^{m_k-1} \mathrm{I}(A_k = j) C_{kj}(H_k)$$

  along with an alternative fitting strategy; see Moodie et al. (2007)
- Just as validity of Q-learning is predicated on correct specification of the Q-functions, validity of A-learning depends on correct specification of the contrast function models
- For Decisions $k < K$ is a nonstandard modeling problem, raising the possibility of model misspecification

Estimation via marginal structural models

Recall: From Slide 309, when scientific interest focuses on regimes with simple rules in terms of low-dimensional η

- Restrict to $\mathcal{D}_\eta \subset \mathcal{D}$ with elements

$$d_\eta = \{d_1(h_1;\eta_1),\ldots,d_K(h_K;\eta_K)\}, \quad \eta = (\eta_1^T,\ldots,\eta_K^T)^T$$

- $\mathcal{D}_\eta$ may be very simple, with

$$\eta = \eta_1 = \cdots = \eta_K$$

  as in the HIV example on Slide 310, where HIV therapy is given if CD4 count is below a common threshold η

$$d_k(h_k;\eta) = \mathrm{I}(\mathrm{CD4}_k \leq \eta), \quad k = 1,\ldots,K$$

  $\mathrm{CD4}_k$= CD4 count (cells/mm^3) immediately prior to Decision k

- *Goal:* Estimate an optimal regime in $\mathcal{D}_\eta$

$$d_\eta^{opt} = \{d_1(h_1;\eta^{opt}),\ldots,d_K(h_K;\eta^{opt})\},$$
$$\eta^{opt} = \arg\max_\eta \mathcal{V}(d_\eta)$$

Estimation via marginal structural models

Marginal structural model: A model for $\mathcal{V}(d_\eta)$ as a function of η

- I.e., $\mathcal{V}(d_\eta) = \mu(\eta)$ for some function $\mu(\,\cdot\,)$, posit a parametric model
$$\mathcal{V}(d_\eta) = E\{Y^*(d_\eta)\} = \mu(\eta;\alpha)$$
- For example, a quadratic model
$$\mu(\eta;\alpha) = \alpha_1 + \alpha_2\eta + \alpha_3\eta^2, \quad \alpha = (\alpha_1,\alpha_2,\alpha_3)^T$$
- Estimate α by $\widehat{\alpha}$ using the approaches on Slides 309–316
- Estimate η^{opt} by maximizing
$$\widehat{\mathcal{V}}_{MSM}(d_\eta) = \mu(\eta;\widehat{\alpha})$$
in η, which is entirely feasible for scalar η

Estimation via marginal structural models

Remarks:

- Estimation of an optimal regime is appealing when regimes of interest have rules characterized by a low-dimensional parameter
- Quality of estimation is clearly predicated on how well the marginal structural model represents the true relationship between $\mathcal{V}(d_\eta)$ and η
- See Robins,Orellana,and Rotnitzky (2008) and Orellana, Rotnitzky, and Robins (2010ab)

Discussion

Numerous approaches: New approaches are being developed daily

- Alternative form of value search/backward iterative strategy via nonparametric regression modeling for $d_{\eta}^{opt} \in \mathcal{D}_{\eta}$ (see Zhang et al., 2018)
- Restricted class of regimes with rules at each deision point in the form of a *decision list* (Zhang et al., 2018)
- Optimal regimes based on alternative criteria
- Optimal regimes when the outcome is a time-to-an-event/survival time, where the observed outcome may be *censored* (number and timing of decision points is *random*)

6. Optimal Multiple Decision Treatment Regimes

References

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63, 447–455.

Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussions). *Journal of the Royal Statistical Society, Series B*, 65, 331–366.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010a). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: Main content. *The International Journal of Biostatistics*, 6.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010b). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part II: Proofs and additional results. *The International Journal of Biostatistics*, 6.

References

Robins, J. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor,*Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics*, pp. 69–117, Springer-Verlag.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, New York. Springer.

Robins, J. M., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27, 4678–4721.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Statistical Science*, 29, 640–661.

References

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrics*, 100, 681–694.

Zhang, B. and Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, 74, 891–899.

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018). Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113, 1541–1549.

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110, 583–598.