Biometrika (2011), **98**, 1, *pp*. 1–20

Advance Access publication on 30 June 2011

Printed in Great Britain

Supplementary Material for Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions

BY BAQUN ZHANG

Department of Preventive Medicine, 680 N. Lakeshore Drive, Suite 1400 Northwestern University, Chicago, Illinois, 60611 U.S.A.

baqun.zhang@northwestern.edu

ANASTASIOS A.TSIATIS, ERIC B. LABER AND MARIE DAVIDIAN

Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695-8203, U.S.A.

tsiatis@ncsu.edu eblaber@ncsu.edu davidian@ncsu.edu

1. DISTRIBUTION OF POTENTIAL OUTCOMES FROM DISTRIBUTION OF OBSERVED DATA

We demonstrate how to deduce the joint distribution $p_{Y^*(\bar{a}_K),\bar{X}^*_K(\bar{a}_{K-1})}(y,\bar{x}_K)$ and conditional distributions $p_{Y^*(\bar{a}_K)|\bar{X}^*_K(\bar{a}_{K-1})}(y \mid \bar{x}_K)$ and $p_{X^*_k(\bar{a}_{k-1})|\bar{X}^*_{k-1}(\bar{a}_{k-2})}(\bar{x}_k \mid \bar{x}_{k-1})$ for a fixed $\bar{a}_K \in \bar{A}_K$ from the distribution of the observed data. Under the consistency and no unmeasured confounders assumptions, the joint density of $(W, \bar{A}_K)$ is

$$
\begin{aligned}
p_{W,\bar{A}_K}(w,\bar{a}_K) &= p_W(w)p_{\bar{A}_K|W}(\bar{a}_K \mid w) \\
&= p_W(w)p_{\bar{A}_K|\bar{A}_{K-1},W}(a_K \mid \bar{a}_{K-1}, w) \times \cdots \times p_{A_1|W}(a_1 \mid w) \\
&= p_W(w)p_{\bar{A}_K|\bar{A}_{K-1},\bar{X}_K,W}(a_K \mid \bar{a}_{K-1}, \bar{x}_K, w) \times \cdots \times p_{A_1|W}(a_1 \mid w) \\
&= p_W(w)p_{\bar{A}_K|\bar{A}_{K-1},\bar{X}_K}(a_K \mid \bar{a}_{K-1}, \bar{x}_K) \times \cdots \times p_{A_1|X_1}(a_1 \mid x_1)
\end{aligned}
$$

$$= p_W(w) p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j).$$

Moreover,

$$p_{W,\bar{A}_K|Y,\bar{X}_K,\bar{A}_K}(w, \bar{a}_K \mid y, \bar{x}_K, \bar{a}_K) = \frac{p_{W,\bar{A}_K}(w, \bar{a}_K)}{\int_{\{u: Y^*(\bar{a}_K)=y, \bar{X}_K^*(\bar{a}_{K-1})=\bar{x}_K\}} p_{W,\bar{A}_K}(u, \bar{a}_K) dv_W(u)}$$
$$= \frac{p_W(w) p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j)}{\int_{\{u: Y^*(\bar{a}_K)=y, \bar{X}_K^*(\bar{a}_{K-1})=\bar{x}_K\}} p_W(u) p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j) dv_W(u)}$$
$$= \frac{p_W(w) p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j)}{p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j) \int_{\{u: Y^*(\bar{a}_K)=y, \bar{X}_K^*(\bar{a}_{K-1})=\bar{x}_K\}} p_W(u) dv_W(u)}$$
$$= \frac{p_W(w)}{\int_{\{u: Y^*(\bar{a}_K)=y, \bar{X}_K^*(\bar{a}_{K-1})=\bar{x}_K\}} p_W(u) dv_W(u)} = p_{W|Y^*(\bar{a}_K),\bar{X}_K^*(\bar{a}_{K-1})}(w \mid y, \bar{x}_K).$$

Thus,

$$p_{Y^*(\bar{a}_K),\bar{X}_K^*(\bar{a}_{K-1})}(y, \bar{x}_K) = \frac{p_W(w)}{p_{W|Y^*(\bar{a}_K),\bar{X}_K^*(\bar{a}_{K-1})}(w \mid y, \bar{x}_K)}$$
$$= \frac{p_W(w)}{p_{W,\bar{A}_K|Y,\bar{X}_K,\bar{A}_K}(w, a_K \mid y, \bar{x}_K, a_K)}$$
$$= \frac{p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_k \mid \bar{a}_{j-1}, \bar{x}_j) p_W(w)}{p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_k \mid \bar{a}_{j-1}, \bar{x}_j) p_{W,\bar{A}_K|Y,\bar{X}_K,\bar{A}_K}(w, \bar{a}_K \mid y, \bar{x}_K, \bar{a}_K)}$$
$$= \frac{p_{W,\bar{A}_K}(w, \bar{a}_K)}{p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_k \mid \bar{a}_{j-1}, \bar{x}_j) p_{W,\bar{A}_K|Y,\bar{X}_K,\bar{A}_K}(w, \bar{a}_K \mid y, \bar{x}_K, \bar{a}_K)}$$
$$= \frac{p_{Y,\bar{X}_K,\bar{A}_K}(y, \bar{x}_K, \bar{a}_K)}{p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{K} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j)}$$
$$= p_{Y|\bar{X}_K,\bar{A}_K}(y \mid \bar{x}_K, \bar{a}_K) p_{X_1}(x_1) \prod_{j=2}^{K} p_{X_j|\bar{X}_{j-1},\bar{A}_{j-1}}(x_j \mid \bar{x}_{j-1}, \bar{a}_{j-1}).$$

Let $W_k = \{X_1, X_2^*(a_1), X_3^*(\bar{a}_2), \ldots, X_k^*(\bar{a}_{k-1}))$ for all $\bar{a}_k \in \bar{\mathcal{A}}_k\}$, $k = 2, \ldots, K$. Using the same argument, $p_{W_k,\bar{A}_k}(w, \bar{a}_k) = p_{W_k}(w_k) p_{A_1|X_1}(a_1 \mid x_1) \prod_{j=2}^{k} p_{\bar{A}_j|\bar{A}_{j-1},\bar{X}_j}(a_j \mid \bar{a}_{j-1}, \bar{x}_j)$ and $p_{\bar{X}_k^*(\bar{a}_{k-1})}(\bar{x}_k) = p_{X_1}(x_1) \prod_{j=2}^{k} p_{X_j|\bar{X}_{j-1},\bar{A}_{j-1}}(x_j \mid \bar{x}_{j-1}, \bar{a}_{j-1})$. It follows that

$$p_{Y^*(\bar{a}_K)|\bar{X}_K^*(\bar{a}_{K-1})}(y \mid \bar{x}_K) = \frac{p_{Y^*(\bar{a}_K),\bar{X}_K^*(\bar{a}_{K-1})}(y, \bar{x}_K)}{p_{\bar{X}_K^*(\bar{a}_{K-1})}(\bar{x}_K)}$$
$$= \frac{p_{Y|\bar{X}_K,\bar{A}_K}(y \mid \bar{x}_K, \bar{a}_K) p_{X_1}(x_1) \prod_{j=2}^{K} p_{X_j|\bar{X}_{j-1},\bar{A}_{j-1}}(x_j \mid \bar{x}_{j-1}, \bar{a}_{j-1})}{p_{X_1}(x_1) \prod_{j=2}^{k} p_{X_j|\bar{X}_{j-1},\bar{A}_{j-1}}(x_j \mid \bar{x}_{j-1}, \bar{a}_{j-1})}$$
$$= p_{Y|\bar{X}_K,\bar{A}_K}(y \mid \bar{x}_K, \bar{a}_K).$$

Similarly, $p_{X_k^*(\bar{a}_{k-1})|\bar{X}_{k-1}^*(\bar{a}_{k-2})}(\bar{x}_k \mid \bar{x}_{k-1}) = p_{X_k|\bar{X}_{k-1},\bar{A}_{k-1}}(x_k \mid \bar{x}_{k-1}, \bar{a}_{k-1})\,, k = 2, \ldots, K.$

2. Details of Q- and A-learning

The Q-learning procedure involves solving ordinary or weighted least squares estimating equations for each k in a backward iterative fashion. Using ordinary least squares for definiteness, solve in β_k for $k = K, \ldots, 1$

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

where $\widetilde{V}_{(K+1)i} = Y_i$, $\widetilde{V}_{(k+1)i} = \max_{a_{k+1} \in \Phi_{k+1}\{\bar{X}_{(k+1)i}, \bar{A}_{ki}\}} Q_{k+1}\{\bar{X}_{(k+1)i}, \bar{A}_{ki}, a_{k+1}; \widehat{\beta}_{k+1}\}$, $k = K-1, \ldots, 1$, and $\widetilde{V}_{1i} = \max_{a_1 \in \Phi_1(X_{1i})} Q_1(X_{1i}, a_1; \widehat{\beta}_1)$; and $\partial/\partial\beta_k\{Q_k(\bar{x}_k, \bar{a}_k; \beta_k)\}$ is the vector of partial derivatives of $Q_k(\bar{x}_k, \bar{a}_k; \beta_k)$ with respect to elements of β_k. The estimated optimal regime is $\widehat{g}_Q^{\text{opt}} = (\widehat{g}_{Q,1}^{\text{opt}}, \ldots, \widehat{g}_{Q,K}^{\text{opt}})$, where $\widehat{g}_{Q,1}^{\text{opt}}(x_1) = g_{Q,1}^{\text{opt}}(x_1; \widehat{\beta}_1) = \arg\max_{a_1 \in \Phi_1(x_1)} Q_1(x_1, a_1; \widehat{\beta}_1)$, and $\widehat{g}_{Q,k}^{\text{opt}}(\bar{x}_k, \bar{a}_{k-1}) = g_{Q,k}^{\text{opt}}(\bar{x}_k, \bar{a}_{k-1}; \widehat{\beta}_k) = \arg\max_{a_k \in \Phi_k(\bar{x}_k, \bar{a}_{k-1})} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k; \widehat{\beta}_k)$, $k = 2, \ldots, K$. An estimator for $E\{Y^*(g^{\text{opt}})\}$ is then $n^{-1}\sum_{i=1}^{n} \widetilde{V}_{1i}$. Note that Q-learning as presented here is straightforward even in the case of arbitrary feasible treatment options $\Phi_k(\bar{X}_k, \bar{A}_{k-1}) \in \mathcal{A}_k$ and is not restricted to two options at each decision k.

The A-learning procedure we consider is a version of g-estimation proposed by Robins (2004) and described in Equation (2) of Moodie et al. (2007). As noted in the main paper, $A_k C_k(\bar{x}_k, \bar{a}_{k-1})$ for decision k is equivalent to the optimal-blip-to-zero function of Robins (2004) in the case of two treatment options at each k considered here. The corresponding value function is $h_k(\bar{x}_k, \bar{a}_{k-1}) + C_k(\bar{x}_k, \bar{a}_{k-1}) I\{C_k(\bar{x}_k, \bar{a}_{k-1}) > 0\}$; $C_k(\bar{x}_k, \bar{a}_{k-1})[I\{C_k(\bar{x}_k, \bar{a}_{k-1}) > 0\} - a_k]$ is the regret of Murphy (2003). Robins (2004) and Moodie et al. (2007) discuss the relationship between regrets and optimal blip functions.

The general approach is as follows. Given posited models as described in the main paper, estimators $\widehat{\psi}_k$ for ψ_k may be found iteratively by solving simultaneously in ψ_k and α_k for $k = K, \ldots, 1$

$$\sum_{i=1}^{n} \lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)\{A_{ki} - \pi_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \widehat{\gamma}_k)\} \tag{S.1}$$

$$\times\{\widetilde{V}_{(k+1)i} - A_{ki}C_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \alpha_k)\} = 0, \tag{S.2}$$

$$\sum_{i=1}^{n} \frac{\partial h_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \alpha_k)}{\partial \alpha_k}\{\widetilde{V}_{(k+1)i} - A_{ki}C_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \alpha_k)\} = 0,$$

where $\widetilde{V}_{(K+1)i} = Y_i$, $\widetilde{V}_{ki} = \widetilde{V}_{(k+1)i} + C_k\{\bar{X}_{ki}, \bar{A}_{(k-1)i}; \widehat{\psi}_k\}\Big(I[C_k\{\bar{X}_{ki}, \bar{A}_{(k-1)i}; \widehat{\psi}_k\} > 0] - A_{ki}\Big)$, $k = K, \ldots, 2$, $\widetilde{V}_{1i} = \widetilde{V}_{2i} + C_1(X_{1i}; \widehat{\psi}_1)[I\{C_1(X_{1i}; \widehat{\psi}_1) > 0\} - A_{1i}]$. In (S.1), $\lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$ are arbitrary functions, and the entire term in (S.1) is analogous to $S_j(A_j) - E\{S_j(A_j)|\text{history}_j\}$ in Equation (2) of Moodie et al. (2007). The term in (S.2) is analogous to $H_j(\psi) - E\{H_j(\psi)|\text{history}_j\}$ in (2) of Moodie et al. (2007). This demonstrates that the approach we refer to as A-learning is equivalent to the form of g-estimation Moodie et al. (2007) cite as being refined to gain efficiency over the inefficient version in their Equation (1). See also Moodie et al. (2009).

As noted in the main paper, the $\widehat{\gamma}_k$, $k = K, \ldots, 1$, are found via solving the maximum likelihood estimating equations for binary regression for each k.

As discussed by in an unpublished article (Schulte et al., 2013), available from the last author, a reasonable choice in practice is to take $\lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$ to be equal to

$$\partial/\partial\psi_k\{C_k(\bar{x}_k, \bar{a}_{k-1}; \psi_k)\}, \tag{S.3}$$

where this expression and $\partial/\partial\alpha_k\{h_k(\bar{x}_k, \bar{a}_{k-1}; \alpha_k)$ are the obvious vectors of partial derivatives. For $k = K$, if $\text{var}(Y|\bar{X}_K, \bar{A}_{K-1})$ is constant, then the optimal choice of $\lambda_K(\bar{X}_{Ki}, \bar{A}_{(K-1)i}; \psi_K)$ is in fact (S.3). For other $k = K-1, \ldots, 1$, the form of the optimal choice of $\lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$ is very complicated, and hence the efficient estimator solving equations of form (S.1), (S.2) is virtually impossible to implement. Accordingly, taking $\lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$ to be equal to (S.3) is a feasible practical alternative. In our simulations, we adopt this choice in our implementation of A-learning in (S.1), (S.2). Indeed, in their simulations, which are based on the same set-up as our first scenario in §5 of the main paper, Moodie et al. (2007) also use this same formulation in their implementations of g-estimation.

In the event that $\text{var}(Y|\bar{X}_K, \bar{A}_{K-1})$ is not constant, the optimal choice for $k = K$ would also be complex; it is not a simple matter of incorporating weights equal to $1/\text{var}(Y|\bar{X}_K, \bar{A}_{K-1})$.

The estimated optimal regime is then $\widehat{g}_A^{\text{opt}} = (\widehat{g}_{A,1}^{\text{opt}}, \ldots, \widehat{g}_{A,K}^{\text{opt}})$, where $\widehat{g}_{A,1}^{\text{opt}}(x_1) = g_{A,1}^{\text{opt}}(x_1; \widehat{\psi}_1) = I\{C_1(x_1; \widehat{\psi}_1) > 0\}$ and $\widehat{g}_{A,k}^{\text{opt}}(\bar{x}_k, \bar{a}_{k-1}) = g_{A,k}^{\text{opt}}(\bar{x}_k, \bar{a}_{k-1}; \widehat{\psi}_k) = I\{C_k(\bar{x}_k, \bar{a}_{k-1}; \widehat{\psi}_k) > 0\}$, $k = 2, \ldots, K$, and $E\{Y^*(g^{\text{opt}})\}$ is estimated by $n^{-1}\sum_{i=1}^n \widetilde{V}_{1i}$.

If the contrast functions and propensity models are correctly specified, then it may be shown (Robins, 2004) that $\widehat{\psi}_k$ will be consistent for ψ_k even if the models $h_k(\bar{x}_k, \bar{a}_{k-1}; \alpha_k)$, $k = K, \ldots, 2$, and $h_1(x_1; \alpha_1)$ are misspecified, and $\widehat{g}_A^{\text{opt}}$ will consistently estimate g^{opt}. Thus, a simpler version of A-learning that will still lead to consistent estimation of the $\widehat{\psi}_k$ and hence the optimal regime when the contrast functions are correctly specified is to set all the h_k to be identically equal to zero. This is analogous to the inefficient version of g-estimation of Robins (2004) in (1) of Moodie et al. (2007). Moodie et al. (2007) also describe the related approach of Murphy (2003), which they refer to as iterative minimization of optimal regimes. This method is based on postulated models for the regrets $C_k(\bar{x}_k, \bar{a}_{k-1})I\{C_k(\bar{x}_k, \bar{a}_{k-1}) > 0\}$ and on taking the h_k to be identically equal to a constant for all k.

As demonstrated by Moodie et al. (2007), iterative minimization of optimal regimes and the inefficient version of g-estimation noted above in (1) of Moodie et al. (2007) yield inefficient estimators for parameters ψ_k in postulated models for the contrast functions. Accordingly, in the main paper we restrict attention to the version of A-learning in (S.1) and (S.2) here, with the $\lambda_k(\bar{X}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$ taken equal to (S.3) as described above. Given the complexity involved in implementing the fully efficient version, and given that in our simulation scenarios var$(Y|\bar{X}_K, \bar{A}_{K-1})$ is in fact constant, we are in all likelihood implementing a version of g-estimation that is as close to the efficient (impossible) version as could be hoped to be obtained in practice.

3. Results under the assumption of coarsening at random

We demonstrate $p_{W_{g\eta}|\mathcal{C}_\eta, G_{\mathcal{C}_\eta}(W_{g\eta})}(w \mid k, v) = p_{W_{g\eta}|G_k(W_{g\eta})}(w \mid v)$ and $p_{W_{g\eta}|\mathcal{C}_\eta \geq k, G_k(W_{g\eta})}(w \mid v) = p_{W_{g\eta}|G_k(W_{g\eta})}(w \mid v)$, $k = 1, \ldots, K$, when the coarsening at random assumption holds. Under this assumption, pr$(\mathcal{C}_\eta = k \mid W_{g\eta}) = \pi_{\mathcal{C}_\eta}\{k, G_k(W_{g\eta})\}$, a function of $W_{g\eta}$ only through $G_k(W_{g\eta})$, for $k = 1, \ldots, K, \infty$. Let $\nu_{W_{g\eta}}$ be the dominating measure for $W_{g\eta}$. First,

$$\begin{aligned}
p_{W_{g\eta}|\mathcal{C}_\eta, G_{\mathcal{C}_\eta}(W_{g\eta})}(w \mid k, v) &= \frac{p_{\mathcal{C}_\eta, W_{g\eta}}(k, w)}{\int_{\{u: G_k(u) = v\}} p_{\mathcal{C}_\eta, W_{g\eta}}(k, u)\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{p_{\mathcal{C}_\eta|W_{g\eta}}(k \mid w) p_{W_{g\eta}}(w)}{\int_{\{u: G_k(u) = v\}} p_{\mathcal{C}_\eta|W_{g\eta}}(k \mid u) p_{W_{g\eta}}(u)\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{\pi_{\mathcal{C}_\eta}(k, v) p_{W_{g\eta}}(w)}{\pi_{\mathcal{C}_\eta}(k, v) \int_{\{u: G_k(u) = v\}} p_{W_{g\eta}}(u)\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{p_{W_{g\eta}}(w)}{\int_{\{u: G_k(u) = v\}} p_{W_{g\eta}}(u) d\nu_{W_{g\eta}}(u)} = p_{W_{g\eta}|G_k(W_{g\eta})}(w \mid v).
\end{aligned}$$

The second result follows because

$$\begin{aligned}
p_{W_{g\eta}|\mathcal{C}_\eta \geq k, G_k(W_{g\eta})}(w \mid v) &= \frac{\int_{k' \geq k} p_{\mathcal{C}_\eta, W_{g\eta}}(k', w) d\nu_{\mathcal{C}_\eta}(k')}{\int_{\{u: G_k(u) = v\}} \{\int_{k' \geq k} p_{\mathcal{C}_\eta, W_{g\eta}}(k', u) d\nu_{\mathcal{C}_\eta}(k')\}\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{\{1 - \sum_{k'=1}^{k-1} p_{\mathcal{C}_\eta|W_{g\eta}}(k' \mid w)\} p_{W_{g\eta}}(w)}{\int_{\{u: G_k(u) = v\}} \{1 - \sum_{k'=1}^{k-1} p_{\mathcal{C}_\eta|W_{g\eta}}(k' \mid u)\} p_{W_{g\eta}}(u)\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{\{1 - \sum_{k'=1}^{k-1} \pi_{\mathcal{C}_\eta}(k', v)\} p_{W_{g\eta}}(w)}{\{1 - \sum_{k'=1}^{k-1} \pi_{\mathcal{C}_\eta}(k', v)\} \int_{\{u: G_k(u) = v\}} p_{W_{g\eta}}(u)\, d\nu_{W_{g\eta}}(u)} \\
&= \frac{p_{W_{g\eta}}(w)}{\int_{\{u: G_k(u) = v\}} p_{W_{g\eta}}(u) d\nu_{W_{g\eta}}(u)} = p_{W_{g\eta}|G_k(W_{g\eta})}(w \mid v).
\end{aligned}$$

Here, $\pi_{\mathcal{C}_\eta}(k, v)$ and $\{1 - \sum_{k'=1}^{k-1} \pi_{\mathcal{C}_\eta}(k', v)\}$ cancel in the numerator and denominator because of coarsening at random.

4. Double robustness property

We demonstrate the double robustness property for estimators of the form of (4) of the main paper with fitted models substituted. Specifically, we show that such an estimator is consistent if either the propensity score models $\pi_1(x_1; \gamma_1)$, $\pi_k(\bar{x}_k, \bar{a}_{k-1}; \gamma_k)$, $k = 2, \dots, K$, or η-dependent regression models $R_{\eta_k}(\bar{x}_k; \xi_k)$, $k = 1, \dots, K$, for $E\{Y^*(g_\eta) \mid \bar{X}_k^*(\bar{g}_{\eta_k}) = \bar{x}_k\}$, are correctly specified. Here, we use the generic notation $R_{\eta_k}(\bar{x}_k; \xi_k)$ to indicate that the models that might be substituted in (4) of the main paper for the $L_k(\bar{X}_k)$ could be correctly or incorrectly specified models $\mu_{\eta_k}(\bar{x}_k, \bar{a}_k; \xi_k)$ as in the construction of the estimator $\text{DR}(\eta)$ in (6) or could be fitted Q-functions $Q_k\{\bar{X}_{ki}, \bar{g}_{\eta_k}(\bar{X}_{ki}); \widehat{\beta}_k\}$ as in the $\text{AIPWE}(\eta)$ in (7).

Whether or not the propensity score models are correctly specified, the maximum likelihood estimator $\widehat{\gamma}$ will converge to some constant γ^*. If the models are correctly specified, then $\gamma^* = \gamma_0$, where $\pi_1(x_1; \gamma_{01}) = \pi_{01}(x_1)$ and $\pi_k(\bar{x}_k, \bar{x}_{k-1}; \gamma_{0k}) = \pi_{0k}(\bar{x}_k, \bar{x}_{k-1})$, $k = K, \dots, 2$, say, the true propensity scores. For the regression models, we also have $\widehat{\xi}$ will converge to a constant ξ^*. If the models are correctly specified, then $\xi^* = \xi_0$, where $R_{\eta_k}(\bar{x}_k; \xi_{0k}) = E\{Y^*(g_\eta) \mid \bar{X}^*(\bar{g}_{\eta_k}) = \bar{x}_k\}$, $k = 1, \dots, K$.

The estimator in (4) of the main paper converges in probability to

$$
\begin{aligned}
& E\left[\frac{I(\mathcal{C}_\eta = \infty)}{K_K(\bar{X}_K; \gamma^*)} Y + \sum_{k=1}^{K} \frac{I(\mathcal{C}_\eta = k) - \lambda_k(\bar{X}_k; \gamma_k^*) I(\mathcal{C}_\eta \geq k)}{K_k(\bar{X}_k; \gamma^*)} R_{\eta_k}(\bar{X}_k; \xi_k^*)\right] \\
&= E\left[\frac{I(\mathcal{C}_\eta = \infty)}{K_K(\bar{X}_K; \gamma^*)} Y^*(g_\eta) + \sum_{k=1}^{K} \frac{I(\mathcal{C}_\eta = k) - \lambda_k(\bar{X}_k; \gamma_k^*) I(\mathcal{C}_\eta \geq k)}{K_k(\bar{X}_k; \gamma^*)} R_{\eta_k}(\bar{X}_k; \xi_k^*)\right] \\
&= E\{Y^*(g_\eta)\} + E\left[\left\{\frac{I(\mathcal{C}_\eta = \infty)}{K_K(\bar{X}_K; \gamma^*)} - 1\right\} Y^*(g_\eta)\right]
\end{aligned}
$$

$$+E\left\{\sum_{k=1}^{K}\left[\frac{I(\mathcal{C}_\eta=k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}R_{\eta_k}(\bar{X}_k;\xi_k^*)\right]\right\}.$$

It can be shown that (Tsiatis, 2006, Chapter 10) that

$$\left\{1-\frac{I(\mathcal{C}_\eta=\infty)}{K_K(\bar{X}_K;\gamma^*)}\right\}=\sum_{k=1}^{K}\left[\frac{I(\mathcal{C}_\eta=k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right],$$

so that the estimator converges to

$$E\{Y^*(g_\eta)\}-E\left\{\sum_{k=1}^{K}\left[\frac{I(\mathcal{C}_\eta=k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right][Y^*(g_\eta)-R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\}.$$

Therefore, to demonstrate the double robustness property, it suffices to show that, for $k=1,\ldots,K$,

$$E\left\{\left[\frac{I(\mathcal{C}_\eta=k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right][Y^*(g_\eta)-R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\}=0$$

if either the propensity score models $\pi_1(X_1;\gamma_1)$, $\pi_k(\bar{X}_k,\bar{A}_{k-1};\gamma_k)$, $k=2,\ldots,K$, or the models $R_{\eta_k}(\bar{X}_k;\xi_k)$, $k=1,\ldots,K$, are correctly specified.

Consider first the case where the propensity score models are correctly specified. Then $\lambda_k(\bar{X}_k;\gamma_k^*)=\lambda_k(\bar{X}_k;\gamma_{0k})=\lambda_{0k}(\bar{X}_k)$, say, the true discrete hazards. For $k=1,\ldots,K$, define the random vector $\mathcal{F}_k=\{I(\mathcal{C}_\eta=1),\ldots,I(\mathcal{C}_\eta=k-1),W\}$. Deriving the expectation by first conditioning on $\mathcal{F}_k$,

$$\begin{aligned}&E\left\{\left[\frac{E\{I(\mathcal{C}_\eta=k)\mid\mathcal{F}_k\}-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right][Y^*(g_\eta)-R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\}\\&=E\left\{\left[\frac{\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right][Y^*(g_\eta)-R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\}=0.\end{aligned}$$

Next consider the case where the $R_{\eta_k}(\bar{X}_k;\xi_k)$, $k=1,\ldots,K$, are correctly specified. Then $R_{\eta_k}(\bar{x}_k;\xi_k^*)=R_{\eta_k}(\bar{x}_k;\xi_{0k})=E\{Y^*(g_\eta)\mid\bar{X}_k^*(\bar{g}_{\eta_k})=\bar{x}_k\}$. Using the coarsened data notation $G_k(W_{g_\eta})$ as in § 4 of the main paper,

$$E\left\{\left[\frac{I(\mathcal{C}_\eta=k)-\lambda_k(\bar{X}_k;\gamma_k^*)I(\mathcal{C}_\eta\geq k)}{K_k(\bar{X}_k;\gamma^*)}\right][Y^*(g_\eta)-R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\}$$

$$= E\left\{\left[\frac{I(\mathcal{C}_\eta = k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\}$$
$$-E\left\{\left[\frac{\lambda_k\{G_k(W_{g_\eta});\gamma_k^*\}I(\mathcal{C}_\eta \geq k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\}.$$

By first conditioning on $\{I(\mathcal{C}_\eta \geq k), G_k(W_{g_\eta})\}$, we have

$$E\left\{\left[\frac{\lambda_k\{G_k(W_{g_\eta});\gamma_k^*\}I(\mathcal{C}_\eta \geq k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\}$$
$$= E\left\{\left[\frac{\lambda_k\{G_k(W_{g_\eta});\gamma_k^*\}I(\mathcal{C}_\eta \geq k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][E\{Y^*(g_\eta) \mid \mathcal{C}_\eta \geq k, G_k(W_{g_\eta})\} - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\}.$$

In §3 of this document, it is shown that $p_{W_{g_\eta}|\mathcal{C}_\eta \geq k, G_k(W_{g_\eta})}(w \mid v) = p_{W_{g_\eta}|G_k(W_{g_\eta})}(w \mid v)$ and $Y^*(g_\eta)$ is a function of W_{g_η}, so $E\{Y^*(g_\eta) \mid \mathcal{C}_\eta \geq k, G_k(W_{g_\eta})\} = E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}$. We have

$$E\left\{\left[\frac{\lambda_k\{G_k(W_{g_\eta});\gamma_k^*\}I(\mathcal{C}_\eta \geq k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\} = 0.$$

A similar argument, conditioning on $\{I(\mathcal{C}_\eta = r), G_k(W_{g_\eta})\}$ and using $p_{W_{g_\eta}|\mathcal{C}_\eta, G_k(W_{g_\eta})}(w \mid k, v) = p_{W_{g_\eta}|G_k(W_{g_\eta})}(w \mid v)$ from §3 of this document, yields

$$E\left\{\left[\frac{I(\mathcal{C}_\eta = k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - E\{Y^*(g_\eta) \mid G_k(W_{g_\eta})\}]\right\} = 0,$$

so that

$$E\left\{\left[\frac{I(\mathcal{C}_\eta = k) - \lambda_k\{G_k(W_{g_\eta});\gamma_k^*\}I(\mathcal{C}_\eta \geq k)}{K_k\{G_k(W_{g_\eta});\gamma^*\}}\right][Y^*(g_\eta) - R_{\eta_k}(\bar{X}_k;\xi_k^*)]\right\} = 0.$$

5. Conditional expectation of potential outcomes from observed data

We wish to derive $E\{Y^*(g_\eta) \mid \bar{X}_k^*(\bar{g}_{\eta_{k-1}}) = \bar{x}_k\}$, $k = 1, \ldots, K$. First consider $k = K$ and the random variable $Y^*\{\bar{A}_{K-1}, g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\}$, and define $f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1}) = E[Y^*\{\bar{A}_{K-1}, g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\} \mid \bar{X}_K, \bar{A}_{K-1}]$. Under the consistency and no unmeasured confounders assumptions,

$$f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1}) = E[Y^*\{\bar{A}_{K-1}, g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\} \mid \bar{X}_K, \bar{A}_{K-1}]$$

$$= E[Y^*\{\bar{A}_{K-1}, g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\} \mid \bar{X}_K, \bar{A}_{K-1}, A_K = g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})]$$

$$= E\{Y \mid \bar{X}_K, \bar{A}_{K-1}, A_K = g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\} = \mu\{\bar{X}_K, \bar{A}_{K-1}, g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\}.$$

When $\mathcal{C}_\eta \geq K$, $\bar{A}_{K-1} = \bar{g}_{\eta_{K-1}}(\bar{X}_{K-1})$ and $\bar{X}^*_K(\bar{g}_{\eta_{K-1}}) = \bar{X}_K$; using $p_{W_{g_\eta}|\mathcal{C}_\eta \geq K, G_K(W_{g_\eta})}(w \mid v) = p_{W_{g_\eta}|G_K(W_{g_\eta})}(w \mid v)$ from §3 of this document, $E\{Y^*(g_\eta) \mid \bar{X}^*_K(\bar{g}_{\eta_{K-1}}) = \bar{x}_K\} = E\{Y^*(g_\eta) \mid \mathcal{C}_\eta \geq K, \bar{X}^*_K(\bar{g}_{\eta_{K-1}}) = \bar{x}_K\} = E\{Y^*(g_\eta) \mid \bar{A}_{K-1} = \bar{g}_{\eta_{K-1}}(\bar{x}_{K-1}), \bar{X}_K = \bar{x}_K\} = f_{\eta_K}\{\bar{x}_K, \bar{g}_{\eta_{K-1}}(\bar{x}_{K-1})\} = \mu\{\bar{x}_K, \bar{g}_{\eta_K}(\bar{x}_K)\}$, where $\mu_{\eta_K}(\bar{x}_K, \bar{a}_K) = E(Y \mid \bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$.

Next consider $E\{Y^*(g_\eta) \mid \bar{X}^*_{K-1}(\bar{g}_{\eta_{K-2}}) = \bar{x}_{K-1}\}$. Under no unmeasured confounders, $Y^*\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot), g_{\eta_K}(\cdot)\} \perp\!\!\!\perp A_{K-1} \mid \bar{X}_{K-1}, \bar{A}_{K-2}$, where $g_{\eta_{K-1}}(\cdot) = g_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2})$ and $g_{\eta_K}(\cdot) = g_{\eta_K}[\bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)]$; similarly, $X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2})\} \perp\!\!\!\perp A_{K-1} \mid \bar{X}_{K-1}, \bar{A}_{K-2}$. Define $f_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2}) = E(f_{\eta_K}[\bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)] \mid \bar{X}_{K-1}, \bar{A}_{K-2})$. Under the consistency and no unmeasured confounders assumptions, $f_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2}) = E\{f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1}) \mid \bar{X}_{K-1}, \bar{A}_{K-2}, A_{K-1} = g_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2})\}$. By the definition of f_{η_K},

$$f_{\eta_K}[\bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)]$$

$$= E[Y^*\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot), g_{\eta_K}(\cdot)\} \mid \bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}, A_{K-1} = g_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2})]$$

$$= E[Y^*\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot), g_{\eta_K}(\cdot)\} \mid \bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}].$$

Therefore, by the definition of $f_{\eta_{K-1}}$,

$$f_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2}) = E(f_{\eta_K}[\bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)] \mid \bar{X}_{K-1}, \bar{A}_{K-2})$$

$$= E(E[Y^*\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot), g_{\eta_K}(\cdot)\} \mid \bar{X}_{K-1}, X^*_K\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot)\}, \bar{A}_{K-2}] \mid \bar{X}_{K-1}, \bar{A}_{K-2})$$

$$= E[Y^*\{\bar{A}_{K-2}, g_{\eta_{K-1}}(\cdot), g_{\eta_K}(\cdot)\} \mid \bar{X}_{K-1}, \bar{A}_{K-2}].$$

When $\mathcal{C}_\eta \geq K-1$, $\bar{A}_{K-2} = \bar{g}_{\eta_{K-2}}(\bar{X}_{K-2})$ and $\bar{X}^*_{K-1}(\bar{g}_{\eta_{K-2}}) = \bar{X}_{K-1}$; thus, using $p_{W_{g_\eta}|\mathcal{C}_\eta \geq K-1, G_{K-1}(W_{g_\eta})}(w \mid v) = p_{W_{g_\eta}|G_{K-1}(W_{g_\eta})}(w \mid v)$ from §3 of this document, $E\{Y^*(g_\eta) \mid \bar{X}^*_{K-1}(\bar{g}_{\eta_{K-2}}) = \bar{x}_{K-1}\} = E\{Y^*(g_\eta) \mid \mathcal{C}_\eta \geq K-1, \bar{X}^*_{K-1}(\bar{g}_{\eta_{K-2}}) = \bar{x}_{K-1}\} = E\{Y^*(g_\eta) \mid \bar{A}_{K-2} = \bar{g}_{\eta_{K-2}}(\bar{x}_{K-2}), \bar{X}_{K-1} = \bar{x}_{K-1}\} = f_{\eta_{K-1}}\{\bar{x}_{K-1}, \bar{g}_{\eta_{K-2}}(\bar{x}_{K-2})\} = \mu\{\bar{x}_{K-1}, \bar{g}_{\eta_{K-1}}(\bar{x}_{K-1})\}$, where $\mu_{\eta_{K-1}}(\bar{x}_{K-1}, \bar{a}_{K-1}) = E\{f_{\eta_K}(\bar{x}_{K-1}, X_K, \bar{a}_{K-1}) \mid \bar{X}_{K-1} = \bar{x}_{K-1}, \bar{A}_{K-1} = \bar{a}_{K-1}\}$. The rest of the argument follows iteratively.

Consequently, the algorithm for building models is as follows. At decision K, specify a model $\mu_{\eta_K}(\bar{X}_K, \bar{A}_K) = E(Y \mid \bar{X}_K, \bar{A}_K)$. Define $f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1}) = \mu_{\eta_K}\{\bar{X}_K, \bar{A}_{K-1}, A_K = g_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})\}$. As $f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1})$ is a function of $\bar{X}_K$, $\bar{A}_{K-1}$, for the ith individual we can derive the corresponding predicted value. Next build a model $\mu_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-1}) = E[f_{\eta_K}(\bar{X}_K, \bar{A}_{K-1}) \mid \bar{X}_{K-1}, \bar{A}_{K-1}]$ and define $f_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2}) = \mu_{\eta_{K-1}}\{\bar{X}_{K-1}, \bar{A}_{K-2}, A_{K-1} = g_{\eta_{K-1}}(\bar{X}_{K-1}, \bar{A}_{K-2})\}$. In general, $\mu_{\eta_k}(\bar{X}_k, \bar{A}_k) = E[f_{\eta_{k+1}}(\bar{X}_{k+1}, \bar{A}_k) \mid \bar{X}_k, \bar{A}_k]$ and $f_{\eta_k}(\bar{X}_k, \bar{A}_{k-1}) = \mu_{\eta_k}\{\bar{X}_k, \bar{A}_{k-1}, A_k = g_{\eta_k}(\bar{X}_k, \bar{A}_{k-1})\}$, $k = K-1, \ldots, 2$; for $k = 1$, $\mu_{\eta_1}(X_1, A_1) = E\{f_{\eta_2}(X_1, X_2, A_1) \mid X_1, A_1\}$, $f_{\eta_1}(X_1) = \mu_{\eta_1}\{X_1, g_{\eta_1}(X_1)\}$.

6. Practical implementation when $\mathcal{G}_\eta$ is directly specified

When the class of regimes of interest $\mathcal{G}_\eta$ is directly specified without reference to models for Q-functions or Q-contrasts, the following is a practical approximate approach to circumvent the computational burden maximizing the estimator $\text{DR}(\eta)$ in (6) of the main paper, analogous to that leading to $\text{AIPWE}(\eta)$ in (7) discussed below (6). Thus, assume here that regimes $g_\eta \in \mathcal{G}_\eta$ do not necessarily correspond to regimes arising from conventional linear or nonlinear models; for example, regimes with rules of the form $g_{\eta_k}(\bar{x}_k, \bar{a}_{k-1}) = I(x_{k1} > \eta_{k1}, x_{k2} > \eta_{k2})$, $x_k = (x_{k1}, x_{k2})^{\mathrm{T}}$, $\eta_k = (\eta_{k1}, \eta_{k2})^{\mathrm{T}}$. Here, the strategy in the main paper of substituting fitted

Q-functions (that is, maximizing AIPWE(η) instead) is not appropriate, as these would almost certainly not be compatible with the form of the regimes.

Instead, to avoid the computationally intensive refitting of the models $\mu_{\eta_k}(\bar{x}_k, \bar{a}_{k-1}; \xi_k)$ for each η encountered in the maximization of DR(η), we suggest the following approach.

(i) Obtain a preliminary estimator for η, $\widehat{\eta}^{(0)}$, say, by maximizing IPWE(η) in (5) or AIPWE(η) in (7) (for some specification of Q-functions) of the main paper. Let $\ell = 0$.

(ii) Fix η at $\widehat{\eta}^{(\ell)}$, and fit the postulated models $\mu_{\eta_k}(\bar{x}_k, \bar{a}_{k-1}; \xi_k)$ as described in the main paper to obtain $\widehat{\xi}_k^{(\ell)}$, $k = 1, \ldots, K$.

(iii) Substitute the fitted models into DR(η) in (6) of the main paper, continuing to hold ξ_k in these models fixed at $\widehat{\xi}_k^{(\ell)}$, and maximize DR(η) in η where it appears elsewhere in the expression for DR(η) (so not refitting the models $\mu_{\eta_k}(\bar{x}_k, \bar{a}_{k-1}; \xi_k)$ for each η encountered). Let $\ell = \ell + 1$, and call the resulting maximizing value $\widehat{\eta}^{(\ell)}$. One can stop and use $\widehat{\eta}^{(\ell)}$ as an approximation to $\widehat{\eta}_{\mathrm{DR}}^{\mathrm{opt}}$, or return to step (ii) and iterate one or more times.

7. IMPLEMENTATION OF GENETIC ALGORITHM

In the third simulation scenario of Section 5 of the main paper with $K = 3$ decision points, the dimension of η makes a simple grid search untenable for maximization of IPWE(η), DR(η), and AIPWE(η) in η. Because of the nonsmooth nature of these quantities as functions of η, standard optimization algorithms may be problematic. Accordingly, to carry out the required maximizations of IPWE(η), DR(η), and AIPWE(η) in η, we used a genetic algorithm discussed by Goldberg (1989), implemented in the `rgenoud` package in R (Mebane & Sekhon, 2011). In the `genoud` function, we adopted all default settings except we took `max=TRUE`; `optim.method = Nelder-Mead`, recommended in the documentation for discontinuous objective functions; and `pop.size = 5000`, which we determined to be sufficiently large to achieve satisfactory results via preliminary testing. We took `starting.values = c(0,0,0)`, and set the `Domains` matrix to be the 3×2 matrix with columns $\{\min(X_1), \min(X_2), \min(X_3)\}^{\mathrm{T}}$ and $\{\max(X_1), \max(X_2), \max(X_3)\}^{\mathrm{T}}$, where each row corresponds to lower and upper bounds on each element of η, so that the algorithm searched in this region.

8. Additional simulation results

We report here on results under a more complex simulation scenario with $K = 2$ decision points; here, the rules at each decision point involve several variables and allow all possible combinations of treatment options. Defining $X_1 = (X_{11}, X_{12}, X_{13})^{\mathrm{T}}$, X_{1j}, $j = 1, 2, 3$, were generated as independent $N(0, 1)$; given X_1, A_1 was Bernoulli with success probability $\mathrm{pr}(A_1 = 1 \mid X_1) = \mathrm{expit}(0.5X_{11} + 0.5X_{12} - 0.25X_{13})$. For $X_2 = (X_{21}, X_{22}, X_{23})^{\mathrm{T}}$, X_{2j}, $j = 1, 2, 3$, were independent $N(0, 1)$; given X_2, A_2 was Bernoulli with $\mathrm{pr}(A_2 = 1 \mid X_2) = \mathrm{expit}(0.5X_{21} - 0.25X_{22} + 0.5X_{23})$, and the outcome was generated as $Y \sim Z - (X_{11} + X_{12} + X_{13} - 1)^2 \times \{A_1 - I(2X_{11} + 2X_{12} - X_{13} - 0.3 > 0)\}^2 - (X_{21} + X_{22} + X_{23} - 1)^2\{A_2 - I(2X_{21} - X_{22} + 2X_{23} - 0.3 > 0)\}^2$, where $Z \sim N(10, 1)$. Thus, $g^{\mathrm{opt}} = (g_1^{\mathrm{opt}}, g_2^{\mathrm{opt}})$ with $g_1^{\mathrm{opt}}(x_1) = I(2x_{11} + 2x_{12} - x_{13} - 0.3 > 0)$ and $g_2^{\mathrm{opt}}(\bar{x}_2, a_1) = I(2x_{21} - x_{22} + 2x_{23} - 0.3 > 0)$, and $E\{Y^*(g^{\mathrm{opt}})\} = 10$.

The true Q- and Q-contrast functions are complicated, and it would be impossible to posit correct models in practice. For A-learning, we took $h_2(x_1, x_2, a_1; \alpha_2) = \alpha_{20} + \alpha_{21}x_{11} + \alpha_{22}x_{12} + \alpha_{23}x_{13} + \alpha_{24}x_{21} + \alpha_{25}x_{22} + \alpha_{26}x_{23} + a_1(\alpha_{27}x_{11} + \alpha_{28}x_{12} + \alpha_{29}x_{13})$, $C_2(x_1, x_2, a_1; \psi_2) = \psi_{20} + \psi_{21}x_{21} + \psi_{22}x_{22} + \psi_{23}x_{23}$, $h_1(x_1; \alpha_1) = \alpha_{10} + \alpha_{11}x_{11} + \alpha_{12}x_{12} + \alpha_{13}x_{13}$ and $C_1(x_1; \psi_1) = \psi_{10} + \psi_{11}X_{11} + \psi_{21}X_{12} + \psi_{13}X_{13}$. For Q-learning, we analogously posited Q-functions $Q_2(x_1, x_2, a_1, a_2; \beta_2) = \beta_{20} + \beta_{21}x_{11} + \beta_{22}x_{12} + \beta_{23}x_{13} + \beta_{24}x_{21} + \beta_{25}x_{22} + \beta_{26}x_{23} + a_1(\beta_{27}x_{11} + \beta_{28}x_{12} + \beta_{29}x_{13}) + a_2(\psi_{20} + \psi_{21}x_{21} + \psi_{22}x_{22} + \psi_{23}x_{23})$ and

$Q_1(x_1, a_1; \beta_1) = \beta_{10} + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13} + a_1(\psi_{10} + \psi_{11}x_{11} + \psi_{12}x_{12} + \psi_{13}x_{13}$. Thus, the Q-contrast and Q-functions are misspecified. For the propensity scores, we considered correctly specified models $\pi_2(x_1, x_2, a_1; \gamma_2) = \text{expit}(\gamma_{20} + \gamma_{21}X_{21} + \gamma_{22}X_{22} + \gamma_{23}X_{23})$ and $\pi_1(x_1; \gamma_1) = \text{expit}(\gamma_{10} + \gamma_{11}X_{11} + \gamma_{12}X_{12} + \gamma_{13}X_{13})$ and incorrect versions $\pi_2(x_1, x_2, a_1; \gamma_2) = \gamma_2$ and $\pi_1(x_1; \gamma_1) = \gamma_1$.

For the proposed methods, we took $\mathcal{G}_\eta$ to have elements $g_\eta = (g_{\eta_1}, g_{\eta_2})$, where $g_{\eta_2}(\bar{x}_2, a_1) = I(\eta_{20} + \eta_{21}x_{21} + \eta_{22}x_{22} + \eta_{23}x_{23} > 0)$, $g_{\eta_1}(x_1) = I(\eta_{10} + \eta_{11}x_{11} + \eta_{12}x_{12} + \eta_{13}x_{13} > 0)$, $\eta_2 = (\eta_{20}, \eta_{21}, \eta_{22}, \eta_{23})^{\mathrm{T}}$, $\eta_1 = (\eta_{10}, \eta_{11}, \eta_{12}, \eta_{13})^{\mathrm{T}}$ and thus $\eta = (\eta_1^{\mathrm{T}}, \eta_2^{\mathrm{T}})^{\mathrm{T}}$. Clearly, $g^{\text{opt}} \in \mathcal{G}_\eta$. Expressed in this form, regimes in $\mathcal{G}_\eta$ do not have a unique representation. For computational convenience in automating the simulations, we achieved uniqueness by normalizing the coefficients in each rule, imposing $\{(\eta_{21}, \eta_{22}, \eta_{23})(\eta_{21}, \eta_{22}, \eta_{23})^{\mathrm{T}}\}^{1/2} = 1$ and $\{(\eta_{11}, \eta_{12}, \eta_{13})(\eta_{11}, \eta_{12}, \eta_{13})^{\mathrm{T}}\}^{1/2} = 1$. Thus, $g^{\text{opt}} \in \mathcal{G}_\eta$ corresponds to $\eta^{\text{opt}} = (\eta_1^{\text{opt}\,\mathrm{T}}, \eta_2^{\text{opt}\,\mathrm{T}})^{\mathrm{T}}$, $\eta_1^{\text{opt}} = (-0.1, 0.67, 0.67, -0.33)^{\mathrm{T}}$, $\eta_2^{\text{opt}} = (-0.1, 0.67, -0.33, 0.67)^{\mathrm{T}}$. We used the same propensity models, and, for (6) and (7) in main paper, the same Q-function models as above. For (6), we posited the same models as those for the Q-functions, as in the previous simulations.

Because the high dimension of η made a grid search infeasible, to carry out the maximizations, we used a genetic algorithm discussed by Goldberg (1989), implemented in the `rgenoud` package in R (Mebane & Sekhon, 2011). In the `genoud` function, we adopted all default settings except we took `max=TRUE`; `optim.method = Nelder-Mead`, recommended in the documentation for discontinuous objective functions; and `pop.size = 5000`, which we determined to be sufficiently large to achieve satisfactory results via preliminary testing. We took `starting.values = c(0,0,0,0,0,0,0,0)`, and set the `Domains` matrix to be the 8×2 matrix with columns $(-1, -1, -1, -1, -1, -1, -1, -1)^{\mathrm{T}}$ and

$(1,1,1,1,1,1,1,1)^{\mathrm{T}}$, where each row corresponds to lower and upper bounds on each element of η, so that the algorithm searched in this region. As above, to identify a unique $\widehat{\eta}^{\text{opt}}$, we imposed $\{(\eta_{21},\eta_{22},\eta_{23})(\eta_{21},\eta_{22},\eta_{23})^{\mathrm{T}}\}^{1/2}=1$ and $\{(\eta_{11},\eta_{12},\eta_{13})(\eta_{11},\eta_{12},\eta_{13})^{\mathrm{T}}\}^{1/2}=1$ at the value of $\widehat{\eta}^{\text{opt}}$ obtained from `genoud` for each Monte Carlo data set.

Table 1 presents the results, which are qualitatively similar to those for the scenarios in the main paper.

Table 1. *Results for the additional simulation scenario, 1000 Monte Carlo data sets, $n = 500$. For the true optimal regime $g^{\rm opt} = g^{\rm opt}_{\eta} \in \mathcal{G}_{\eta}$, $\eta_1^{\rm opt} = (-0.1, 0.67, 0.67, -0.33)^{\rm T}$, $\eta_2^{\rm opt} = (-0.1, 0.67, -0.33, 0.67)^{\rm T}$ and $E\{Y^*(g^{\rm opt}_{\eta})\} = 10$.*

Estimator	$\widehat{\eta}_{10}$	$\widehat{\eta}_{11}$	$\widehat{\eta}_{12}$	$\widehat{\eta}_{13}$	$\widehat{\eta}_{20}$	$\widehat{\eta}_{21}$	$\widehat{\eta}_{22}$	$\widehat{\eta}_{23}$	$\widehat{E}(\widehat{\eta}^{\rm opt})$	SE	Cov.	$E(\widehat{\eta}^{\rm opt})$
Q-learning	-0.46 (0.11)	0.70 (0.07)	0.69 (0.07)	0.09 (0.13)	-0.46 (0.12)	0.70 (0.07)	0.10 (0.13)	0.69 (0.07)	9.50 (0.20)	–	–	8.85 (0.10)
						Propensity score correct						
A-learning	-0.37 (0.09)	0.69 (0.06)	0.69 (0.06)	0.16 (0.11)	-0.37 (0.09)	0.69 (0.06)	0.16 (0.12)	0.69 (0.06)	9.97 (0.23)	–	–	8.82 (0.08)
AIPWE	-0.11 (0.16)	0.68 (0.08)	0.64 (0.09)	-0.30 (0.16)	-0.01 (0.14)	0.67 (0.10)	-0.22 (0.12)	0.68 (0.10)	10.05 (0.15)	0.16	95.5	9.69 (0.18)
DR	-0.12 (0.16)	0.68 (0.09)	0.63 (0.10)	-0.29 (0.19)	-0.01 (0.13)	0.68 (0.10)	-0.21 (0.12)	0.68 (0.10)	10.04 (0.16)	0.16	94.9	9.67 (0.19)
IPWE	-0.18 (0.33)	0.67 (0.25)	0.42 (0.28)	0.11 (0.47)	-0.07 (0.25)	0.63 (0.25)	-0.06 (0.24)	0.64 (0.24)	11.01 (0.37)	0.52	50.3	9.04 (0.33)
						Propensity score incorrect						
A-learning	-0.46 (0.11)	0.70 (0.07)	0.69 (0.07)	0.09 (0.13)	-0.46 (0.12)	0.70 (0.07)	0.10 (0.13)	0.69 (0.07)	9.50 (0.20)	–	–	8.85 (0.10)
AIPWE	-0.10 (0.15)	0.68 (0.07)	0.65 (0.07)	-0.32 (0.10)	-0.02 (0.12)	0.68 (0.07)	-0.24 (0.11)	0.68 (0.07)	10.19 (0.19)	0.20	88.8	9.73 (0.14)
DR	-0.10 (0.15)	0.68 (0.06)	0.64 (0.07)	-0.32 (0.10)	0.01 (0.12)	0.68 (0.07)	-0.23 (0.11)	0.68 (0.07)	10.25 (0.20)	0.21	78.7	9.72 (0.13)
IPWE	-0.06 (0.18)	0.71 (0.12)	0.61 (0.13)	-0.28 (0.14)	-0.01 (0.13)	0.69 (0.12)	-0.18 (0.13)	0.67 (0.13)	17.11 (0.82)	0.88	0.00	9.52 (0.16)

AIPWE, DR, and IPWE, estimators based on maximizing $\text{AIPWE}(\eta)$ $\text{DR}(\eta)$, and $\text{IPWE}(\eta)$, respectively; $\widehat{\eta}_{10}$, $\widehat{\eta}_{11}$, $\widehat{\eta}_{12}$, $\widehat{\eta}_{13}$, $\widehat{\eta}_{20}$, $\widehat{\eta}_{21}$, $\widehat{\eta}_{22}$, $\widehat{\eta}_{23}$, Monte Carlo average estimates (standard deviations); $\widehat{E}(\widehat{\eta}^{\rm opt})$, Monte Carlo average and standard deviation of the estimated values of the true $E\{Y^*(g^{\rm opt}_{\eta})\}$; SE, Monte Carlo average of sandwich standard errors; Cov.,coverage of associated 95% Wald-type confidence intervals for $E(\eta^{\rm opt})$; $E(\widehat{\eta}^{\rm opt})$, the Monte Carlo average and standard deviation of values $E\{Y^*(\widehat{g}^{\rm opt}_{\eta})\}$ obtained using 10^6 Monte Carlo simulations for each data set.

9. APPLICATION TO STAR*D

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was a multi-site, multistep clinical trial enrolling 4041 patients with nonpsychotic major depressive disorder to compare treatment options for patients who do not attain a satisfactory response with citalopram. The trial involved four levels, each consisting of a 12-week follow-up phase. All patients received citalopram during level 1. Patients without sufficient symptomatic benefit were randomized to level 2 treatments, classified as either (i) switch: sertraline, bupropion, venlafaxine, or cognitive therapy, or (ii) augment: citalopram plus one of bupropion, buspirone, or cognitive therapy. Patients assigned to cognitive therapy (switch or augment options) at level 2 were eligible, in case of inadequate improvement, to be randomized to level 2A switch options (bupropion or venlafaxine). All patients without adequate response at levels 2 and 2A were randomized to level 3 options, (i) switch: mirtazepine or nortriptyline or (ii) augment: add lithium or triiodothyronine. Patients without adequate improvement at level 3 continued to be randomized to level 4 switch options (tranylcypromine or mirtazepine combined with venlafaxine). The decision to proceed to the next level depended on clinician-rated versions of the Quick Inventory of Depressive Symptomatology (QIDS) score. At the end of each level, patients deemed to have sufficient improvement using that level's treatment did not move to future levels, where sufficient improvement was defined by 12-week QIDS score $\leq$ 5 (remission) or showing a 50$\%$ or greater decrease from the baseline score at the beginning of level 1 (successful reduction). See Rush et al. (2004) for details.

Following Schulte et al. (2013), we take level 2A to be part of level 2 and consider only levels 2 and 3, denoting entry to levels 2 and 3 as decision points 1 and 2, respectively ($K = 2$), and consider the 1260 patients who entered level 2, 330 of whom continued to level 3. Let A_k, $k = 1, 2$, be assigned treatment at decision k, taking values 0 (augment) or 1 (switch); both options are feasible for all subjects. Let X_{10} and X_{11} denote QIDS score at baseline and at decision $k = 1$, and define X_{12} to be the slope of QIDS score based on X_{10} and X_{11}, so that $X_1 = (X_{11}, X_{12})^{\mathrm{T}}$ is the information available immediately prior to the first decision. Let X_{21} denote QIDS score at decision $k = 2$ and X_{22} be the QIDS score slope based on X_{11} and X_{21}, so that $X_2 = (X_{21}, X_{22})^{\mathrm{T}}$ is the information available between decision points 1 and 2. Letting T be QIDS score at the end of level 3 and $L_0 = \max(5, X_{10}/2)$, define the outcome as $Y = -I(X_{21} \leq L_0)X_{21} - I(X_{21} > L_0)(X_{21} + T)/2$, the cumulative average negative QIDS score.

It can be deduced that $Q_2(\bar{x}_2, \bar{a}_2) = E(Y \mid \bar{X}_2 = \bar{x}_2, \bar{A}_2 = \bar{a}_2) = \{I(x_{21} \leq l_0) + I(x_{21} > l_0)/2\}x_{21} + I(x_{21} > l_0)E(-T \mid \bar{X}_2 = \bar{x}_2, \bar{A}_2 = \bar{a}_2, x_{21} > l_0)/2$. As in Schulte et al. (2013), who carried out preliminary exploratory analysis, we posited $Q_2(\bar{x}_2, \bar{a}_2; \beta_2) = -\{I(x_{21} \leq l_0) + I(x_{21} > l_0)/2\}x_{21} + I(X_{21} > l_0)(\beta_{20} + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}a_2)/2$, where β_2 can be estimated using the data from patients continuing to level 3; thus, $V_2(\bar{x}_2, a_1; \beta_2) = \{I(x_{21} \leq l_0) + I(x_{21} > l_0)/2\}x_{21} + I(x_{21} > l_0)\{\beta_{20} + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}I(\beta_{23} > 0)\}$. At decision 1, we specified $Q_1(x_1, a_1; \beta_1) = \beta_{10} + \beta_{11}x_{11} + \beta_{12}x_{12} + a_1(\beta_{13} + \beta_{14}x_{12})$. For A-learning, we analogously took $h_2(\bar{x}_2, a_1; \alpha_2) = \alpha_{20} + \alpha_{21}x_{21} + \alpha_{22}x_{22}$, $C_2(\bar{x}_2, a_1; \psi_2) = \psi_2$, $h_1(x_1; \alpha_1) = \alpha_{10} + \alpha_{11}x_{11} + \alpha_{12}X_{12}$ and $C_1(x_1; \psi_1) = \psi_{10} + \psi_{11}X_{12}$ and specified $\pi_2(\bar{x}_2, a_1; \gamma_2) = \text{expit}(\gamma_{20} + \gamma_{21}x_{21} + \gamma_{22}x_{22} + \gamma_{23}a_1)$ and $\pi_1(x_1; \gamma_1) = \text{expit}(\gamma_{10} + \gamma_{11}x_{11} + \gamma_{12}x_{12})$.

Q-learning suggests a treatment switch for patients with decision 1 QIDS slope greater than -0.97; A-learning assigns a switch for QIDS slope greater than -1.07. At the second decision, both suggest that all patients should switch. Using DR(η) to estimate $E\{Y^*(g^{\text{opt}})\}$ for all methods for consistency yields -7.97 (-8.50, -7.45) and -8.03 (-8.56,-7.50) for Q- and A-learning, respectively.

Analogous to the above, we define the class of regimes $\mathcal{G}_\eta$ with elements $g_\eta = (g_{\eta_1}, g_{\eta_2})$, where $g_{\eta_1}(x_1) = (x_{12} > \eta_1)$ and $g_{\eta_2}(\bar{x}_2, a_1) = I(x_{22} > \eta_2)$. Using the same Q-function and propensity models as above and estimating $\eta = (\eta_1, \eta_2)^{\mathrm{T}}$ by maximizing DR(η) and AIPWE(η)

in η via a grid search over all possible jump points $(x_{12,i}, x_{22,j})$, $i, j = 1, \ldots, n$, as in the simulations, $\widehat{\eta}^{\text{opt}}_{\text{AIPWE},1} = \widehat{\eta}^{\text{opt}}_{\text{DR},1} = -1.78$, suggesting that patients with larger decision 1 QIDS slope switch treatments, and $\widehat{\eta}^{\text{opt}}_{\text{AIPWE},2} = \widehat{\eta}^{\text{opt}}_{\text{DR},2} = -7.50$, the minimum value of X_{22}, indicating that all patients should switch treatments at the second decision. Using $\text{DR}(\eta)$ to estimate $E\{Y^*(g^{\text{opt}}_{\eta})\}$ yields -7.85 (-8.36, -7.33).

Acknowledgement

Data used in the analyses of § 9 were obtained from the limited access datasets distributed from the National Institutes of Health-supported Sequenced Treatment Alternatives to Relieve Depression (STAR*D). STAR*D focused on non-psychotic major depressive disorder in adults seen in outpatient settings. The primary purpose of this research study was to determine which treatments work best if the first treatment with medication does not produce an acceptable response. The study was supported by National Institute of Mental Health Contract # N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is NCT00021528.

This manuscript reflects the views of the authors and may not reflect the opinions or views of the STAR*D Study Investigators or the National Institutes of Health.

References

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison-Wesley.

Mebane, W. R. & Sekhon, J. S. (2011). Genetic optimization using derivatives: the rgenoud package for R. *J. Statist. Soft.* **42**, 1–26.

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

MOODIE, E. E. M., PLATT, R. W. & KRAMER, M. S. (2009). Estimating response-maximized decision rules with applications to breastfeeding. *J. Am. Statist. Assoc.* **104**, 155–165.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *J. Royal Statist. Soc., Ser. B* **58**, 331–366.

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

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–866.

RUSH, A. J., FAVA, M,, WISNIEWSKI, S. R., LAVORI, P.W., TRIVEDI, M.H., SACKEIM, H. A., THASE, M. E., NIERENBERG, A. A., QUITKIN, F.M., KASHNER, T. M., KUPFER, D. J., ROSENBAUM, J. F., ALPERT, J., STEWART, J. W., MCGRATH, P. J., BIGGS, M. M., SHORES-WILSON, K., LEBOWITZ, B. D., RITZ, L., NIEDEREHE, G. (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): rationale and design. *Control. Clin. Trials* **25**, 119–142.

SCHULTE, P.J., TSIATIS, A.A., LABER, E.B. & DAVIDIAN, M. (2013). Q- and A-learning methods for estimating optimal dynamic treatment regimes. Pre-print, arXiv:1202.4177. (In revision for *Statist. Sci.*)

TSIATIS, A.A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

[*Received January* 2011. *Revised June* 2011]