Web-based Supplementary Materials for

"C-learning: a New Classification Framework to Estimate Optimal Dynamic Treatment Regimes"

by Baqun Zhang and Min Zhang

Baqun Zhang

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P.R.China.

and

Min Zhang*

Department of Biostatistics, University of Michigan, Ann Arbor, U.S.A.

mzhangst@umich.edu

A: Backward Induction in terms of Potential Outcomes

The optimal dynamic treatment regime, g^{opt}, is a regime that maximizes expected outcomes were all patients in the population to follow it. As discussed in Schulte et al (2014) and Zhang et al (2013a), the optimal regime g^{opt} may be determined via dynamic programming, also referred to as backward induction. In the main manuscript, the backward induction is in terms of observed variables and here we define the optimal dynamic treatment regime using backward induction, as well as V-functions, in terms of potential outcomes. At the Kth decision point, for any $\bar{x}_K \in \bar{\mathcal{X}}_K$, $\bar{a}_{K-1} \in \bar{\mathcal{A}}_{K-1}$, define

$$g_K^{opt}(\bar{x}_K, \bar{a}_{K-1}) = \arg\max_{a_K \in \{0,1\}} E\{Y^*(\bar{a}_{K-1}, a_K) | \bar{X}_K^*(\bar{a}_{K-1}) = \bar{x}_K\}$$

$$V_K(\bar{x}_K, \bar{a}_{K-1}) = \max_{a_K \in \{0,1\}} E\{Y^*(\bar{a}_{K-1}, a_K) | \bar{X}_K^*(\bar{a}_{K-1}) = \bar{x}_K\}.$$

For $k = K-1, \ldots, 2$ and any $\bar{x}_k \in \bar{\mathcal{X}}_k$, $\bar{a}_{k-1} \in \bar{\mathcal{A}}_{k-1}$, define

$$g_k^{opt}(\bar{x}_k, \bar{a}_{k-1}) = \arg\max_{a_k \in \{0,1\}} E[V_{k+1}\{\bar{x}_k, X_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{X}_k^*(\bar{a}_{k-1}) = \bar{x}_k]$$

$$V_k(\bar{x}_k, \bar{a}_{k-1}) = \max_{a_k \in \{0,1\}} E[V_{k+1}\{\bar{x}_k, X^*_{k+1}(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{X}^*_k(\bar{a}_{k-1}) = \bar{x}_k].$$

For $k = 1$, $x_1 \in \mathcal{X}_1$, $g_1^{opt}(x_1) = \arg\max_{a_1 \in \{0,1\}} E[V_2\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1]$ and $V_1(x_1) = \max_{a_1 \in \{0,1\}} E[V_2\{x_1, X_2^*(a_1), a_1\} | X_1 = x_1]$. Note, here the V-functions are defined in terms of potential outcomes, whereas in the main manuscript they are defined in terms of observed variables. The two sets of definitions are equivalent under the assumed assumptions. See Schulte et al (2014) for details. Recall that $L_k \equiv (\bar{X}_k, \bar{A}_{k-1})$, by the definition of V-functions and g_k^{opt}, $k = K, \ldots, 1$, we also have

$$V_K(L_K) = E\{Y^*(\bar{A}_{K-1}, g_K^{opt}) | L_K\},$$
$$V_k(L_k) = E\{Y^*(\bar{A}_{k-1}, g_k^{opt}, g_{k+1}^{opt}, ..., g_K^{opt}) | L_k\}, \text{where} \quad k = K-1, \ldots, 1.$$

B: Proof of Theorem 1

At the K-th decision point, we have that

$$\begin{aligned}
g_K^{opt}(L_K) &= \arg\max_{g_K \in \mathcal{G}_K} E\{Y^*(\bar{A}_{K-1}, g_K)\} \\
&= \arg\max_{g_K \in \mathcal{G}_K} E[E\{Y^*(\bar{A}_{K-1}, g_K) | L_K\}] \\
&= \arg\max_{g_K \in \mathcal{G}_K} E[E\{Y^*(\bar{A}_{K-1}, g_K) | L_K, A_K = g_K\}] \quad (1) \\
&= \arg\max_{g_K \in \mathcal{G}_K} E[E\{Y | L_K, A_K = g_K\}] \quad (2) \\
&= \arg\max_{g_K \in \mathcal{G}_K} E\{E(Y | L_K, A_K = 1) g_K + E(Y | L_K, A_K = 0)(1 - g_K)\} \\
&= \arg\max_{g_K \in \mathcal{G}_K} E\{Q_K(L_K, 1) g_K + Q_K(L_K, 0)(1 - g_K)\} \\
&= \arg\max_{g_K \in \mathcal{G}_K} E[g_K\{Q_K(L_K, 1) - Q_K(L_K, 0)\}] \\
&= \arg\max_{g_K \in \mathcal{G}_K} E\{g_K(L_K) C_K(L_K)\}. \quad (3)
\end{aligned}$$

where equality (1) is due to the no unmeasured confounders (or sequential ignorability) assumption, equality (2) due to the consistency assumption.

Considering the term $g_K(L_K)C_K(L_K)$, a key step is to express $C_K(L_K)$ in terms of the magnitude and the sign, ie, $C_K(L_K) = Z_K|C_K(L_K)| - (1 - Z_K)|C_K(L_K)|$, where $Z_K = I(C_K(L_K) > 0)$. Substituting this into $g_K(L_K)C_K(L_K)$, as $g(L_K)$ takes values $\{0,\ 1\}$, straightforward algebra shows and it is easy to check that

$$g(L_K)C_K(L_K) = Z_K|C_K(L_K)| - |C_K(L_K)|I(Z_K \neq g(L_K)).$$

Substituting this back into to (3), it follows then $g_K^{opt}(L_K)$

$$\begin{aligned} &= \arg\max_{g_K \in \mathcal{G}_K} E\{Z_K|C_K(L_K)| - |C_K(L_K)|I(Z_K \neq g(L_K))\} \\ &= \arg\min_{g_K \in \mathcal{G}_K} E\{|C_K(L_K)|I(Z_K \neq g(L_K))\}. \end{aligned}$$

Then recursively for stage k, $k = K-1, \ldots, 1$, the optimal decision rule can be written as

$$\begin{aligned} g_k^{opt}(L_k) &= \arg\max_{g_k \in \mathcal{G}_k} E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})\} \\ &= \arg\max_{g_k \in \mathcal{G}_k} E[E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})|L_k\}] \\ &= \arg\max_{g_k \in \mathcal{G}_k} E[E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})|L_k, A_k = g_k\}], \end{aligned} \tag{4}$$

due to no unmeasured confounders assumption. We also have that

$$\begin{aligned} & E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})|L_k, A_k = g_k\} \\ = \ & E[E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})|L_k, A_k = g_k, X_{k+1}^*(\bar{A}_{k-1}, g_k)\}|L_k, A_k = g_k] \\ = \ & E[E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, ..., g_K^{opt})|L_k, A_k = g_k, X_{k+1}\}|L_k, A_k = g_k] \qquad (5) \\ = \ & E\{V_{k+1}(L_k, g_k, X_{k+1})|L_k, A_k = g_k\}, \qquad (6) \end{aligned}$$

where the last equality is due to the definitions of the value functions $V_k, k = K, \ldots, 1$, in

terms of potential outcomes (see Appendix A). Substituting (6) back to (4), we have

$$\begin{aligned} g_k^{opt}(L_k) &= \arg\max_{g_k \in \mathcal{G}_k} E[E\{V_{k+1}(L_k, g_k, X_{k+1})|L_k, A_k = g_k\}] \\ &= \arg\max_{g_k \in \mathcal{G}_k} E[E\{V_{k+1}(L_k, 1, X_{k+1})|L_k, A_k = 1\}g_k \\ &+ E\{V_{k+1}(L_k, 0, X_{k+1})|L_k, A_k = 0\}(1 - g_k)] \\ &= \arg\max_{g_k \in \mathcal{G}_k} E\{Q_k(L_k, 1)g_k + Q_k(L_k, 0)(1 - g_k)\} \\ &= \arg\max_{g_k \in \mathcal{G}_k} E\{g_k(L_k)C_k(L_k)\}, \end{aligned}$$

where $C_k(L_k) = Q_k(L_k, 1) - Q_k(L_k, 0)$. Then by the same derivation as for stage K, it follows that

$$g_k^{opt} = \arg\min_{g_k \in \mathcal{G}_k} E\{|C_k(L_k)|I(Z_k \neq g(L_k))\}, \text{where } Z_k = I(C_k(L_k) > 0), k = K-1, \ldots, 1.$$

C: Proof of Proposition 1

At the K-th decision point, we have

$$\begin{aligned} V_K(L_K) &= E\{Y^*(\bar{A}_{K-1}, g_K^{opt})|L_K\} \\ &= E\{Y|L_K, A_K = g_K^{opt}\} \\ &= Q_K(L_K, 1)g_K^{opt} + Q_K(L_K, 0)(1 - g_K^{opt}) \\ &= Q_K(L_K, 0) + \{Q_K(L_K, 1) - Q_K(L_K, 0)\}g_K^{opt}, \end{aligned} \tag{7}$$

where the second equality is due to the consistency and nounumeasured confounders as-

sumptions. We also have that

$$\begin{aligned}
& E[Y + \{Q_K(L_K,1) - Q_K(L_K,0)\}\{g_K^{opt}(L_K) - A_K\}|L_K] \\
= \; & E\{E[Y + \{Q_K(L_K,1) - Q_K(L_K,0)\}\{g_K^{opt}(L_K) - A_K\}|L_K, A_K]|L_K\} \\
= \; & E[Q_K(L_K,A_K) + \{Q_K(L_K,1) - Q_K(L_K,0)\}\{g_K^{opt}(L_K) - A_K\}|L_K] \\
= \; & E[Q_K(L_K,0) + \{Q_K(L_K,1) - Q_K(L_K,0)\}A_K \\
& + \{Q_K(L_K,1) - Q_K(L_K,0)\}\{g_K^{opt}(L_K) - A_K\}|L_K] \\
= \; & Q_K(L_K,0) + \{Q_K(L_K,1) - Q_K(L_K,0)\}g_K^{opt}(L_K) \\
= \; & V_K(L_K),
\end{aligned}$$

where the last equality is due to (7).

Similarly at stage $k, k = K-1, \ldots, 1$, we have

$$\begin{aligned}
V_k(L_k) &= E\{Y^*(\bar{A}_{k-1}, g_k^{opt}, g_{k+1}^{opt}, ..., g_K^{opt})|L_k\} \\
&= E\{Y^*(\bar{A}_{k-1}, g_k^{opt}, g_{k+1}^{opt}, ..., g_K^{opt})|L_k, A_k = g_k^{opt}\} \\
&= E\{V_{k+1}(L_k, g_k^{opt}, X_{k+1})|L_k, A_k = g_k^{opt}\} \qquad (8) \\
&= Q_k(L_k,1)g_k^{opt} + Q_k(L_k,0)(1 - g_k^{opt}) \\
&= Q_k(L_k,0) + \{Q_k(L_k,1) - Q_k(L_k,0)\}g_k^{opt}, \qquad (9)
\end{aligned}$$

and

$$\begin{aligned}
& E[V_{k+1}(L_{k+1}) + \{Q_k(L_k,1) - Q_k(L_k,0)\}\{g_k^{opt}(L_k) - A_k\}|L_k] \\
= \; & E(E[V_{k+1}(L_{k+1}) + \{Q_k(L_k,1) - Q_k(L_k,0)\}\{g_k^{opt}(L_k) - A_k\}|L_k, A_k]|L_k) \\
= \; & E[Q_k(L_k,A_k) + \{Q_k(L_k,1) - Q_k(L_k,0)\}\{g_k^{opt}(L_k) - A_k\}|L_k] \\
= \; & E[Q_k(L_k,0) + \{Q_k(L_k,1) - Q_k(L_k,0)\}A_k \\
& + \{Q_k(L_k,1) - Q_k(L_k,0)\}\{g_k^{opt}(L_k) - A_k\}|L_k] \\
= \; & Q_k(L_k,0) + \{Q_k(L_k,1) - Q_k(L_k,0)\}g_k^{opt}(L_k) \\
= \; & V_k(L_k),
\end{aligned}$$

where the last step is due to (9).

D: Other Methods and Relationship

We discuss the connections and distinctions between the proposed C-learning and other direct optimization methods (method of Zhang et al., 2013, OWL-based methods) and outcome regression based methods (Q-and A- learning).

The AIPWE based method of Zhang et al.(2012a, 2013) is a direct optimization method that directly maximizes the expected potential outcomes under a regime, $E\{Y^*(g)\}$, across a class of regimes. It estimates $E\{Y^*(g)\}$ using the AIPWE estimator

$$AIPWE(g) = \sum_{i=1}^{n} \frac{\prod_{k=1}^{K} I\{g_k(L_{ki}) = A_{ki}\}}{\prod_{k=1}^{K} \widehat{\pi}_k(A_{ki}, L_{ki})} Y_i + \sum_{k=1}^{K} \text{Augmentation term}_k, \tag{10}$$

where $\widehat{\pi}_k(a_k, L_{ki})$ estimates $P(A_{ki} = a_k | L_{ki})$ and the augmentation terms involve fitted values for Q-functions or other outcome regression models (see Zhang et al., 2013 for explanation). By appropriate choosing the augmentation term, the AIPWE has the double-robustness property in the sense that if the treatment probabilities (propensity scores) or outcome regression models, but not necessarily both, are correctly specified, then AIPWE(g) is consistent for $E\{Y^*(g)\}$ and the augmentation terms can be used to improve efficiency even if they are misspecified. If the augmentation terms are taken to be zero, this reduces to IPWE, which is also studied in Zhang et al.(2012ab, 2013) and is known to be inefficient. The method of Zhang et al.(2012a, 2013) then estimates the optimal treatment regimes at all stages simultaneously by directly maximizing $AIPWE(g)$ across regimes in a restricted class of regimes indexed by a finite number of parameters.

The proposed C-learning is also a direct optimization method by essentially directly estimating $E\{Y^*(g)\}$ (see explanation below) and estimation is based on AIPWE as well. It differs from Zhang et al. (2012a, 2013) in two important aspects. First, C-learning transforms the problem of estimating the optimal treatment regime into a classification

problem. Existing powerful classification algorithms (eg., CART, SVM, etc.) can be used to facilitate optimization, leading to more powerful and flexible estimation of regimes. This point was also discussed in Zhang et al (2012b) in the single decision setting. In Zhang et al, (2013), the optimization of regimes is among a restricted class of regimes indexed by a finite number of parameters. However, in C-learning it can accommodate larger classes, not necessarily a parametric class. For example, we illustrated optimization using CART among all regimes of the form of a decision tree. Second, in contrast to Zhang et al. (2013) which estimates regimes at all stages simultaneously, C-learning estimates the optimal regime at each stage sequentially, which has important implications for improved performance. That is, Zhang et al. (2013) estimates $E\{Y^*(g)\}$ using AIPWE, where $g = (g_1, \ldots, g_K)$ is a regime with multiple stages, and then optimize it across a class of regimes $\mathcal{G}$ to estimate g^{opt}. However, from the proof for Theorem 1 and discussion in Zhang et al. (2012b) on AIPWE of contrast functions, C-learning sequentially estimates $E\{Y^*(\bar{A}_{K-1}, g_K)\}$, $\ldots$, $E\{Y^*(\bar{A}_{k-1}, g_k, g_{k+1}^{opt}, \ldots, g_K^{opt})\}$, $\ldots$, $E\{Y^*(g_1, g_2^{opt}, \ldots, g_K^{opt})\}$ using a series of AIPWEs and then optimizes each expected potential outcomes sequentially across $\mathcal{G}_k$ to estimate g_k^{opt}, $k = K, \ldots, 1$. As our simulation shows (Table 1), even when both C-learning and the method of Zhang et al. (2013) are based on AIPWEs with same propensity score and augmentation term models and the optimization are carried out across the same class of regimes, C-learning still outperforms Zhang et al. (2013) considerably. This is due to the difference in maximization (sequential vs. simultaneous) and the sample size used in estimating the optimal regime at each stage. Consider the last stage (stage K), data on all subjects are used to estimate the optimal regime in the proposed C learning. From the missing data perspective, the potential outcome of a subject is observed as long as the treatment at stage K is consistent with a regime, regardless of treatments received prior to K because covariate and treatment histories at previous stages are treated as baseline covariates. Once we estimate the optimal treatment regime at stage K, we move

backward. Consider estimation at stage k, in C-learning essentially we impute the expected potential outcomes for stage k assuming the optimal decisions are made in the future. We then estimate the optimal treatment regime at stage k using data from all subjects. Intuitively, in C-learning, at each stage, the best effort has been made to estimate the optimal treatment regime at that stage. In the method of Zhang et al (2013), however, the optimal treatment regimes at different stages are estimated simultaneously. From the coarsening data perspective on which the method of Zhang et al. (2013) was derived from, the potential outcome of a subject is observed only if the treatments at all stages are consistent with a regime, and therefore, we have more missing information as compared with the C-learning method. This explains why even when both C-learning and the method of Zhang et al (2013) use AIPWEs with same propensity score and augmentation term models and optimize across the same class of regimes, C-learning still has better performance.

The simultaneous outcome weighted learning (SOWL) of Zhao et al.(2015) is in principle the same as the method of Zhang et al. (2013) with the use of the particular IPWE. Following OWL (Zhao et al. 2012), by mimicking the idea of SVM, instead of directly optimizing IPWE as in Zhang et al.(2013), they substitute a continuous and concave surrogate function to replace the indicator function (see also the discussion for BOWL below) in the original objective function to facilitate optimization.

The backward outcome weighted learning (BOWL) of Zhao et al.(2015) is also based on optimizing IPWE of $E\{Y^*(g)\}$. However, instead of learning decision rules at all stages simultaneously, optimization is carried out sequentially. After obtaining the estimated optimal regimes at stages $K, ..., k+1$, denoted as $\widehat{g}_K^{opt}, ..., \widehat{g}_{k+1}^{opt}$, to estimate $\widehat{g}_k^{opt}$ it aims to maximize the IPWE, ie,

$$\sum_{i=1}^{n} \frac{Y_i \prod_{j=k+1}^{K} I\{\widehat{g}_j^{opt}(L_{ji}) = A_{ji}\}}{\prod_{j=k}^{K} \widehat{\pi}_j(A_{ji}, L_{ji})} I\{g_k(L_{ki}) = A_{ki}\},$$

which is equivalent to minimizing

$$\sum_{i=1}^{n} \frac{Y_i \prod_{j=k+1}^{K} I\{\widehat{g}_j^{opt}(L_{ji}) = A_{ji}\}}{\prod_{j=k}^{K} \widehat{\pi}_j(A_{ji}, L_{ji})} I\{g_{ik}(L_{ik}) \neq A_{ik}\}.$$

For simplicity taking $K = 1$, the IPWE estimator is the empirical analogue of $E[YI\{g(X) = A\}/\pi(A, X)]$, where $\pi(a, X) = Pr(A = a|X)$, and maximizing it is equivalent to minimizing $E[YI\{g(X) \neq A\}/\pi(A, X)]$. Because of the particular form of IPWE, where a term $I\{g(X) \neq A\}$ is involved, $I\{g(X) \neq A\}$ can be viewed as a zero-one loss in a classification problem and $Y/\pi(A, X)$ can be viewed as the weight when Y is positive (or when Y is bounded and can therefore be transformed to a positive random variable). It is easy to see that the classification perspective in BOWL is to classify patients based on his/her characteristics to classes that actually received treatment A=0 or 1, i.e., a classification error is made if $g(X) \neq A$. This is indeed the idea behind IPWE by viewing the problem as a missing data problem in the sense that the potential outcome for a subject under a regime is missing if the actually received treatment is not the one determined by the regime, i.e., $g(X) \neq A$; see Zhang et al. (2012a) for details. Viewing this as a classification problem and, instead of using the 0-1 loss, BOWL uses a hinge loss to facilitate the optimization, ie., replacing the indicator function by a continuous and concave surrogate function. The optimization at stage k is among subjects who follow $\widehat{g}_K^{opt}, ..., \widehat{g}_{k+1}^{opt}$ and the number of subjects in learning optimal decision rules is decreasing geometrically as k decreases, which further loses information and leads to inefficient estimation.

Therefore, both OWL (BOWL and SOWL) and the robust AIPWE based method of Zhang et al.(2012a and 2013) are direct optimization approach. The difference is that the method of Zhang et al.(2012 and 2013) uses the more general and efficient AIPWE but OWL uses IPWE, which is less efficient as it does not incorporate information in outcome regression models and less robust due to lack of the double robustness property. By replacing indicator function by a continuous and concave function, OWL is more flexible in

optimizing among a large class of regimes. However, the use of SVM in OWL and the transformation of the problem to a classification problem is predicated on an IPWE estimator and cannot readily generalize to other more efficient estimators. Also the classification idea of SVM is used purely as a convenient tool for optimization instead of being a meaningful classification targeting the goal of individualizing treatments since in this classification perspective it tries to classify patients based on his/her characteristics to classes that actually received treatment A=0 or 1. As also pointed out by Zhou, et al. (2015), due to this, the estimated treatment regime from OWL-based methods tries to keep treatment assignments that subjects actually received, which is an undesirable feature as theoretically by definition the optimal treatment regime should not depend on how treatment is actually assigned in observed data. The proposed C-learning does not suffer from this feature as our classification perspective is different from that of OWL-based methods. In the classification perspective of C-learning, it is easy to see from Theorem 1 that it aims to classify patients, based on patients characteristics, to the class that would potentially benefit from one treatment relative to the other and hence should receive the particular treatment, i.e., a classification error is made if $g(X) \neq I\{C(X) > 0\}$. As a matter of fact, in Theorem 1, it is clear that the optimal treatment regime in our classification perspective does not depend on the observed treatment A, as it should be by the original definition, which is in contrast with OWL-based methods. Therefore, the meaning of classification is consistent with the goal of optimizing treatment and it does not suffer from the issue of OWL-based methods.

In contrast with outcome regression based methods (Q- and A-learning methods), C-learning is a direct optimization method that estimates the optimal treatment regime by directly optimizing $E\{Y^*(g)\}$. Direct optimization methods enjoy some robustness property against outcome model misspecification. They lead to good estimate of the optimal treatment regime as long as $E\{Y^*(g)\}$ is consistently estimated, not necessarily requir-

ing the outcome regression models to be correctly specified. For example, in randomized clinical trials $E\{Y^*(g)\}$ can always be consistently estimated using AIPWE and IPWE methods and in observational studies $E\{Y^*(g)\}$ can be consistently estimated when either propensity score models or outcome regression models are correctly specified. Outcome regression models do not dictate the performance but good outcome regression models can be used to improve performance. In contrast, Q- and A-learning methods are outcome regression based methods. Q-learning models conditional expectation of outcomes and A-learning models the contrast of expectations between treatments. Estimated contrast functions directly determine the estimated regimes and, as a result, the performance of the resulting estimate depends on whether the outcome regression models are correctly specified. A-learning is more robust than Q-learning since it does not require the model for conditional means to be fully correctly specified, but it still requires the parametric form of the contrast functions to be correctly specified. See Zhang et al. (2012a and 2013), Zhao et al. (2012 and 2015) and Kang, et al.(2014) for more discussion on direct optimization methods.

To summarize, C-learning is a direct optimization method as opposed to outcome regression-based methods (Q-and A-learning) and it enjoys more protection against model misspecification. As the direct optimization method of Zhang et al.(2012a and 2013), outcome regression models (Q-functions) are used to improve efficiency, in contrast with OWL; however, outcome regression models in C-learning do not dictate the form of the optimal treatment regime, in contrast with Q- and A-learning. C-learning successfully transforms the problem of identifying the optimal treatment regime into a sequential classification problem. In C-learning, the meaning of classification is consistent with that of individualizing decision rules based on patient characteristics, which is in contrast with OWL-based methods and offers considerable improvement in performance over BOWL and other OWL-based methods. Instead of using simultaneous optimization as in Zhang et al.

(2013), C-learning uses backward induction to backward sequentially identify the optimal decision rules, which leads to big improvement in performance. In our backward sequential optimization, at each stage it is able to optimize decision rules based on all subjects, whereas the sample size in BOWL decreases geometrically as it can only use subjects who have followed the estimated optimal decisions rules at the future stages. To the best our knowledge, this is the first direct optimization method that is able to use backward induction to sequentially estimate regimes without having to decrease sample size at later stages.

E: C-learning Algorithm

We summarize the C-learning algorithm as follows.

1. *At stage K, based on data $(Y_i, L_{Ki}, A_{Ki}), i = 1, \ldots, n$,*

   1.1 *Build model for $P(A_K = 1|L_K)$ to obtain estimate of the propensity score $P(A_{Ki} = 1|L_{Ki})$, denoted as $\widehat{\pi}_K(L_{Ki})$, for each subject.*

   1.2 *Build models for $Q_K(L_K, A_K)$ and obtain estimates $\widehat{Q}_K(L_{Ki}, a_K)$, $a_K = 0, 1$.*

   1.3 *Estimate $C_K(L_{Ki})$ for each subject by $\widehat{C}_K(L_{Ki})$ as in (3) or more generally (6) of the main manuscript.*

   1.4 *Estimate g_K^{opt} according to (7) by some optimization/classification technique, denoted as $\widehat{g}_{C,K}^{opt}$.*

   1.5 *Estimate V_{Ki} by $\widetilde{V}_{Ki}$ according to (5).*

2. *Similarly for stage $k = K - 1, \ldots, 1$ sequentially, repeat 1.1-1.5 based on "data" $(\widetilde{V}_{(k+1)i}, L_{ki}, A_{ki}), i = 1, \ldots, n$, to obtain estimate of g_k^{opt}, denoted as $\widehat{g}_{C,k}^{opt}$*

F: Additional Results

We conducted additional simulations to evaluate the performance of C-learning relative to BOWL and other OWL-based methods under scenarios 1 and 2 of Zhao et al (2015);

results are shown in Table s1. Implementation of C-learning is similar to that in Table 2 (C-learning-RF) of the main manuscript. In these two scenarios, C-learning also outperforms OWL-based methods for reasons explained in the main manuscript and in Supplementary Material D, namely, the difference in classification perspective, whether or not incorporating information from outcome regression models and sample size used in optimization at later stages. We note that here results for BOWL, IOWL and SOWL are directly copied from Zhao et al, (2015) because we were unable to reproduce results and the performance of OWL-based methods from our simulations are not as good. One reason for not being able to reproduce results of OWL-based methods is due to how one handles negative outcomes as OWL-based methods may be ill-behaved when outcomes can be negative (Chen et al., 2017). OWL-based methods are assuming outcomes are positive. In simulation studies, as outcomes are generated from normal distributions, one needs to shift the outcomes to make it positive. As pointed out by Zhou et al.(2015) and also noted by ourselves and Chen et al. (2017), results of OWL-based methods are very sensitive to a simple shift and also shifting further impacts performance.

In our main manuscript, BOWL is modified to overcome the issue of dealing with negative outcomes by using the IPWE method of Zhang et al. (2012b) to get the objective function and then replacing the indicator function to a concave surrogate function, i.e., weight is $|Y|/\{A\pi + (1-A)(1-\pi)\}$ and class label is $(1-A)$. As discussed by Zhang et al. (2012b), this is equivalent to OWL when Y is positive. However, it can handle both positive and negative outcomes and is not ill-behaved when Y is negative. This modification is similar to the remedy proposed by Chen et al. (2017) and is the same when $\pi = 0.5$. We report results of the original BOWL under scenario 1 of the main manuscript in Table s2. To implement the original BOWL, we have followed the recommendation (personal communication) of the first author to handle negative outcomes, i.e., shifting outcomes by subtracting the minimum observed outcome, and used code kindly shared by the first author

Table s1: Simulation results under scenarios 1 and 2 of Zhao et al. (2015).

	n	BOWL	IOWL	SOWL	C-learning-RF
Scenario 1	200	4.50(0.77)	4.21(0.84)	5.93(0.75)	6.43(0.22)
	400	5.81(0.33)	5.00(0.60)	6.19(0.39)	6.57(0.10)
Scenario 2	200	2.85(0.27)	2.87(0.28)	3.12(0.08)	3.68(0.06)
	400	3.10(0.13)	3.05(0.20)	3.21(0.09)	3.71(0.03)

for the optimization. As shown in Table s2, original BOWL behaves significantly worse than the modified BOWL (denoted as BOWL). Similar pattern of relative performances of different versions of BOWL is observed under scenarios 2 and 3 of the main manuscript (results not shown).

We report simulation results under scenarios in the main manuscript using sample size $n = 100$ in Table s3. Results under this smaller sample size are overall consistent with results reported in the main manuscript and demonstrate the superior performance of the proposed method relative to BOWL, Q-learning and the method of Zhang et al. (2013). We note that when the sample size is as small as $n = 100$, under scenarios 2 and 3 Q-learning† and Zhang et al. $(2013)^{\dagger}$ seem to have slightly better performance. However, this is due to that in our implementation Q-learning† and Zhang et al. $(2013)^{\dagger}$ only used relevant variables among the high dimensional set of covariates to construct the optimal treatment regime (see main manuscript for explanation) to illustrate their ideal performance, which is not feasible in real data, whereas in the implementation of C-learning we did not use any a priori information on which variables are important. Under more reasonable sample sizes $n = 200, 400$, the impact of variable selection is less and C-learning outperforms other methods even when information this is not available in practice is used in the implementation of other methods.

For one simulation data set (n=200), Figure s1 plots the classification data points used

Table s2: Results for the first simulation scenario using 500 Monte Carlo data sets . $E\{Y^*(g^{opt})\} = 20$. $E(\widehat{g}^{opt})$ shows the Monte Carlo average and standard deviation of values $E\{Y^*(\widehat{g}^{opt})\}$ obtained using 10^6 Monte Carlo simulations for each data set.

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
Original BOWL	3.33(3.61)	4.71(3.27)	5.83(2.69)
BOWL	10.84(1.85)	12.13(1.54)	13.02(1.36)
Q-learning	12.49(1.83)	12.76(1.46)	13.05(1.14)
Zhang et al.(2013)	13.25(2.12)	15.08(1.46)	16.28(1.01)
C-learning	17.27(0.97)	18.52(0.74)	19.37(0.41)

for estimation in each stage for C-learning and BOWL. It provides some further insight on how the weighted classification in C-learning can facilitate estimation and on the difference between C-learning and BOWL. Note, in BOWL, the number of data points used for estimation decreases with stages.

For data application, in Figure s2 we report the classification data sets used for finding the optimal treatment regime. After obtaining estimated contrast functions for each subject using the AIPWE, the class label (indicated by color) for each subject is the sign of the estimated contrast and the weight (indicated by size) is the magnitude of the estimated contrast. Unlike the case in Figure s1, Figure s2 suggests that, for both stages, we cannot separate subjects to two classes and the optimal treatment decision does not seem to depend on patient characteristics available in our data.

References

Rush, A. J., Fava, M,, Wisniewski, S. R., Lavori, P.W., Trivedi, M.H., Sackeim, H. A.,

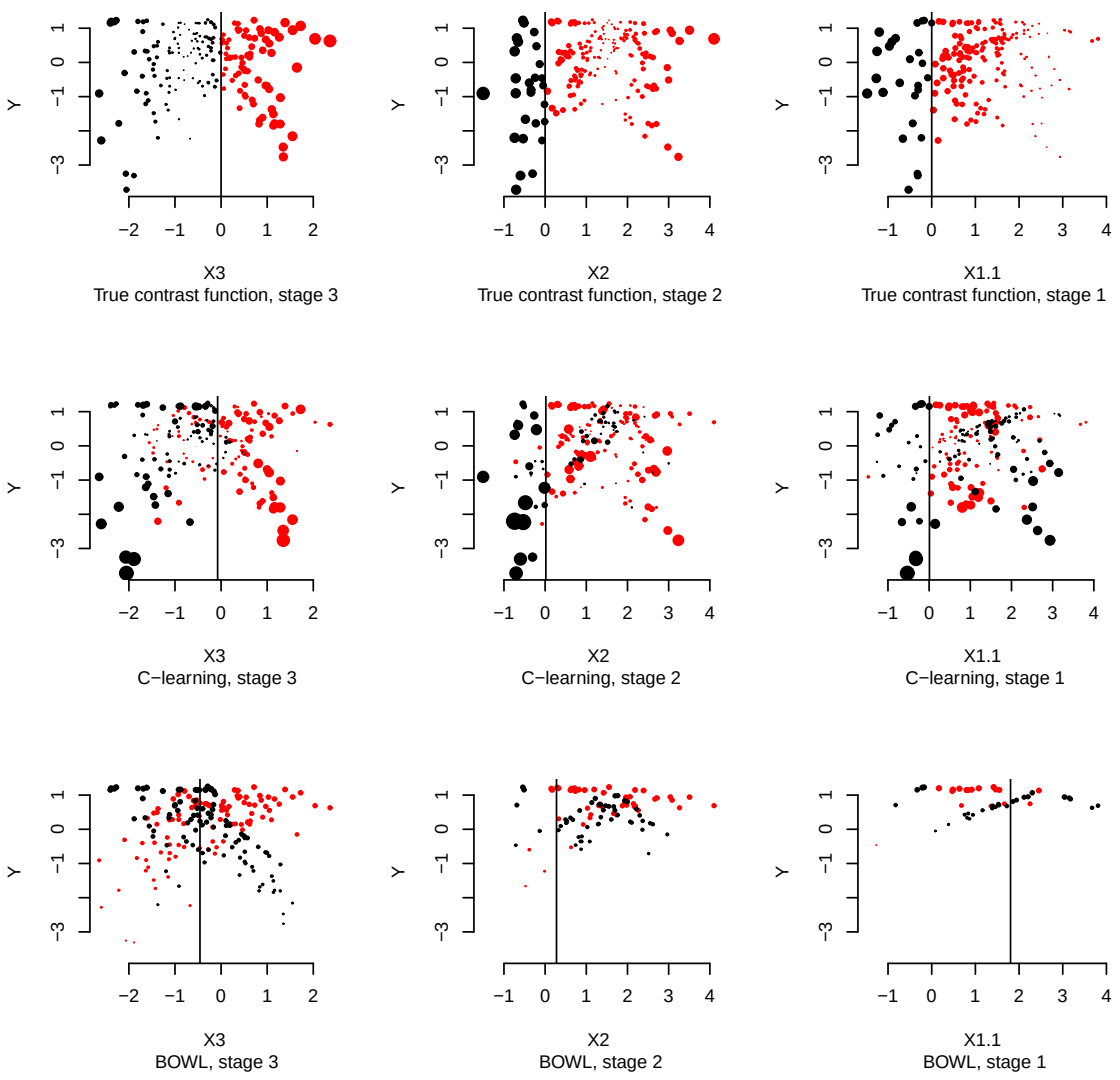

Figure s1: Classification data sets used for estimating the optimal decision rule at each stage in C-learning (second row) and BOWL (third row) for one simulated data set from simulation setting one (n=200). Each dot corresponds to a data point for a subject and the vertical bar is the true or estimated regime. The first row is the truth, with size indicating the magnitude of true treatment contrast and color indicating true optimal treatment. In row one and row two, only dots for those patients whose data are used in estimation in the corresponding stage are plotted, with size indicating weight (e.g., in C-learning the weight is the magnitude of estimated contrast functions) and color indicating class label (e.g., in C-learning, label is the sign of estimated contrast functions) in the corresponding weighted classification algorithm.

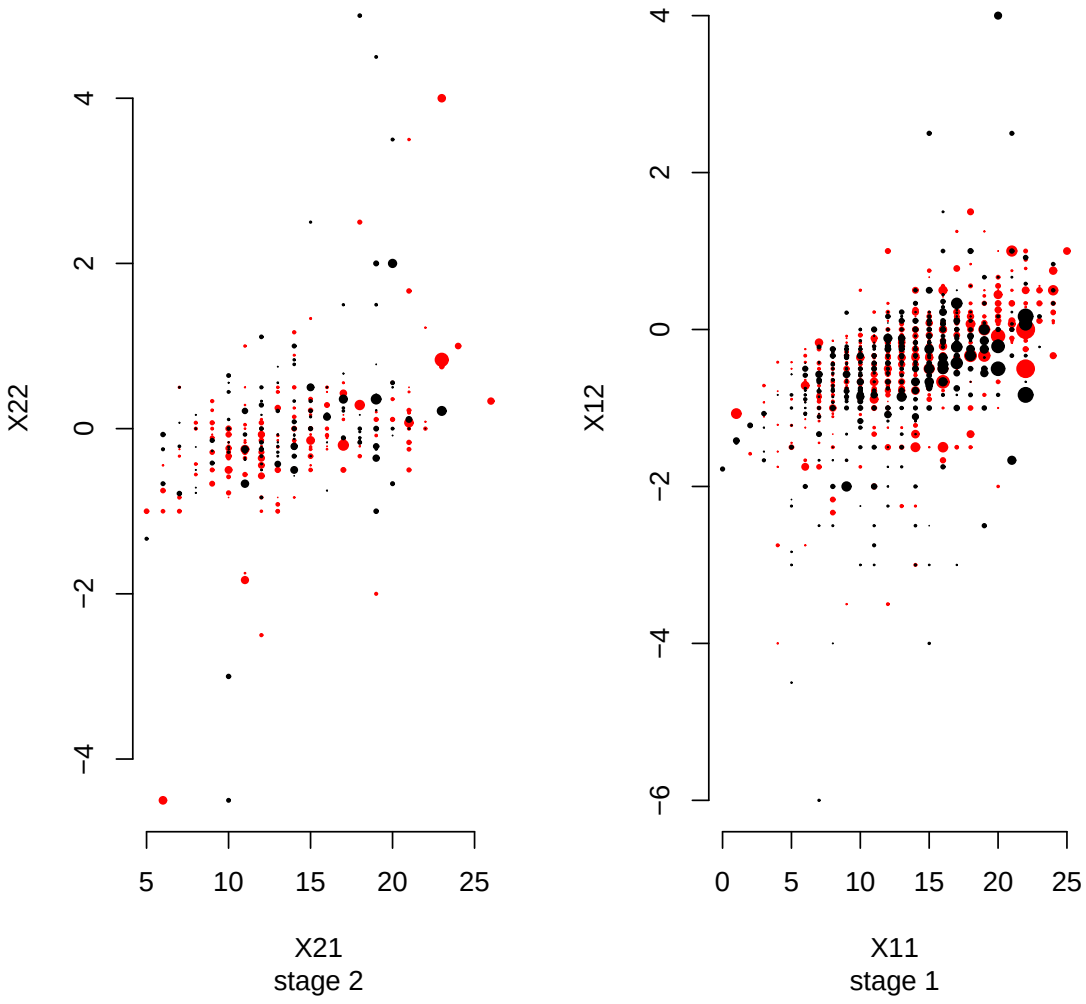

Figure s2: Classification data sets used for estimating the optimal decision rule at each stage in C-learning for data application. Each dot corresponds to a subject, with color indicating class label and size indicating weight.

Table s3: Results for n=100 using 500 Monte Carlo data sets for scenarios reported in the main manuscript . $E\{Y^*(g^{opt})\} = 20$. $E(\widehat{g}^{opt})$ shows the Monte Carlo average and standard deviation of values $E\{Y^*(\widehat{g}^{opt})\}$ obtained using 10^6 Monte Carlo simulations for each data set. Superscript"†" indicates that only relevant variables among the high dimensional set of covariates are used to construct the optimal treatment regime, which is not feasible in reality. Methods without "†" are searching the optimal treatment regimes without any a priori information on which variables are important.

	Scenario 1	Scenario 2	Scenario 3
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	8.77 (2.42)	2.41(1.9)	2.23 (2.08)
BOWL†	-	13.55 (2.25)	11.5 (2.18)
Q-learning	11.78 (2.14)	-	-
Q-learning†	-	13.94 (1.27)	12.92 (0.76)
Zhang et al.(2013)	11.12 (2.69)	-	-
Zhang et al.(2013)†		16.46 (2.41)	15.82 (1.74)
C-learning-Q	15.43 (1.68)	13.24 (2.89)	14.41 (2.57)
C-learning-RF	-	10.03 (3.21)	11.82 (3.29)

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.