Biometrics 74, 891–899
September 2018

DOI: 10.1111/biom.12836

C-Learning: A New Classification Framework to Estimate Optimal Dynamic Treatment Regimes

Baqun Zhang[1] **and Min Zhang**[2,*]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P.R.China.
[2]Department of Biostatistics, University of Michigan, Ann Arbor, U.S.A.
*email: mzhangst@umich.edu

Summary. A dynamic treatment regime is a sequence of decision rules, each corresponding to a decision point, that determine that next treatment based on each individual's own available characteristics and treatment history up to that point. We show that identifying the optimal dynamic treatment regime can be recast as a sequential optimization problem and propose a direct sequential optimization method to estimate the optimal treatment regimes. In particular, at each decision point, the optimization is equivalent to sequentially minimizing a weighted expected misclassification error. Based on this classification perspective, we propose a powerful and flexible C-learning algorithm to learn the optimal dynamic treatment regimes backward sequentially from the last stage until the first stage. C-learning is a direct optimization method that directly targets optimizing decision rules by exploiting powerful optimization/classification techniques and it allows incorporation of patient's characteristics and treatment history to improve performance, hence enjoying advantages of both the traditional outcome regression-based methods (Q- and A-learning) and the more recent direct optimization methods. The superior performance and flexibility of the proposed methods are illustrated through extensive simulation studies.

Key words: A-learning; Augmented inverse probability weighted estimator; CART; Dynamic treatment regime; Precision medicine; Q-learning.

1. Introduction

Treatment of patients may involve a series of decisions and it is important that decisions are adaptive with time-dependent information on patients over time. A dynamic treatment regime is a sequence of decision rules that determine the next treatment for a patient based on his/her own available information up to that time (Murphy, 2003; Robins, 2004) and has received much attention lately (Moodie et al., 2007; Song et al., 2011; Zhang et al., 2012ab, 2013; Zhao et al., 2012, 2015; Geng et al., 2015; Wallace and Moodie, 2015). It explicitly takes into account patient heterogeneity and the evolving nature of a disease. The goal is to identify the optimal set of decision rules that, if followed by the entire patient population, would yield the most favorable outcome on average.

Two common approaches to estimate the optimal dynamic treatment regime are Q- and A-learning (Watkins and Dayan, 1992; Murphy, 2003; Robins, 2004). Both approaches involve modeling the outcome and then the optimal treatment regime is identified by inverting the relationship between outcome, patient information, and treatment. Q- and A-learning work well under good regression models for outcomes. However, if the regression models are misspecified the estimated regime may be far from optimal. This is due to the fact that there is a mismatch between the target of outcome regression-based methods and the goal of learning the optimal treatment regime, as firstly pointed out by Murphy (2005). Outcome regression-based methods target good models for the outcome instead of optimizing decision rules to yield the maximum expected potential outcomes.

More recent efforts have been made to mitigate the concern of outcome model misspecification and several approaches have been proposed to directly optimize population expected outcomes across regimes. The advantage of direct optimization has been discussed in detail in literature mentioned below; see also Kang et al. (2014) and discussion articles. The direct optimization approach includes the work of Zhang et al. (2012a, 2013), outcome weighted learning of Zhao et al. (2012, 2015), and residual weighted learning of Zhou et al. (2015). These methods essentially directly estimate the population mean outcome under a regime using doubly robust augmented inverse probability weighted estimators (AIPWE) or simple inverse probability weighted estimator (IPWE). One other relevant work is Tian et al. (2014), which proposes a robust method for estimating interactions of treatment and a large number of covariates, with applications in estimating the optimal treatment regimes.

For the single decision point setting, Zhang et al. (2012b) proposed a general framework within which identifying the optimal treatment regime is equivalent to minimizing a weighted misclassification error, weighted by the contrast in outcome regression between treatments. It allows one to take advantage of existing powerful classification techniques. Equally importantly, this framework allows the optimization step for optimizing decision rules to be decoupled from modeling outcomes, alleviating the mismatch issue pointed out by Murphy (2005). We propose to extend the classification framework to the multiple decision point setting, which requires important methodological developments. The proposed method is a direct optimization method, where the optimization can be viewed as a classification problem. In addition, it allows for incorporating information from outcome regression models to improve efficiency, hence enjoying advantages of both types of approaches.

2. Notation and Dynamic Treatment Regimes

Consider a multistage decision problem where decisions are made at K decision points. We denote the decision at stage k as a_k, with $a_k \in \{0, 1\}$, the treatment actually received at stage k as A_k, and the covariate information observed between decision $k-1$ and k as X_k. Treatment history up to and including the kth decision is denoted as $\bar{a}_k = (a_1, \ldots, a_k)$, and similarly we can define the observed treatment history $\bar{A}_k$ and the observed covariate history $\bar{X}_k$. The overall outcome of interest is $Y \in \mathcal{R}$, which can be a function of intermediate information collected across all K decisions or a measurement ascertained after the Kth decision. Without loss of generality suppose a larger value of outcome is preferred.

A dynamic treatment regime is a sequence of decision rules, $g = (g_1, \ldots, g_K)$, that determine how to treat a patient over time. The kth decision rule $g_k(\bar{x}_k, \bar{a}_{k-1})$, denoted as $g_k \in \mathcal{G}_k$, assigns a treatment for a subject based on his/her covariate and treatment history up to decision k. The potential outcome associated with any regime g is denoted as $Y^*(g)$, that is, the outcome that would result if the subject followed g. The optimal treatment regime $g^{opt} = (g_1^{opt}, \ldots, g_K^{opt}) \in \mathcal{G}$ is the one that would yield the maximum expected outcome if were followed by all patients in the population. That is, g^{opt} satisfies $E\{Y^*(g^{opt})\} \geq E\{Y^*(g)\}$ for all $g \in \mathcal{G}$. We make some standard assumptions that make g^{opt} identifiable from the observed data (Schulte et al., 2014). That is, we assume the consistency assumption, the stable unit treatment value assumption, and the no unmeasured confounders assumption. Under these assumptions, g^{opt} can be expressed in terms of the observed data via backward induction. Defining $Q_K(\bar{x}_K, \bar{a}_K) = E(Y|\bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$, referred to as Q-functions with "Q" for "quality," the optimal decision rule at the K-th decision point satisfies $g_K^{opt}(\bar{x}_K, \bar{a}_{K-1}) = \arg\max_{a_K \in \{0,1\}} Q_K(\bar{x}_K, \bar{a}_{K-1}, a_K)$. Recursively we can define the value function (V-function) as $V_k(\bar{x}_k, \bar{a}_{k-1}) = \max_{a_k \in \{0,1\}} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$ for $k = K, \ldots, 2$, with $\bar{a}_0$ being null, and Q-functions as $Q_k(\bar{x}_k, \bar{a}_k) = E\{V_{k+1}(\bar{x}_k, X_{k+1}, \bar{a}_k)|\bar{X}_k = \bar{x}_k, \bar{A}_k = \bar{a}_k\}$ for $k = K-1, \ldots, 1$. The optimal decision rule at the k-th point satisfies $g_k^{opt}(\bar{x}_k, \bar{a}_{k-1}) = \arg\max_{a_k \in \{0,1\}} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$. Supplementary Material A provides more background.

3. C-Learning

3.1. *Main Results*

To provide some intuition first consider the single decision point setting ($K = 1$), for which Zhang et al. (2012b) proposed a general framework for estimating the optimal regime from a classification perspective. We omit the subscript denoting stage below. Recall the Q-function is defined as $Q(x, a) = E(Y|X = x, A = a)$ and define a contrast function $C(x) = Q(x, 1) - Q(x, 0)$, which is the difference in expected potential outcomes for a subject with covariate x were she/he to receive treatment 1 versus 0. Zhang et al. (2012b) show that g^{opt} minimizes an expected weighted misclassification error; that is,

$$g^{opt} = \arg\min_{g \in \mathcal{G}} E[|C(X)|I\{g(X) \neq Z\}], \text{ where } Z = I\{C(X) > 0\}. \tag{1}$$

This allows one to recast the problem of estimating the optimal treatment regime as a weighted classification problem. Consider viewing each subject as belonging to one of the two (latent) classes defined by $Z = I\{C(X) > 0\}$, where class $Z = a$ compose those subjects who would benefit from treatment a and therefore should be treated with treatment a. If $g(X) = I\{C(X) > 0\}$, a correct treatment decision is made and there is no loss incurred. However, if $g(X) \neq I\{C(X) > 0\}$, the decision is not optimal and the corresponding loss is $W = |C(X)|$; that is, the larger the difference in expected potential outcomes between two treatment options, the larger the loss. As it only involves patient characteristics (covariates) and the true treatment contrast but not the observed treatment assignment, (1) can be viewed as an alternative definition of the optimal treatment regime.

In this article, we provide an alternative definition of the optimal dynamic treatment regime in the multiple decision point setting from the classification perspective and, based on this perspective, propose a new and powerful statistical learning method. We term our approach as C-learning, where "C" stands for classification. As in the single decision point setting, we define a contrast function for each decision point; that is, for stage k, $k = 1, \ldots, K$, the contrast function is defined as $C_k(\bar{x}_k, \bar{a}_{k-1}) = Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k = 1) - Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k = 0)$, where $Q_k(\bar{x}_k, \bar{a}_k)$ are defined recursively in Section 2. The contrast function at stage k represents the difference in expected potential outcomes between treatment option 1 and 0 at stage k, assuming that optimal decisions are made in the future. To simplify notation, we define $L_k \equiv (\bar{X}_k, \bar{A}_{k-1})$, which is the covariate and treatment history available at decision point k. We discuss how one can embed the classification approach in backward induction to find the optimal dynamic treatment regime. The key lies in the following Theorem 1 and Proposition 1, proofs of which are given in Web Supplementary Materials B and C.

Theorem 1. *Let $g^* = (g_1^*, \ldots, g_K^*)$, be a treatment regime that satisfies*

$$g_k^*(L_k) = \arg\min_{g_k \in \mathcal{G}_k} E[|C_k(L_k)|I\{g_k(L_k) \neq Z_k\}],$$
$$\textit{where } Z_k = I\{C_k(L_k) > 0\}$$

$k = K, \ldots, 1$, then g^* is the optimal dynamic treatment regime.

Theorem 1 states that the optimal treatment decision rule at each stage minimizes an objective function that can be interpreted as a weighted misclassification error, where the goal of classification is to classify subjects at each stage to one of two latent classes, denoted by $Z_k = I\{C_k(L_k) > 0\}$, for whom the optimal decision at the stage is 0 and 1, respectively. That is, class $Z_k = 1$ include subjects for whom treatment $a_k = 1$ leads to a larger expected potential outcome than decision 0, given that optimal decisions are made in the future. If $g_k(L_k)$ is not the optimal decision at stage k, that is, $g_k(L_k) \neq Z_k$, then the loss incurred is $|C_k(L_k)|$; otherwise, the loss is zero. Theorem 1 is a general result that recasts the problem of identifying the optimal dynamic treatment regime into a meaningful sequential classification problem. We note that classification technique is used in the backward outcome weighted learning (BOWL) of Zhao et al. (2015) to sequentially estimate the optimal treatment regime. Our result differs from that in two important ways. First, BOWL is based on the particular IPWE estimator of $E\{Y^*(g)\}$ and the use of classification techniques is possible because of the form of the IPWE estimator, whereas the classification perspective of Theorem 1 is a general result that holds regardless how one estimates $E\{Y^*(g)\}$ or $C_k(L_k)$. For simplicity taking $K = 1$, BOWL essentially estimates $E\{Y^*(g)\}$ by the IPWE estimator, $\sum_{i=1}^{n}[Y_i I\{g(X_i) = A_i\}/\pi(A_i, X_i)]$, where $\pi(a, X) = Pr(A = a|X)$, and then maximizes it across a class of regimes, which is equivalent to minimizing $\sum_{i=1}^{n}[Y_i I\{g(X_i) \neq A_i\}/\pi(A_i, X_i)]$. 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 to classify patients to $A = 0$ or 1, and $Y/\pi(A, X)$ can be viewed as the weight if Y is positive. The classification idea of BOWL cannot easily generalize to other estimator of $E\{Y^*(g)\}$, whereas based on Theorem 1 one can transform the problem into a weighted classification problem using any estimators of $E\{Y^*(g)\}$ (or equivalently $C_k(L_k)$), say IPWE, AIPWE, regression estimator (see Zhang et al., 2012b for discussion in the $K = 1$ setting). Second, the interpretation of classification is different, which has important implications on the performance of the resulting learning method as demonstrated by simulation studies. In BOWL as well as other OWL-based methods, if $g(X_i) = A_i$ then no loss is incurred and a misclassification loss is incurred if $g(X_i) \neq A_i$; that is, this classification aims to classify patients to classes that are defined by the actually received treatment A. Due to this classification perspective, the estimated classifier (treatment regime) tries to minimize the weighted misclassification error by keeping treatment assignments that subjects actually received, which is an issue of OWL-based methods as pointed out by Zhou et al. (2015). In our classification perspective, a loss is incurred if $g(X) \neq I\{C(X) > 0\}$; that is, the classifier aims to classify patients to classes corresponding to the optimal treatment decisions. The interpretation of this classification corresponds exactly to the intuitive meaning of optimizing individual treatment decisions and the resulting method does not suffer from the same issue as BOWL. In Supplementary Material D, we provide a more comprehensive discussion on these issues.

Proposition 1. *The value functions defined recursively in Section 2 satisfy the following condition:*

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

$k = K, \ldots, 1$, $V_{K+1} \equiv Y$, *where* g_k^{opt} *is the optimal decision rule at stage* k.

3.2. *Estimation Procedure*

Based on Theorem 1 and Proposition 1, we propose a flexible and powerful new learning method using backward induction. We start at the last decision point K. Then covariate and treatment history $(\bar{X}_K, \bar{A}_{K-1}) \equiv L_K$ before stage K can be regarded as baseline covariate vector and data can be rewritten as (Y, L_K, A_K). As in the single decision point setting, by separating the contrast function into two parts, with one part representing the magnitude and the other representing the sign, we show in the proof of Theorem 1 that equivalently the optimal treatment rule at K minimizes a weighted misclassification error; that is,

$$g_K^{opt} = \arg\min_{g_K \in \mathcal{G}_K} E[|C_K(L_K)| I\{g_K(L_K) \neq Z_K\}]. \tag{2}$$

Therefore, g_K^{opt} can be estimated by

$$\widehat{g}_{C,K}^{opt} = \arg\min_{g_K \in \mathcal{G}_K} \sum_{i=1}^{n}[\widehat{W}_{Ki} I\{g_K(L_{Ki}) \neq \widehat{Z}_{Ki}\}],$$

where $\widehat{Z}_{Ki} = I\{\widehat{C}_K(L_{Ki}) > 0\}$, $\widehat{W}_{Ki} = |\widehat{C}_K(L_{Ki})|$, and $\widehat{C}_K(L_{Ki})$ is an estimator of $C_K(L_{Ki})$. The contrast function can be estimated using various ways as discussed in Zhang et al. (2012b) and the doubly robust AIPWE method has superior performance relative to other methods. Therefore, we recommend estimating $C_K(L_{Ki})$ by the AIPWE estimator

$$\begin{aligned}\widehat{C}_K(L_{Ki}) = {} & \frac{A_{Ki}}{\widehat{\pi}_K(L_{Ki})} Y_i - \frac{A_{Ki} - \widehat{\pi}_K(L_{Ki})}{\widehat{\pi}_K(L_{Ki})} \widehat{Q}_K(L_{Ki}, 1) \\ & - \left\{ \frac{1 - A_{Ki}}{1 - \widehat{\pi}_K(L_{Ki})} Y_i + \frac{A_{Ki} - \widehat{\pi}_K(L_{Ki})}{1 - \widehat{\pi}_K(L_{Ki})} \widehat{Q}_K(L_{Ki}, 0) \right\},\end{aligned} \tag{3}$$

where $\widehat{\pi}_K(L_{Ki})$ is the estimated probability (propensity score) of receiving treatment $A_K = 1$ at stage K conditional on covariate and treatment history L_K using, for example, a logistic regression model; and $\widehat{Q}_K(L_{Ki}, A_K = a_K), a_K = 0, 1$, are estimates based on parametric or nonparametric models for $E(Y|L_K)$, further discussed in Section 4. From the proof for Theorem 1 and discussion in Zhang et al. (2012b), essentially this is equivalent to firstly estimating $E\{Y^*(\bar{A}_{K-1}, g_K)\}$ by the AIPWE estimator and then optimizing AIPWE across a class of regimes. We acknowledge that other estimators of contrast functions can also be used within this framework; for example, one can directly estimate $C_K(L_{Ki})$ by the difference in Q-functions. The minimization can be viewed as a typical classification problem with $\widehat{Z}_K$ as the binary "response," L_K the "predictor," $\widehat{W}_K$ the "weight," and g_K the "classification rule." In simulation studies in Section 4, we show various ways to implement this optimization step. We denote the estimated regime as $\widehat{g}_{C,K}^{opt}$.

After obtaining $\widehat{g}_{C,K}^{opt}$, C-learning moves backward sequentially until the first stage to estimate the optimal decision rule at stage k, $k = K - 1, \ldots, 1$. By Theorem 1, the optimal decision rule at stage k satisfies

$$g_k^{opt} = \arg\min_{g_k \in \mathcal{G}_k} E[|C_k(L_k)| I\{g_k(L_k) \neq Z_k\}], \tag{4}$$

where $C_k(L_k) = Q_k(L_k, 1) - Q_k(L_k, 0)$ is the contrast function at stage k. Therefore, if one can estimate $C_k(L_k)$ or equivalently $Q_k(L_k, a_k)$, then we can proceed similarly as in stage K. Recall that $Q_k(L_k, a_k) = E\{V_{k+1}(L_{k+1})|L_k, a_k\}$, and if $V_{k+1}(L_{k+1})$ is available, one can estimate $Q_k(L_k, a_k)$ by treating $V_{k+1}(L_{k+1})$ as the response. However, except for the last stage, $V_{k+1}(L_{k+1})$ is not directly observable and has to be estimated. By Proposition 1, $V_k(L_{ki})$ can be estimated recursively by

$$\widetilde{V}_{ki} \equiv \widetilde{V}_k(L_{ki}) = \widetilde{V}_{(k+1)i} + \{\widehat{Q}_k(L_{ki}, 1) - \widehat{Q}_k(L_{ki}, 0)\}\{\widehat{g}^{opt}_{C,k}(L_{ki}) - A_{ki}\}, \qquad (5)$$

for $k = K, K-1, \ldots, 2$, and $\widetilde{V}_{(K+1)i} \equiv Y_i$. Then one can estimate $Q_k(L_k, a_k)$ and the contrast function $C_k(L_k)$ based on "optimal responses" $\widetilde{V}_{(k+1)i}$, as discussed below. This strategy is similar in spirit to the contrast-based A-learning (Schulte, 2014). For example, after we obtain $\widehat{g}^{opt}_{C,K}$, the value function $V_K(L_{Ki}), i = 1, \ldots, n$, can be estimated by

$$\widetilde{V}_{Ki} \equiv \widetilde{V}_K(L_{Ki}) = Y_i + \{\widehat{Q}_K(L_{Ki}, 1) - \widehat{Q}_K(L_{Ki}, 0)\}\{\widehat{g}^{opt}_{C,K}(L_{Ki}) - A_{Ki}\},$$

which is Y_i if the estimated optimal treatment at K is the same as the actual received treatment A_{Ki} and is Y_i plus the absolute difference in expected potential outcomes if A_{Ki} is not the estimated optimal treatment option.

Similar to stage K, recursively at stage $k, k = K-1, \ldots, 1$, treating $(\widetilde{V}_{(k+1)i}, L_{ki}, A_{ki}), i = 1, \ldots, n$, as "data," where $L_{ki} = (\bar{X}_{ki}, \bar{A}_{(k-1)i})$ is regarded as the baseline covariate vector, $\widetilde{V}_{(k+1)i}$ as response, and A_{ki} as treatment, we estimate $C_k(L_{ki})$ by the AIPWE estimate

$$\widehat{C}_k(L_{ki}) = \frac{A_{ki}}{\widehat{\pi}_k(L_{ki})}\widetilde{V}_{(k+1)i} - \frac{A_{ki} - \widehat{\pi}_k(L_{ki})}{\widehat{\pi}_k(L_{ki})}\widehat{Q}_k(L_{ki}, 1) - \left\{\frac{1 - A_{ki}}{1 - \widehat{\pi}_k(L_{ki})}\widetilde{V}_{(k+1)i} + \frac{A_{ki} - \widehat{\pi}_k(L_{ki})}{1 - \widehat{\pi}_k(L_{ki})}\widehat{Q}_k(L_{ki}, 0)\right\}, \qquad (6)$$

where $\widehat{\pi}_k(L_{ki})$ are estimated propensity score $P(A_{ki} = 1|L_{ki})$ based on, say, a logistic regression model, and $\widehat{Q}_k(L_{ki}, a_k), a_k = 0, 1$, are estimates of $Q_k(L_{ki}, a_k) = E\{V_{(k+1)i}|L_{ki}, A_{ki} = a_k\}$ based on parametric or nonparametric models. The main difference from stage K is that here the estimated value function $\widetilde{V}_{(k+1)i}$ plays the role of Y_i as in the Kth decision point. We then obtain the corresponding $\widehat{Z}_{ki} = I\{\widehat{C}_k(L_{ki}) > 0\}$ and $\widehat{W}_{ki} = |\widehat{C}_k(L_{ki})|$ and, according to (4), $g^{opt}_k(L_k)$ can be estimated by

$$\widehat{g}^{opt}_{C,k} = \arg\min_{g_k \in \mathcal{G}_k} \sum_{i=1}^{n} \widehat{W}_{ki} I\{\widehat{Z}_{ki} \neq g_k(L_{ki})\} \qquad (7)$$

using some classification or optimization technique. The final estimated optimal regime is $\widehat{g}^{opt}_C = (\widehat{g}^{opt}_{C,1}, \ldots, \widehat{g}^{opt}_{C,K})$. The steps for implementing C-learning are summarized in Web Supplementary Material E.

In Web Supplementary Material D, we discuss in detail the connection with existing methods and clarify the theoretical advantages and differences of the proposed method. Here, we summarize the main points. As the method of Zhang et al. (2013), C-learning is a doubly robust AIPWE-based, direct optimization method and enjoys more protection against model misspecification. This is in contrast with outcome regression-based methods (e.g., Q- and A-learning), where outcome regression models directly determine estimated optimal treatment regimes and as a result their performance heavily depends on correct specification of the model for Q-functions or contrast functions. Instead direct optimization methods aim to directly optimize estimate of $E\{Y^*(g)\}$ (simultaneously or sequentially) across a class of regimes and are robust as long as $E\{Y^*(g)\}$ is consistently estimated. From our proof for Theorem 1, it is clear that at each stage C-learning essentially optimizes AIPWE estimate of expected potential outcomes across a class of regimes and AIPWE is known to have the double robustness property, that is, is robust when either propensity score models or outcome regression models, but not necessarily both, are correctly specified. When treatment is randomized as in a sequentially randomized study, the propensity score models are always correct and AIPWE estimates of expectations of potential outcomes are consistent, leading to robust estimate of treatment regimes even when outcome regression models are misspecified. Outcome regression models, even misspecified, are useful for improving efficiency of estimates and lead to improved performance of estimated regimes, especially when covariates are strongly predictive of outcomes. See Zhang et al. (2012ab, 2013) for more discussion on robustness under model misspecification.

C-learning shares similar robustness property as Zhang et al. (2013) and enjoys additional appealing features. First, C-learning transforms the problem into a sequential classification problem, which has several advantages. For example, modern powerful and flexible classification algorithms can be used and optimization can be carried out among a much larger class of regimes (e.g., decision trees), whereas in Zhang et al. (2013) the optimization of regimes is carried out among a restricted class of regimes indexed by a finite number of parameters. Second, unlike Zhang et al. (2013) that uses simultaneous optimization across stages, the proposed method uses sequential optimization which leads to considerable improvement in performance. The method of Zhang et al. (2013) is based on an AIPWE estimator of $E\{Y^*(g)\}$ for monotone coarsened (missing) data. In the missing data perspective, the potential outcome of a subject is observed only if the observed treatments at all stages are consistent with a regime as regimes at all stages are estimated simultaneously. In C-learning, however, at stage K, 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 since 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 and, intuitively, in C-learning the best effort can be made to only estimate the optimal regime at that stage. In addition, we note that one has to optimize across a large number of parameters for parameterizing the whole dynamic treatment regime in the simultaneous optimization, whereas in C-learning at each stage the optimization is among a smaller number of parameters relevant only to that stage. This difference in handling multiple stages leads to big improvement in performance for C-learning. Finally, as illustrated in our second simulation scenario, C-learning accommodates variable selection targeting selection of prescriptive variables, that is, variables relevant for decision making.

Compared with BOWL, unlike the proposed method that uses AIPWE, BOWL is based on IPWE which does not incorporate outcome regression model and is less efficient and less robust (lack of the double-robustness property). Second, as discussed below Theorem 1, our classification perspective is fundamentally different from BOWL and leads to considerable improvement in performance. Lastly, to achieve sequential estimation, BOWL must reduce sample size geometrically at later stages (see Figure S1). However, the proposed method is able to use all subjects in estimation at all stages and leads to much better performance.

4. Simulation Studies

We report results on simulation studies under scenarios imitating a multi-stage randomized trial with $K = 3$ for sample size n=200, 400, and 800 using 500 Monte Carlo replicates.

4.1. *Data Generation and Methods Implementation*

The first setting was adopted from Zhao et al. (2015). Treatments A_1, A_2, and A_3 are randomly generated from $\{1, 0\}$ with equal probability. Baseline covariates $X_{1,1}, X_{1,2}, X_{1,3}$ are generated from $N(45, 15^2)$, X_2 is generated according to $X_2 \sim N(1.5X_{1,1}, 10^2)$, and X_3 is generated according to $X_3 \sim N(0.5X_2, 10^2)$. Outcomes are generated as $Y = \mu(\bar{A}_3, \bar{X}_3) + \epsilon$ for ϵ standard normal and $\mu(\bar{A}_3, \bar{X}_3) = 20 - |0.6X_{1,1} - 40|(A_1 - g_1^{opt})^2 - |0.8X_2 - 60|(A_2 - g_2^{opt})^2 - |1.4X_3 - 40|(A_3 - g_3^{opt})^2$, where $g_1^{opt} = I(X_{1,1} - 30 > 0)$, $g_2^{opt} = I(X_2 - 40 > 0)$, and $g_3^{opt} = I(X_3 - 40 > 0)$. The optimal treatment regime is $g^{opt} = (g_1^{opt}, g_2^{opt}, g_3^{opt})$ and $E\{Y^*(g^{opt})\} = 20$.

For Q-learning, we posited Q-functions

$$\begin{aligned}
Q_3(\bar{x}_3, \bar{a}_3; \beta_3) &= \beta_{3,0} + \beta_{3,1}x_{1,1} + \beta_{3,2}x_{1,2} + \beta_{3,3}x_{1,3} \\
&\quad + a_1(\beta_{3,4} + \beta_{3,5}x_{1,1}) + \beta_{3,6}x_2 + a_2(\beta_{3,7} + \beta_{3,8}x_2) \\
&\quad + \beta_{3,9}x_3 + a_3(\beta_{3,10} + \beta_{3,11}x_3), \\
Q_2(\bar{x}_2, \bar{a}_2; \beta_2) &= \beta_{2,0} + \beta_{2,1}x_{1,1} + \beta_{2,2}x_{1,2} + \beta_{2,3}x_{1,3} \\
&\quad + a_1(\beta_{2,4} + \beta_{2,5}x_{1,1}) + \beta_{2,6}x_2 + a_2(\beta_{2,7} + \beta_{2,8}x_2), \\
Q_1(x_1, a_1; \beta_1) &= \beta_{1,0} + \beta_{1,1}x_{1,1} + \beta_{1,2}x_{1,2} + \beta_{1,3}x_{1,3} \\
&\quad + a_1(\beta_{1,4} + \beta_{1,5}x_{1,1}).
\end{aligned}$$

For the AIPWE-based method of Zhang et al. (2013), we took $\mathcal{G}_\eta$ to have elements $g_\eta = (g_{\eta_1}, g_{\eta_2}, g_{\eta_3})$, where $g_{\eta_3}(\bar{x}_3, \bar{a}_2) = I(\eta_{3,0} + \eta_{3,1}x_{1,1} + \eta_{3,2}x_{1,2} + \eta_{3,3}x_{1,3} + \eta_{3,4}x_2 + \eta_{3,5}x_3 > 0)$, $g_{\eta_2}(\bar{x}_2, a_1) = I(\eta_{2,0} + \eta_{2,1}x_{1,1} + \eta_{2,2}x_{1,2} + \eta_{2,3}x_{1,3} + \eta_{2,4}x_2 > 0)$, $g_{\eta_1}(x_1) = I(\eta_{1,0} + \eta_{1,1}x_{1,1} + \eta_{1,2}x_{1,2} + \eta_{1,3}x_{1,3} > 0)$. Clearly, $g^{opt} \in \mathcal{G}_\eta$ and all available covariates at each stage were considered in parameterizing the treatment regime. In BOWL and C-learning, we estimated $\pi_k(L_k)$ by $\hat{\pi}_k(L_k) = \sum_{i=1}^n A_{ki}/n, k = 1, 2, 3$. For C-learning and method of Zhang et al. (2013), one also needs to specify model for the outcome and we used the same Q-function models as in Q-learning. To carry out minimization in C-learning, we used a genetic algorithm discussed by Goldberg (1989), implemented in the `rgenoud` package in R (Mebane and Sekhon, 2011).

In the second set of simulations, we increased the dimension of covariates to 50. At baseline, 40 covariates $X_{1,1}, \ldots, X_{1,40}$ are generated from $N(45, 15^2)$. At stage 2, $X_{2,j}$ is generated according to $X_{2,j} \sim N(1.5X_{1,j}, 10^2)$, $j = 1, \ldots, 5$. At stage 3, $X_{3,j}$ is generated according to $X_{3,j} \sim N(0.5X_{2,j}, 10^2)$, $j = 1, \ldots, 5$. The outcome was generated as $Y = \mu(\bar{A}_3, \bar{X}_3) + \epsilon$ for ϵ standard normal and $\mu(\bar{A}_3, \bar{X}_3) = 20 - |0.6X_{1,1} - 40|(A_1 - g_1^{opt})^2 - |0.8X_{2,1} - 60|(A_2 - g_2^{opt})^2 - |1.4X_{3,1} - 40|(A_3 - g_3^{opt})^2$, where $g_1^{opt} = I(X_{1,1} - X_{1,2} > 0)$, $g_2^{opt} = I(X_{2,1} - X_{2,2} > 0)$, $g_3^{opt} = I(X_{3,1} - X_{3,2} > 0)$. This scenario is similar to scenario 1, but we further made the optimal decision rule at each stage depends on a linear combination of two covariates.

For Q-learning, we posited Q-functions

$$\begin{aligned}
Q_3(\bar{x}_3, \bar{a}_3; \beta_3) &= \beta_{3,0} + \beta_{3,1}x_{1,1} + \beta_{3,2}x_{1,2} \\
&\quad + a_1(\beta_{3,3} + \beta_{3,4}x_{1,1} + \beta_{3,5}x_{1,2}) + \beta_{3,6}x_{2,1} + \beta_{3,7}x_{2,2} \\
&\quad + a_2(\beta_{3,8} + \beta_{3,9}x_{2,1} + \beta_{3,10}x_{2,2}) + \beta_{3,11}x_{3,1} \\
&\quad + \beta_{3,12}x_{3,2} + a_3(\beta_{3,13} + \beta_{3,14}x_{3,1} + \beta_{3,15}x_{3,2}), \\
Q_2(\bar{x}_2, \bar{a}_2; \beta_2) &= \beta_{2,0} + \beta_{2,1}x_{1,1} + \beta_{2,2}x_{1,2} + a_1(\beta_{2,3} + \beta_{2,4}x_{1,1} \\
&\quad + \beta_{2,5}x_{1,2}) + \beta_{2,6}x_{2,1} + \beta_{2,7}x_{2,2} \\
&\quad + a_2(\beta_{2,8} + \beta_{2,9}x_{2,1} + \beta_{2,10}x_{2,2}), \\
Q_1(x_1, a_1; \beta_1) &= \beta_{1,0} + \beta_{1,1}x_{1,1} + \beta_{1,2}x_{1,2} \\
&\quad + a_1(\beta_{1,3} + \beta_{1,4}x_{1,1} + \beta_{1,5}x_{1,2}).
\end{aligned}$$

Note, these model specifications favor the Q-learning method in that they only include the correct interaction of treatment and covariate and main effects of important covariates, leaving out those unimportant interaction terms and main effect terms, although the Q-functions are still misspecified. For the method of Zhang et al. (2013), we took $\mathcal{G}_\eta$ to have elements $g_\eta = (g_{\eta_1}, g_{\eta_2}, g_{\eta_3})$, where $g_{\eta_3}(\bar{x}_3, \bar{a}_2) = I(\eta_{3,0} + \eta_{3,1}x_{3,1} + \eta_{3,2} x_{3,2} > 0)$, $g_{\eta_2}(\bar{x}_2, a_1) = I(\eta_{2,0} + \eta_{2,1}x_{2,1} + \eta_{2,2}x_{2,2} > 0)$, $g_{\eta_1}(x_1) = I(\eta_{1,0} + \eta_{1,1}x_{1,1} + \eta_{1,2}x_{1,2} > 0)$. Clearly, $g^{opt} \in \mathcal{G}_\eta$. Similarly for BOWL, in one implementation we considered only important variables in searching for the optimal regimes and considered all variables in the other implementation. Of course, in a real application, it is difficult to pre-specify the right variables and forms for the true optimal regime and the results on these methods (marked by † in Tables 2–3) in the presence of high-dimensional covariates are too optimistic. We intend to illustrate their ideal performance in the presence of high-dimensional covariates for the purpose of comparing with the proposed method.

Unlike the other methods, in the implementation of the C-learning, we did not pre-specify the correct variables in the form of the treatment regime, but instead we use a data-driven way to choose the important covariates from the high-dimensional set of covariates. Therefore, the C-learning considers all linear decision rules constructed by the high-dimensional set of covariates, which is a much larger class than $\mathcal{G}_\eta$. Specifically, in the minimization step for each time point k, we used a forward selection algorithm to sequentially choose important covariates in forming the treatment regime, where the forward selection is on the basis of the proportion of reduction in the weighted misclassification error. Hence, the variable selection algorithm for the optimization step directly targets the goal of finding the optimal treatment regimes, in contrast to the model selection in the Q-learning method, where the selection targets the optimal model for the Q-functions. The details of the forward selection algorithm are given in this technical report (Zhang and Zhang, 2016). We implemented C-learning using two different ways that differ in how AIPWE is constructed: in C-learning-Q, we used a parametric model for the Q-functions and the parametric forms are the same as in the Q-learning method, and in C-learning-RF, we used random forest to nonparametrically model the Q-functions using the R function *random Forests* with default settings. In both ways, all linear decision rules constructed by the high-dimensional set of covariates are considered, as opposed to Q-learning, the method of Zhang et al. (2013) and BOWL in one implementation.

In the third set of simulations, the data generating scenario is the same as the second one except that $g_1^{opt} = I(X_{1,1} > 40)I(X_{1,2} < 60)$, $g_2^{opt} = I(X_{2,1} > 60)I(X_{2,2} < 90)$, and $g_3^{opt} = I(X_{3,1} > 30)I(X_{3,2} < 50)$ in $\mu(\bar{A}_3, \bar{X}_3)$. Here, the optimal decision rule at each stage is of the form of a tree. For the method in Zhang et al. (2013), we took $\mathcal{G}_\eta$ to have elements $g_\eta = (g_{\eta 1}, g_{\eta 2}, g_{\eta 3})$, where $g_{\eta k}(\bar{x}_k, \bar{a}_{k-1}) = I(X_{k,1} > \eta_{k,1})I(X_{k,2} < \eta_{k,2})$, $k = 1, 2, 3$. For C-learning, once we get the classification data set $(\widehat{Z}_{ki}, L_{ki}, \widehat{W}_{ki})$, we input this new data set into the CART algorithm to find the estimated optimal treatment regime among all tree decision rules constructed by the high-dimensional set of covariates. We used the `R` function `rpart` with default settings, except that we set the weights as the estimated weight $\widehat{W}$. Other methods are implemented as in the second set of simulations.

4.2. *Results and Discussion*

Results from scenarios 1–3 are shown in Tables 1–3, respectively. Table 1 shows that C-learning out-performs all other methods in this scenario. C-learning performs considerably better than Q-learning even though it used the same (misspecified) models for Q-functions in augmentation terms. This illustrates the advantage of AIPWE-based direct optimization methods over outcome regression-based methods as discussed in Section 3. It is also interesting to note that, although C-learning and the method of Zhang et al. (2013) are based on AIPWEs with the same propensity and augmentation term models and consider optimization across the same class of regimes, the performance of C-learning is still much better than that of Zhang et al. (2013). This is due to the difference in estimation across stages and the amount of information used in estimation; that is, Zhang et al. (2013) simultaneously estimates regimes at all stages and C-learning backward sequentially estimates the regime at each stage. C-learning has better performance than BOWL due to several reasons. For example, C-learning uses outcome regression models to improve efficiency, whereas BOWL is not able to incorporate outcome regression models. Also C-learning and BOWL differ in their way to handle multiple stages. C-learning is able to use information on all subjects at all stages. However, to sequentially estimate the regimes BOWL has to lose sample size geometrically with stages. For one simulation data set ($n = 200$), Figure S1 in Supplementary Material F plots the classification data points used 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.

Table 1

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	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
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)

Table 2 shows the performance of various methods when the dimension of covariates is relatively high. We comment that, in Table 2 as well as Table 3, performances of Q-learning†, BOWL†, and the method of Zhang et al. (2013)† are too optimistic due to the implementation and we should take this into account when comparing their performance with other methods and with results in Table 1. C-learning (both implementations) as well as BOWL consider all regimes constructed by linear combinations of the high-dimensional set of covariates, whereas methods marked with † only consider regimes constructed by relevant covariates, which is a much smaller class. This is because we try to give the best advantage to our comparison methods in implementation since their performances are dependent on the chosen parametric models or the class of regimes indexed by a finite number of parameters. Although our implementation unrealistically favors other methods by eliminating the burden for dealing with the high-dimensional set covariates, the performance of C-learning, combined with suitable variable selection algorithm in the

Table 2

Second simulation scenario (500 Monte Carlo data sets, $E\{Y^*(g^{opt})\} = 20$). *Superscript "†" indicates that only relevant variables among the high-dimensional set of covariates are used to construct the optimal treatment regime. Methods without "†" are searching the optimal treatment regimes without any a priori information on which variables are important.*

	$n = 200$	$n = 400$	$n = 800$
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	3.38(1.62)	5.93(1.37)	7.79(1.10)
BOWL†	14.76(1.74)	15.43(1.38)	15.74(1.12)
Q-learning†	14.01(1.05)	13.94(0.78)	13.78(0.56)
Zhang et al.(2013)†	17.98(1.42)	18.83(0.87)	19.35(0.45)
C-learning-Q	17.70(1.75)	19.45(0.61)	19.78(0.22)
C-learning-RF	16.59(2.14)	19.21(0.80)	19.75(0.14)

Table 3

Third simulation scenario (500 Monte Carlo data sets, $E\{Y^*(g^{opt})\} = 20$). Superscript "†" indicates that only relevant variables among the high-dimensional set of covariates are used to construct the optimal treatment regime. Methods without "†" are searching the optimal treatment regimes without any a priori information on which variables are important.

Estimator	$n = 200$ $E(\widehat{g}^{opt})$	$n = 400$ $E(\widehat{g}^{opt})$	$n = 800$ $E(\widehat{g}^{opt})$
BOWL	3.01(1.63)	5.02(1.42)	6.73(1.15)
BOWL†	12.55(1.28)	12.91(0.95)	13.12(0.72)
Q-learning†	13.12(0.45)	13.08(0.35)	13.07(0.23)
Zhang et al.(2013)†	17.02(1.25)	18.02(0.90)	18.71(0.63)
C-learning-Q	17.44(1.29)	18.91(0.73)	19.52(0.32)
C-learning-RF	16.94(1.48)	18.92(0.63)	19.61(0.24)

optimization step, is still considerably better than BOWL† and Q-learning† and is comparable to the method of Zhang et al. (2013)† when $n = 200$ and slightly better when $n = 400$, 800. BOWL, however, cannot handle high dimensionality well and has dramatically worse performance in this case. The C-learning framework can naturally accommodate variable selection methods targeted for optimal treatment regimes instead of prediction to improve performance in the presence of high-dimensional covariates. This (in addition to those discussed for Table 1) explains the dramatically better performance of C-learning than BOWL when they both consider the same class of regimes.

Table 3 shows the results when the true treatment regime is of the form of a decision tree and the dimension of covariates is relatively high. C-learning-RF, with both outcome regression models and important variables in the regimes chosen data-adaptively using existing off-the-shelf algorithms and softwares (Random Forest and CART), has superior performance and is comparable to C-learning-Q, where the Q-functions are modeled parametrically but important variables in the regimes are still chosen data-adaptively. The performance of C-learning is much better than BOWL and even the unrealistic BOWL† because, in addition to reasons explained above for Table 2, BOWL cannot handle regimes of the form of decision trees. For the same reasons explained for Table 1, when $n = 400$ and 800, C-learning performs even better than the overly optimistic benchmark, the method of Zhang et al. (2013)†, and the performance is close to that of the true optimal treatment regime.

Our simulation scenarios are either adopted from scenario 3 of Zhao et al. (2015) or further built upon it, and in this scenario, BOWL has overall better performance than two other OWL-based methods. In our additional simulations using scenarios 1 and 2 of Zhao et al. (2015), we see the same pattern of relative performances. Finally, we point out that OWL-based methods may be ill-behaved when Y can be negative, which is also noted by Chen et al. (2017). In our implementation, we have modified BOWL using a connection between OWL and IPWE as discussed in Zhang et al. (2012b) to overcome this difficulty, which is similar to a remedy proposed by Chen et al. (2017). See Supplementary Material Section F for more details, performance of the original BOWL and additional simulation studies.

5. Application

We applied the method to the data from the Sequential Treatment Alternatives to Relieve Depression (STAR*D) clinical trial. STAR*D was a multi-site, multi-step randomized clinical trial on 4041 patients with nonpsychotic major depressive disorder. To goal is to compare treatment options on the basis of severity of depression, assessed using the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2004), with higher values corresponding to higher severity. The trial involved four levels, each with a 12-week follow-up phase, and severity of depression was assessed at scheduled clinic visits at weeks 0, 2, 4, 6, 9, 12 during each level. At the end of each level, patients with adequate clinical response to that level's treatment did not move to future levels and patients without adequate response continued to future levels. During level 1, all patients received citalogram, at levels 2 and 3 patients were randomized to either "augment" previous treatments or "switch" to new treatments, and at level 4 patients were randomized to one of two switch options. The actual design is fairly complicated. See Rush et al. (2004)for more details and Schulte et al. (2014) for a schematic of the study design.

Following Schulte et al. (2014) and Zhang et al. (2013), we only consider level 2 and 3 and simplified the decision options at each stage. We consider the 1260 patients who entered level 2, redefining levels 2 and 3 as decision point 1 and 2 ($K = 2$). At decision points 1 and 2, we consider two treatment options, $A_k = 0$ or 1, where 0 (augment) means augmenting citalogram with one of other treatments and 1 (switch) means switching to one of other treatments. As for patients information used in decision making, we define $X_1 = (X_{11}, X_{12})^T$, where X_{11} is QIDS score at decision $k = 1$ and X_{12} is the slope of QIDS score based on QIDS score at baseline, X_{10}, and decision 1, and X_1 denotes the information available immediately prior to the first decision. Similarly, $X_2 = (X_{21}, X_{22})^T$ is the information available between decision points 1 and 2, where X_{21} denotes QIDS score at decision $k = 2$ and X_{22} is the QIDS score slope based on X_{11} and X_{21}. The outcome is the cumulative average negative QIDS score defined as $Y = -I(X_{21} \leq L_0)X_{21} - I(X_{21} > L_0)(X_{21} + T)/2$, where T is QIDS score at the end of decision point 2 and $L_0 = \max(5, X_{10}/2)$.

To implement C-learning, one needs to specify models for treatment assignment at each stage. Since in this study the treatment assignment (switch or augment) is not randomized and might depend on information available by that stage, we 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 $\pi_1(x_1; \gamma_1) = \text{expit}(\gamma_{10} + \gamma_{11}x_{11} + \gamma_{12}x_{12})$. Random forest was used to estimate $Q_2(L_{2i}, a)$ and $Q_1(L_{1i}, a)$, $a = 0, 1$. After obtaining the classification data set, we considered regimes of linear form and regimes of the form of a decision tree. For the former, a genetic algorithm was used and for the latter CART was used for optimization. Both implementations lead to the same estimated optimal treatment regime, which suggests that, for both stages, for patients who proceeded to that stage patients should switch. This result is consistent with that from BOWL. The estimated expected potential outcome under this regime is −7.91 (95% CI: −8.42, −7.39) using the

AIPWE method, as described in Zhang et al. (2013). As in Figure S1, we also plotted the classification data set used for classification in C-learning in Supplementary Material F (Figure S2), which suggests that the optimal treatment decision does not depend on patient characteristics that we have.

In Zhang et al. (2013), it restricted consideration to 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 $g_{\eta 2}(\bar{x}_2, a_1) = I(x_{22} > \eta_2)$. The optimal treatment regimes identified was to switch at decision point 1 if the decision 1 QIDS slope, x_{12}, is greater than -1.78 and switch at decision point 2 for all patients who proceeded to stage 2. The estimated expected potential outcome under this regime is $-7.85(-8.36, -7.33)$. Actually, in our data set only 48 patients among 1260 patients had QIDS slope $x_{12} \leq -1.78$ at stage 1 and under this estimated treatment regimes, most of the patients would switch at stage 1. For this particular data set, all methods lead to similar estimated regimes, which basically suggest that for patients who proceeded to that stage they should switch treatment.

6. Discussion

We show a general result that identifying the optimal dynamic treatment regime can be recast as a sequential classification problem that aims to minimize a weighted misclassification error at each stage; that is, at stage k, each subject can be viewed as belonging to one of two classes for whom the optimal decision at stage k given available patient characteristics and treatment history is 0 or 1. Based on this result, we proposed a powerful and flexible learning algorithm to learn the optimal treatment regime and the classification perspective allows us to exploit the wealth of existing/new powerful classification algorithms. As discussed in Section 3.1 and Supplementary Material D, this classification perspective is fundamentally different from that of Zhao et al. (2015) and offers considerable advantage in performance as illustrated by simulation studies, especially when the dimensionality of covariates is high. Moreover, this dramatically better performance is not due to modeling assumptions since in C-learning-RF the implementation is completely data-adaptive without any parametric model assumptions.

It is a direct optimization method that enjoys more robustness against model misspecification and it is also able to exploit outcome regression models (Q-functions) to improve efficiency. As discussed in Section 1, there is a mismatch between outcome regression-based approaches (Q- and A-learning) and the goal of optimizing decision rules. Nevertheless, intuitively and theoretically the optimal treatment decision should depend on how outcomes are related to patient characteristics and treatments and information from outcome regression models (even if incorrect or only approximately true) should be exploited to estimate the optimal treatment regime. Being a direct optimization approach, C-learning is able to alleviate the mismatch problem and exploit outcome models simultaneously and the two goals are achieved in C-learning by decoupling the optimization steps from the modeling steps.

C-learning is a flexible methodology. Within this framework, first, data analysts have the freedom to use all existing model building/selection techniques to best model Q-functions to improve efficiency. For example, one can model Q-functions using parametric regression models or nonparametric regression models (e.g., random forest), and all available model selection techniques (e.g., forward selection, Lasso, etc.) that target predictions can be readily incorporated. Second, existing powerful off-the-shelf optimization/classification tools (e.g., CART and genetic algorithm) can easily be accommodated in this framework to carry out the optimization to learn the optimal decision rules. In addition, as illustrated by our second simulation study, new and sophisticated variable selection techniques, targeting optimizing decision rules as opposed to predictions, can be developed within this framework to best select the important sets (and combination) of covariates and treatment history from among a high-dimensional set of covariates to form the optimal decision rules. Furthermore, this framework allows decision rules of different forms. In our simulations we illustrated this flexibility by considering both linear and tree decision rules. Other forms of decision rules can also be accommodated in this framework, making C-learning a flexible and general approach. Finally, we comment that the proposed method can be used to learn the optimal decision rule using data obtained from both clinical trials and observational studies.

7. Supplementary Materials

Web Appendices referenced in Sections 2–5 and R code implementing the method are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank Marie Davidian and Anastasios (Butch) A. Tsiatis for their valuable comments. This work was supported by National Natural Science Foundation of China (71701120 and 61472205), Program for Innovative Research Team of Shanghai University of Finance and Economics.

References

Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* https://doi.org/10.1111/biom.12676

Geng, Y., Lu, W., and Zhang, H. H. (2015). On optimal treatment regimes selection for mean survival time. *Statistics in Medicine* **34**, 1169–1184.

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

Kang, C., Janes, H., and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics* **70**, 695–720.

Mebane, W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software* **42**, 1–26.

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 discussion). *Journal of the Royal Statistical Society, Series B* **58**, 331–366.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24**, 1455–1481.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39** 1180–1210.

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.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., et al. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Controlled Clinical Trials* **25**, 119–142.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science* **29** 640–661.

Song, R.s, Wang, W., Zeng, D., and Kosorok, M. R. (2011). Penalized q-learning for dynamic treatment regimes. *Pre-Print, arXiv:1108.5338.*

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* **109**, 1517–1532

Wallace, M. P. and Moodie, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71**, 636–644.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning* **8**, 279–292.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012a). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.

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

Zhang, B. and Zhang, M. (2016). Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariate. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 1128.

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

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

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* **510**, 583–598.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association* **112**, 169–187.

Received August 2016. *Revised September* 2017.
Accepted October 2017.