University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2015 *Paper* 116

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

Baqun Zhang[*] Min Zhang[†]

[*] Renmin University of China, School of Statistics, zhangbaqun@ruc.edu.cn

[†] University of Michigan - Ann Arbor, mzhangst@umich.edu

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

Baqun Zhang and Min Zhang

Abstract

Personalizing treatment to accommodate patient heterogeneity and the evolving nature of a disease over time has received considerable attention lately. A dynamic treatment regime is a set 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 classification problem and is equivalent to sequentially minimizing a weighted expected misclassification error. This general classification perspective targets the exact goal of optimally individualizing treatments and is new and fundamentally different from existing methods. Based on this fresh classification perspective, we propose a novel, powerful and flexible C-learning algorithm to learn the optimal dynamic treatment regimes backward sequentially from the last stage till 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 dramatically improves performance, hence enjoying the 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.

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

Baqun Zhang
School of Statistics, Renmin University of China,Beijing, China
and
Min Zhang
Department of Biostatistics, University of Michigan, Ann Arbor

February 6, 2016

Abstract

A dynamic treatment regime is a set 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 novel, powerful and flexible C-learning algorithm to learn the optimal dynamic treatment regimes backward sequentially from the last stage till 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 dramatically improves performance, hence enjoying the 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.

Keywords: Personalized medicine; Dynamic treatment regime; Augmented Inverse Probability Weighted Estiamtor; Q-learning; A-learning; CART.

1 Introduction

Personalized medicine, which recognizes individual heterogeneity and focuses on making treatment decisions for a patient based on his/her own characteristics (eg., demographic, clinical, genetic information etc.) has received much attention lately (Moodie et al., 2007; Chakraborty et al., 2010; Song et al., 2011; Zhang et al., 2012ab, 2013; Zhao et al., 2012 and 2015; Geng et al., 2015; Wallace and Moodie, 2015). Treatment of patients may involve a series of decisions over time, especially in the case of chronic diseases, and the disease and conditions of a patient are also evolving. Therefore, it is important that the treatment decisions are adaptive with time-dependent information on patients over time. A dynamic treatment regime is a set of sequential decision rules, each corresponding to a decision point, that determine the next treatment from among possible options for an individual patient based on his/her own available information up to that time (Murphy, 2003; Robins, 2004). A dynamic treatment regime approach explicitly takes into account the heterogeneity among individuals and the evolving nature of a disease over time. The goal is to identify the optimal dynamic treatment regime, ie., the 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 in a sequential decision-making setting are Q- and A-learning(Watkins and Dayan, 1992; Murphy, 2003; Robins, 2004). Both approaches involve modeling for the outcome given covariate and treatment history to that point and treatment at the decision point. and the identification of the optimal treatment regime is through inverting the relationship between outcome, patient information and treatment. Q- and A-learning work well under good regression models for the outcomes. However, if the regression models are misspecified the estimated regime may far from optimal. This is due to that there is a mismatch between the target of the outcome regression based methods and the goal of learning the optimal treatment regime, as firstly pointed out by Murphy (2005). The 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 proposed to directly maximizing population mean outcome across regimes, assuming larger values are preferred. The advantage of direct optimization has been discussed in detail in literature mentioned below; also see Kang et al.(2014) and discussion papers for more discussions. Zhang et al.(2012a and 2013) proposed to estimate the population mean outcome under a given regime using a doubly robust augmented inverse probability weighted estimator (AIPWE) and then directly maximize AIPWEs across all regimes in a restricted class indexed by a finite number of parameters. Zhao et al.(2012 and 2015) proposed to estimate the population mean outcome using the simple inverse probability weighted estimator (IPWE), and then maximize IPWEs across regimes by taking advantage of support vector machine (SVM) techniques, referred to as outcome weighted learning (OWL). One other relevant work is that of 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. This framework allows one to take advantage of existing powerful classification techniques and, equally importantly, this framework allows the optimization step for optimizing decision rules decoupled from modeling outcomes as a function of patient characteristics and treatments. Therefore, it solves the mismatch issue pointed out by Murphy (2005) and in the meantime is able to take advantage of outcome regression modeling.

In this paper, we propose to extend the classification framework to the much more complicated multiple decision point setting and, as in aforementioned methods, this extension requires nontrivial and important methodological developments. This general framework builds upon existing work on outcome regression-based methods (Q- and A-learning) and the direct optimization methods discussed above and unifies them. In particular, the proposed classification framework is a direct optimization method, where the optimization can be viewed as a classification problem, and also allows incorporating information from outcome regression models, as in Q- and A-learning, to improve efficiency, hence enjoying the advantages of both types of approaches. This classification framework leads to a novel and general learning method that is very flexible in implementation and powerful in performance, as illustrated by comprehensive simulations studies. In addition, this general methodology allows sophisticated variable selection algorithms be developed within it, leading to a wealth of future learning methods.

2 Notation and Dynamic Treatment Regimes

Consider a multistage decision problem where decisions are made at K decision points. Assume there is a set of treatment options $\mathcal{A}_k = \{0, 1\}$, corresponding to each decision point $k = 1, \ldots, K$, and the element of $\mathcal{A}_k$ is denoted by a_k. A treatment history up to and including the kth decision is denoted as $\bar{a}_k = (a_1, \ldots, a_k)$, taking values in $\bar{\mathcal{A}}_k = \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$. Denote the treatment actually received at stage k as A_k and the observed treatment history up to decision k as $\bar{A}_k = (A_1, \ldots, A_k)$. Let X_k, taking values x_k in a set $\mathcal{X}_k$, be the covariate information observed between decision $k - 1$ and k, ie., after treatment A_{k-1} but prior to A_k. Similarly, we denote the observed covariate history up to k as $\bar{X}_k = (X_1, \ldots, X_k)$. The overall outcome of interest is Y, 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 and the goal is to identify the optimal sequential treatment decision rule leading to overall maximum expected outcome if the decision rule is followed by the entire population. The observed data are $(\bar{A}_{Ki}, \bar{X}_{Ki}, Y_i)$, assumed to be independent and identically distributed across subject $i, i = 1, \ldots, n$.

A dynamic treatment regime is a set of sequential decision rules, $g = (g_1, \ldots, g_K)$, that determine how to treat a patient over time, where g_k is a decision rule corresponding to stage k. The kth decision rule g_k assigns a treatment among $\mathcal{A}_k$ for a subject based on his/her covariate and treatment history up to decision k and hence is a function of $\bar{x}_k$ and $\bar{a}_{k-1}$, denoted as $g_k(\bar{x}_k, \bar{a}_{k-1})$. We define the potential outcome associated with any regime $g = (g_1, \ldots, g_K) \in \mathcal{G}$, denoted as $Y^*(g)$, ie., 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 maximum expected outcome if were followed by all patients in the population. That is, g^{opt} satisfies

$$E\{Y^*(g^{opt})\} \geq E\{Y^*(g)\} \text{ for all } g \in \mathcal{G}. \tag{1}$$

We make some standard assumptions that make g^{opt} identifiable from observed data (Schulte et al., 2014). First, we make the consistency assumption that $Y = Y^*(\bar{A}_K) = \sum_{\bar{a}_K \in \bar{\mathcal{A}}_K} Y_K^*(\bar{a}_K) I(\bar{A}_K = \bar{a}_K)$ and $X_k = X_k^*(\bar{A}_{k-1}) = \sum_{\bar{a}_{k-1} \in \bar{\mathcal{A}}_{k-1}} X_{k-1}^*(\bar{a}_{k-1}) I(\bar{A}_{k-1} = \bar{a}_{k-1})$ for $k = 2, \ldots, K$. We also assume that a patient's covariates and outcome are not affected by treatments received by other patients (the stable unit treatment value assumption; Rubin, 1978). Finally, we make the no unmeasured confounders (or sequential ignorability) assumption (Robins, 1994), i.e., given past treatment and covariate history, the treatment assignment at stage k is independent of potential outcomes. This assumption is reasonable if all information used in making treatment decisions in an observational study is recorded and is satisfied by design for data from a sequentially randomized clinical trial. Under these assumptions, g^{opt} can be expressed in terms of observed data via backward induction, also referred to as dynamic programming (eg., Zhang et al., 2013). Denoting $Q_K(\bar{x}_K, \bar{a}_K) = E(Y|\bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$, referred to as Q-functions with "Q" for "quality", then the optimal decision rule at the K-th decision point satisfies

$$g_K^{opt}(\bar{x}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Phi_K(\bar{x}_K, \bar{a}_{K-1})} Q_K(\bar{x}_K, \bar{a}_{K-1}, a_K). \tag{2}$$

Recursively we can define the value function (V-function) as

$$V_k(\bar{x}_k, \bar{a}_{k-1}) = \max_{a_k \in \mathcal{A}_k} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k), \tag{3}$$

for $k = K, \ldots, 2$, with $\bar{a}_0$ being null, and Q-functions as

$$Q_k(\bar{x}_k, \bar{a}_k) = E\{V_{k+1}(\bar{x}_k, X_{k+1}, \bar{a}_k)|\bar{X}_k = \bar{x}_k, \bar{A}_k = \bar{a}_k\} \tag{4}$$

for $k = K-1, \ldots, 1$. The optimal decision rule at the k-th point satisfies $g_k^{opt}(\bar{x}_k, \bar{a}_{k-1}) = \arg\max_{a_k \in \Phi_k(\bar{x}_k, \bar{a}_{k-1})} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$, which maximizes the expected potential outcomes that would be achieved if optimal decisions were made in the future. Appendix A provides some further background on this.

3 The proposed C-learning

To provide some intuition consider first the single decision point setting ($K = 1$), for which Zhang et al.(2012b) proposed a general framework for estimating the optimal regime from the 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 s/he to receive treatment 1 versus 0. By definition, $g^{opt} = \arg\max_{g \in \mathcal{G}} E\{Y^*(g)\}$. Because $E\{Y^*(g)\} = E\{g(X)C(X)\} + E\{Q(X, 0)\}$, it follows that $g^{opt} = \arg\max_{g \in \mathcal{G}} E\{g(X)C(X)\}$, ie., the optimal regime should assign treatment 1 to a subject if the expected potential outcome under treatment 1 is greater than that under 0. By separating information in $C(X)$ into sign $I(C(X) > 0)$ and magnitude $|C(X)|$, Zhang et al. (2012b) show that g^{opt} minimizes an expected weighted misclassification error; that is,

$$g^{opt} = \arg\min_{g \in \mathcal{G}} E\{|C(X)|I(Z \neq g(X))\}, \text{ where } Z = I\{C(X) > 0\}. \tag{5}$$

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 $Z = a$ include those subjects who would benefit from treatment a relative to the other 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. Note, (5) only involves patient characteristics (covaraites) and the true treatment contrast but not the observed treatment assignment and therefore can be viewed as an alternative definition of the optimal treatment regime.

In this article, we provide a novel and alternative definition of the optimal dynamic treatment regime in the multiple decision point setting from the classification perspective and, based on this fresh 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; ie., for stage k, $k = 1, \dots, K$, 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 $Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$ 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 and the proofs are given in the Appendix B and C.

Theorem 1. *A sequence of decision rules $g^* = (g_1^*, \dots, g_K^*)$, that satisfy the following conditions,*

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

$k = K, \dots, 1$, *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, ie., $I(C_k(L_k) > 0) \neq g_k(L_k)$, then the loss incurred is $|C_k(L_k)|$; otherwise, the loss is zero. In Theorem 1, the optimal dynamic treatment regime depends only on treatment contrast and patient characteristics at each stage, but not on the observed treatments in the data, and therefore Theorem 1 can be viewed as an alternative definition of the optimal dynamic treatment regime from the classification perspective. This is a general result that recasts the problem of identifying the optimal dynamic treatment regime into a meaningful sequential classification problem where the interpretation of the classification at each step is consistent with the definition of the optimal treatment regimes. We note that classification technique is first 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 classification technique is possible because of the form of the IPWE estimator, whereas the classification perspective of Theorem 1 is a general result that does not depend on any estimator of $E\{Y^*(g)\}$ or $C(X)$ or even the observed treatment A. For simplicity taking $K = 1$, the IPWE estimator is the empirical analogue of $E\{YI(A = g(X))/\pi(A, X)\}$, where $\pi(a, X) = Pr(A = a|X)$, and maximizing it is equivalent to minimizing $E\{YI(A \neq g(X))/\pi(A, X)\}$. Because of the particular form of IPWE, where a term $I(A \neq g(X))$ is involved, $I(A \neq g(X))$ 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. Second, the interpretation of classification is different, which has important implications on the properties of the resulting learning method. The classification in BOWL is to classify patients based on his/her characteristics to classes that actually received treatment A=0 or 1, ie., an error is made if $A \neq g(X)$. 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, ie., $A \neq g(X)$; see Zhang et al. (2012a) for details. In our classification perspective, based on patients characteristics we aim to classify patients to the class that would potentially benefit from one treatment relative to the other and hence should receive the particular treatment, ie., an error is made if $I(C(X) > 0) \neq g(X)$. The interpretation of this classification corresponds exactly to the intuitive meaning of optimizing individual treatment decisions. As pointed out by Zhou, et al. (2015), the estimated treatment regime from OWL-based methods tries to keep treatment assignments that subjects actually received and, from the discussion above, it is clear that this is due to that in the classification perspective of OWL-based methods it tempts to classify patients to the group corresponding to A, ie., an error is made if $g(X) \neq A$. The proposed classification framework does not suffer from this feature.

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.

Theorem 1, combined with proposition 1, leads to a very flexible and powerful learning method based on 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 the 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 shown in the proof of Theorem 1 that equivalently the optimal treatment rule at K minimizes a weighted misclassification error; that is

$$g_K^{opt} = \arg\min_{g_K \in \mathcal{G}_K} E\{|C_K(L_K)|I(Z_K \neq g_K(L_K))\}. \tag{6}$$

Therefore, g_K^{opt} can be estimated by

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

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 estimate of $C_K(L_{Ki})$. The contrast function can be estimated using various ways as discussed in Zhang et al.(2012b) and the AIPWE method has superior performance relative to other methods. Therefore, we recommend estimating $C_K(L_{Ki})$ by the AIPWE estimate

$$\begin{aligned} \widehat{C}_K(L_{Ki}) &= \frac{A_{Ki}}{\widehat{\pi}_K(L_{Ki})}Y_i - \frac{A_{Ki} - \widehat{\pi}_K(L_{Ki})}{\widehat{\pi}_K(L_{Ki})}\widehat{Q}_K(L_{Ki}, 1) \\ &- \{\frac{1 - A_{Ki}}{1 - \widehat{\pi}_K(L_{Ki})}Y_i + \frac{A_{Ki} - \widehat{\pi}_K(L_{Ki})}{1 - \widehat{\pi}_K(L_{Ki})}\widehat{Q}_K(L_{Ki}, 0)\}, \end{aligned} \tag{7}$$

where $\widehat{\pi}_K(L_{Ki})$ is the estimated probability of receiving treatment $A_K = 1$ at time 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. We acknowledge that other estimators of the 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 $\widehat{g}^{opt}_{C,K}$.

After obtaining $\widehat{g}^{opt}_{C,K}$, the C-learning moves backward sequentially till 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

$$g^{opt}_k = \arg\min_{g_k \in \mathcal{G}_k} E\{|C_k(L_k)|I(Z_k \neq g_k(L_k))\}, \tag{8}$$

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 and estimate g^{opt}_k by minimizing a weighted misclassification error. 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

$$\widetilde{V}_{ki} \equiv \widetilde{V}_k(L_{ki}) = \widetilde{V}_{(k+1)i} + \{\widehat{Q}_k(L_{ki}, 1) - \widehat{Q}_k(L_{ki}, 0)\}\{\widehat{g}^{opt}_{C,k}(L_{ki}) - A_{ki}\}, \tag{9}$$

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 $L_k = (\bar{X}_k, \bar{A}_{k-1})$ is regarded as 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

$$\begin{aligned} \widehat{C}_k(L_{ki}) &= \frac{A_{ki}}{\widehat{\pi}_k(L_{ki})}\widetilde{V}_{(k+1)i} - \frac{A_{ki} - \widehat{\pi}_k(L_{ki})}{\widehat{\pi}_k(L_{ki})}\widehat{Q}_k(L_{ki}, 1) \\ &- \{\frac{1 - A_{ki}}{1 - \widehat{\pi}_k(L_{ki})}\widetilde{V}_{(k+1)i} + \frac{A_{ki} - \widehat{\pi}_k(L_{ki})}{1 - \widehat{\pi}_k(L_{ki})}\widehat{Q}_k(L_{ki}, 0)\}, \end{aligned} \tag{10}$$

where $\widehat{\pi}_k(L_{ki})$ are estimated propensity score $P(A_{ki} = 1|L_{ki})$ based on, say, 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 (8), $g_k^{opt}(L_k)$ can be estimated by

$$\widehat{g}_{C,k}^{opt} = \arg\min_{g_k \in \mathcal{G}_k} \sum_{i=1}^{n} \widehat{W}_{ki} I(\widehat{Z}_{ki} \neq g_k(L_{ki})) \tag{11}$$

using some classification or optimization technique. The final estimated optimal regime is $\widehat{g}_C^{opt} = (\widehat{g}_{C,1}^{opt}, \ldots, \widehat{g}_{C,K}^{opt})$.

C-learning is a very flexible approach. First, all existing modeling building/selection techniques can be used to best estimate the Q-function to improve efficiency, for example, parametric regression (see discussions in Zhang et al., 2012a and 2013 on augmentation terms) as in Q-learning or nonparametric regression. Second, existing powerful optimization/classification techniques can easily be used in the optimization step to optimize decision rules. Moreover, new variable selection methods targeting optimizing decision rules instead of Q-functions can be developed within this framework in thet optimization step, for example, to handle high dimensional covariates and improve performance. Last, decision rules can be linear decision rules, decision trees or of other forms. This flexibility and the resulting superior performance is illustrated by various examples in simulation studies.

4 Simulation Studies

We report results on simulation studies under various scenarios. Specifically, **(i).** scenario 1 is adopted from Zhao et al.(2015). **(ii).** Scenario 2 is otherwise similar to scenario 1 except that the number of covariates are of a much higher dimension and the optimal decision rule at each stage depends on more covariates. **(iii).** Scenario 3 also considers a high dimensional set of covariates, but the true optimal treatment regime is of a tree form. Scenarios 1-3 imitate a multi-stage randomized trial with three stages ($K = 3$). We vary the sample size n (n=200, 400 or 800) and use 500 Monte Carlo Replicates for each scenario. For each simulated data set, we apply the proposed C-learning method as well as other methods including BOWL (Zhao et al., 2015), Q-learning and the method of Zhang et al. (2013).

4.1 Data Generation and Methods Implementation

The first simulation was adopted from Zhao et al. (2015). Treatments A_1, A_2 and A_3 are randomly generated from $\{1, 0\}$ with equal probability 0.5. Three 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 $X_3 \sim N(0.5X_2, 10^2)$. The outcome was generated as $Y = \mu(\bar{A}_3, \bar{X}_3) + \epsilon$ for ϵ standard normal and $\mu(\bar{A}_3, \bar{X}_3) = 20 - |0.6X_{1,1} - 40|(A_1 - g_1^{opt})^2 - |0.8X_2 - 60|(A_2 - g_2^{opt})^2 - |1.4X_3 - 40|(A_3 - g_3^{opt})^2$, 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

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

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, 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 so that the total number of covariates is 50. Treatments A_1,A_2 and A_3 are randomly generated from $\{1, 0\}$ with equal probability 0.5. At baseline, 40 covariates $X_{1,1}, ..., X_{1,40}$ are generated from $N(45, 15^2)$. At stage 2, $X_{2,j}$ is generated according to $X_{2,j} \sim N(1.5X_{1,j}, 10^2)$, $j = 1, ..., 5$. At stage 3, $X_{3,j}$ is generated according to $X_{3,j} \sim N(0.5X_{2,j}, 10^2)$, $j = 1, ..., 5$. 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 depend on a linear combination of two covariates instead of a single covariates as in scenario 1.

For Q-learning, we posited Q-functions

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

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

Note, these model specifications favor the Q-learning method in that they only include the correct interaction of treatment and covariate at each stage 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 real application, although possible, it is difficult to prespecify the right variables and forms for the true optimal regime and the results on these methods 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 C-learning method.

Unlike the other methods, in the implementation of the C-learning, we did not prespecify 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-learning. The forward selection algorithm is detailed in a unpublished technical report. We implemented C-learning using two different ways that differ in how AIPWE is constructed: in C-learning-Q, we used 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. In both ways, all linear decision rules constructed by the high dimensionl set of covariates are considered, as opposed to Q-learning, the method of Zhang et al. (2013) and BOWL in one implementation. For our purpose the random forest is simply a black box predictor which takes as input covariate values $(L_{ki},\, A_{ki})$, and gives as output an estimate of $E(\widetilde{V}_{k+1}|L_k, A_k) = Q_k(L_k, A_k)$ for that set of covariate values. Fitting of the random forest is done using the R function *randomForest* with default settings.

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, which is more familiar to clinicians and perhaps more in line with the current practice in medicine. In implementation, the posited Q-function models were taken to be the same as in the second simulation. Therefore, out of a relative high dimensional candidate variables, the correct sets of important variables were used to favor the performance of this method, but the form of the optimal treatment regime was still misspecified. 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 $g_{\eta_k}(\bar{x}_k, \bar{a}_{k-1}) = I(X_{k,1} > \eta_{k,1})I(X_{k,2} < \eta_{k,2})$, $k = 1, 2, 3$. Again, in BOWL, we also limited the search among regimes constructed by relevant variables in one implementation. As above we implemented C-learning using two different ways, C-learning-Q and C-learning-RF. 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 sets of covariates, instead of $\mathcal{G}_\eta$ in the implementation of Zhang et al. (2013). We used the `R` function `rpart` with default settings, except that we set the weights as the estimated weight $\widehat{W}$.

4.2 Results and Discussion

Results from scenarios 1-3 are shown in Tables 1-3 respectively. Table 1 shows that the proposed C-learning out-performs all other methods in this scenario. We first note that, in Table 1, all methods consider the same class of regimes, which is different from Table2 and 3 discussed below. C-learning performs considerably better than Q-learning even though it used the same models for Q-functions in the augmentation terms and this is because that the performance of C-learning is not dictated by the specification of the Q-function and the optimization step directly targets the optimization of treatment regimes. It is also interesting to note that, although C-learning and the method of Zhang et al. (2013) are based on the same AIPWEs, and consider optimization over 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; ie., Zhang et al. (2013) simultaneously estimates regimes at all stages and C-learning backward sequentially estimates the regime at each stage. The method of Zhang et al. (2013) is based on an AIPWE estimator of $E\{Y^*(g)\}$ for monotone coarsening (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 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. C-learning has better performance than BOWL due to two reasons. First, C-learning uses outcome regression model in the augmentation terms to improve efficiency whereas BOWL does not. Second, C-learning and BOWL differ in their way to handle multiple stages. The extension to multiple stages in C-learning is based on proposition 1 which allows us to use information on all subjects at all stages. In BOWL, the extension to multiple stages is based on an IPWE for monotone coarsening data, as in Zhang et al. (2013), and to sequentially estimate the regimes it has to lose sample size geometrically with stages.

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, the performance of Q-learning, BOWL (in one implementation), and the method of Zhang, et al. (2013) is 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. A superscript † is used to indicates the difference in implementation. C-learning (both implementations) as well as BOWL consider all regimes constructed by linear combinations of the high dimensional set of covariates, whereas Q-learning†, BOWL† and the method of Zhang et al. (2013)† only consider regimes constructed by relevant covariates, which is a much smaller class. As explained previously, this is because we try to give the best advantage to our comparison methods in implementation since the performance of Q-learning and the method of Zhang et al. (2013) is highly 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 of dealing with the high dimensional set of covariates, the performance of the C-learning (both implementations), combined with suitable variable selection algorithm in the 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 for reasons explained previously. The C-learning framework can naturally accommodate variable selection methods targeted towards optimal treatment regime instead of prediction to improve performance in the presence of high dimensional covariates. This (in addition to reasons discussed previously for Table 1) explains the dramatically better performance 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. The pattern of relative performance is similar to that in Table 2. As in Table 2, here BOWL† , Q-learning† and the method of Zhang et al. (2013)† consider only important covariates in optimizing regimes and hence results on these methods are overly optimistic. The C-learning-RF, with both outcome regression models and important variables in the regimes chosen data-adaptively using existing off-the-shelf algorithms and software (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† where the important variables are taken to be known and dramatically better than BOWL when BOWL searches the optimal treatment regimes among the same class as C-learning. 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 approaches that of the true optimal treatment regime.

Finally, we comment that 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 SOWL and IOWL (other two OWL-based methods). In our additional simulations using scenarios 1 and 2 of Zhao, et al. (2015), we see the same pattern of

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)

relative performances of C-learning versus BOWL . We have already discussed reasons that lead to the difference in performance and they are further summarized in Section 5.

5 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; ie., 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. This equivalence leads to a novel and general learning method that allows us to exploit the wealth of existing/new powerful classification algorithms. We comment that Zhao, et al. (2015) first exploited powerful classification techniques to estimate the optimal dynamic treatment regime in the multiple stage setting. Our classification framework differs from their work in several aspects. First, as discussed below Theorem 1, Theorem 1 is a general result and this equivalence does not

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.

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
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)

depend on any particular estimator of the contrast functions or the observed treatment assignment, whereas the OWL-based method of Zhao, et al. (2015) is based on a specific IPWE estimator and the classification perspective is possible due to the IPWE estimator. As a result, although we focus on using AIPWE to estimate the contrast functions in our presentation, the framework can accommodate other estimators of the contrast functions, for example, the difference in estimators of the Q-functions. Second, the interpretation of the classification is different and this has important implications. In OWL-based methods, an error is made if $g(X) \neq A$, ie., if the prescribed treatment is different from the observed treatment, and in our framework an error is made if $g(X) \neq I(C(X) > 0)$, ie., if the prescribed treatment is different from the one that leads to larger expected potential outcomes. As a result, our approach does not suffer from the feature of OWL-based methods as noted by Zhou, et al. (2015), namely, OWL-based methods tempts to classify patients to the group corresponding to the actually assigned treatment A. Third, as discussed in Section 4.2, the way to handle multiple stages are different. In direct optimization methods including BOWL and other OWL-based methods studied in Zhao, et al. (2015) and the robust AIPWE-based method of Zhang, et al. (2013), the extension to multiple stages are based on the monotone coarsening idea for IPWE/AIPWE estimators. This would naturally suggest simultaneous optimization across stages as in SOWL of Zhao, et al. (2015) and to achieve sequential optimization, BOWL needs to decrease sample size geometrically. In the proposed approach, the extension to multiple stages is based on Proposition 1 and this is in spirit more similar to outcome-regression based methods (Q- and A-learning). As discussed in Section 4.2, this approach has advantages and leads to better performance. We are not aware of any direct optimization methods that are able to handle multiple stages this way. Our approach, being a direct optimization method, is able to take advantage of this and combine the benefits of both direct optimization methods and outcome-regression based methods. Finally, our classification framework can naturally accommodate model selection targeted for optimizing decisions instead of prediction, whereas the classification method of Zhao, et al. (2015) cannot, as noted by Zhou, et al. (2015).

Based on this general result, we proposed a novel, powerful and flexible C-learning algorithm to learn the optimal treatment regime. It is a direct optimization method that targets the goal of optimizing decision rules 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 approach (Q- and A-learning) and the goal of optimizing decision rules. Nevertheless, outcome regression based approaches are appealing as 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. The proposed C-learning is able to address the mismatch problem and exploit outcome regression models simultaneously and the two goals are achieved in C-learning by decoupling the optimization steps from the modeling steps.

The C-learning is a very flexible general methodology. As illustrated by our simulations studies, within this framework, first, data analysts have the freedom to use all existing model building/selection techniques to best model the Q-functions to improve efficiency. For example, one can model the Q-function using parametric regression models or nonparametric regression models (eg, random forest), and all available model selection techniques (eg., forward selection, Lasso., etc) that target predictions can be readily incorporated. Second, existing powerful off-the-shelf optimization tools can easily be accommodated in this framework to carry out the optimization to learn the optimal decision rules. In addition, new and sophisticated variable selection techniques, targeting optimizing decision rules in contrast 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. This point is illustrated by our simulation studies and will be the focus of future work. 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 the C-learning a very flexible and general approach.

Because of the flexibility and the advantages discussed above, the proposed C-learning has superior performance relative to existing methods. We have devoted a whole subsection 4.2 in the simulation section to discuss and explain the results.

Acknowledgement

The authors would like to thank Marie Davidian and Anastasios (Butch) A. Tsiatis for their valuable comments.

Supplementary material

Supplementary material includes Proofs to Theorem 1 and Proposition 1 and a data application.

References

Chakraborty, B., Murphy, S. A. & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statist. Meth. Med. Res.* **19**, 317–343.

Geng, Y., Lu, W. & 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. & Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations *Biometrics* **70**, 695–720.

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.

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

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

Qian, M., & 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.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

Schulte, P.J., Tsiatis, A.A., Laber, E.B. & Davidian, M. (2014). Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Statist. Sci.* **29** 640–661.

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

Tian, L., Alizadeh, A.A., Gentles, A.J. & 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., & Moodie, E.E.M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares *Biometrics* **71**, 636-644.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Mach. Learn.* **8**, 279–292.

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

Zhang, B.,Tsiatis, A. A., Laber, E. B., & 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. & 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 & 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. & Kosorok, M. R. (2015). Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association* DOI: 10.1080/01621459.2015.1093947.