Electronic Journal of Statistics
Vol. 12 (2018) 3601–3638
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1480

D-learning to estimate optimal individual treatment rules

Zhengling Qi

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: qizl1027@live.unc.edu

Yufeng Liu

Department of Statistics and Operations Research, Department of Genetics,
Department of Biostatistics, Carolina Center for Genome Science,
Lineberger Comprehensive Cancer Center
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: yfliu@email.unc.edu

Abstract: Recent exploration of the optimal individual treatment rule (ITR) for patients has attracted a lot of attentions due to the potential heterogeneous response of patients to different treatments. An optimal ITR is a decision function based on patients' characteristics for the treatment that maximizes the expected clinical outcome. Current literature mainly focuses on two types of methods, model-based and classification-based methods. Model-based methods rely on the estimation of conditional mean of outcome instead of directly targeting decision boundaries for the optimal ITR. As a result, they may yield suboptimal decisions. In contrast, although classification based methods directly target the optimal ITR by converting the problem into weighted classification, these methods rely on using correct weights for all subjects, which may cause model misspecification. To overcome the potential drawbacks of these methods, we propose a simple and flexible one-step method to directly learn (D-learning) the optimal ITR without model and weight specifications. Multi-category D-learning is also proposed for the case with multiple treatments. A new effect measure is proposed to quantify the relative strength of an treatment for a patient. We show estimation consistency and establish tight finite sample error bounds for the proposed D-learning. Numerical studies including simulated and real data examples are used to demonstrate the competitive performance of D-learning.

Keywords and phrases: Precision medicine, multiple treatments, kernel learning, prescriptive variable selection.

Received September 2017.

1. Introduction

Precision Medicine has recently gained increasing attention in scientific research. The goal of precision medicine is to identify the optimal individual treatment rule (ITR) by considering the patients' heterogeneity, such as demographics, background and genetic information, to maximize each patient's expected clin-

ical outcome. Mathematically speaking, ITR is a function mapping from the covariate space into the treatment space.

There is a fast growing literature on estimating ITRs based on observational studies or randomized clinical trials. Existing approaches could be categorized into two main types, model-based and classification-based methods. Q-Learning ([33], [23], [26], [24]) and A-learning ([22], [25]) are two representative model-based methods in precision medicine. Q-learning models the conditional mean of clinical outcomes given the patients' covariates and treatment, while A-learning, which can be more robust to model misspecification than Q-learning, models the contrast function of outcome between treatments. Recently, [35] proposed a doubly robust augmented inverse propensity score weighted (IPSW) estimator to estimate the ITR. [30] proposed a modified covariates regression method to estimate the ITR, and [9] proposed a new concordance-assisted learning method to estimate the optimal ITR.

As an interesting alternative approach, the classification-based method was first proposed by [37]. They showed that maximizing the individual clinical outcome is equivalent to minimizing a weighted classification error, and they proposed the outcome weighted learning (OWL) method by using clinical outcomes as the classification weights. To further improve the finite sample performance of OWL, [20] proposed an augmented OWL method, and [39] considered a residual weighted learning method (RWL). Tree-based methods ([16, 7]) under the classification framework were considered to enhance the interpretability of ITR.

In precision medicine, variables that have qualitative interactions with the treatment are called prescriptive variables ([12]). Correctly identifying prescriptive variables can help save time and the cost of collecting unnecessary information in clinical practice. Although modern variable selection techniques have been used in model-based methods, they mainly focus on variables for prediction and may neglect the prescriptive variables that have weak predictive power but are important for decision making. This may cause the mismatch between predicting clinical outcomes and optimizing ITRs for model-based methods. [12] and [8] proposed methods using an additional step to fill the gap between prescriptive variables and prognostic variables.

For classification-based methods, OWL effectively formulated the problem as a weighted classification framework ([37]). Despites its success, OWL may be affected by a constant shift of the clinical outcome and tends to keep the treatment assignments that patients actually received in randomized trials ([39]). RWL by [39] uses a model-based method to compute the weights to improve the finite sample performance of OWL and also incorporates a variable selection procedure. Although RWL could alleviate some potential issues of OWL, it may suffer from the potential main effect model misspecification problem in calculating residuals as the weights for classification. In addition, the computational cost of RWL can be high due to the use of the non-convex ramp loss function, especially when the dimension is large. Recently, [27] proposed a sparse OWL under the classification framework for variable selection. Despite these existing methods, more developments are needed for effective ITR estimation.

In this paper, we propose a novel one-step method to directly learn (D-learning) the optimal ITR without specifying the main effect model and weights for both binary and multiple treatment settings, and simultaneously perform variable selection on prescriptive variables for linear models. The extensions to nonlinear models by kernel regression are discussed as well. The proposed D-learning is very simple and flexible. It combines the advantages of both model-based and classification-based methods. Furthermore, we propose a new measure to quantify the relative strength of all treatments. Such a measure can provide additional information among the treatments beyond ITRs for doctors and patients to make better decisions. We also present comprehensive theoretical results of D-learning under both linear and nonlinear models.

The remainder of this paper is organized as follows. In Section 2, we briefly review some existing methods and introduce D-learning for estimating the optimal ITR. In Section 3, we establish estimation consistency and convergence rates of D-learning under various settings. In Section 4, we conduct an extensive simulation study to evaluate D-learning by comparing it with several alternative methods. In Section 5, we analyze acquired immune deficiency syndrome (AIDS) randomized clinical trial data ([13]) using D-learning and compare with several other alternative methods. We conclude the paper with some discussion in Section 6.

2. Direct learning for individual treatment rules

For notation, we use boldface capital and lowercase symbols to denote matrices and vectors respectively, with the exception of the random vector $\mathbf{X}$ defined below. We first consider the framework of a binary treatment randomized trial. For each patient, we observe a treatment $A \in \mathcal{A} = \{1, -1\}$, the baseline information $\mathbf{X} = (1, X_1, \cdots, X_p)^T \in \mathcal{X}$, treatment assignment of patients during the study and the clinical outcome R after receiving the treatment. Without loss of generality, we assume the larger R is, the better condition a patient is in. The treatment rules are the set of deterministic decision functions that map the patient's covariate space into the treatment space. We define $\pi(a, x) = \mathbf{P}[A = a|\mathbf{X} = x]$ to be the probability of a patient being assigned to treatment a conditioning on the covariates $\mathbf{X}$ under the randomized trial framework. For observational studies, $\pi(a, x)$ denotes the propensity score and can be estimated via various methods such as logistic regression. We assume $\pi(a, x) > 0$ for any $a \in \mathcal{A}$, given $\mathbf{X} \in \mathcal{X}$ almost surely. For simplicity, we assume $\pi(a, x)$ is known for our discussion.

An optimal ITR is defined as the decision rule that maximizes the expected clinical outcome among all candidate rules. According to [24], the expected clinical outcome under the rule d could be written as

$$V(d) = \mathbf{E}[R|A = d(\mathbf{X})] = \mathbf{E}[\frac{R}{\pi(A, \mathbf{X})}\mathbb{I}(A = d(\mathbf{X}))], \tag{2.1}$$

where $\mathbb{I}(\bullet)$ is the indicator function. This quantity is called the value function, which we denote as $V(d)$ associated with the treatment rule d. Then the optimal

rule is defined

$$d_0(\mathbf{X}) = \text{argmax}_{d\in\mathcal{D}} V(d) \tag{2.2}$$

for a specific class of treatment rules $\mathcal{D}$. Before introducing D-learning, we highlight three previous methods for comparison.

2.1. Previous related methods

[24] proposed l_1-PLS as one of fundamental model based methods. They developed a two-stage procedure to estimate the optimal ITR. The first step is to compute the conditional mean of the clinical outcome, $\mathbf{E}[R|\mathbf{X}, A]$, given covariates and treatment by using l_1 penalized regression. The second stage is to compare this computed conditional mean under different treatments to derive the estimated ITR. The l_1-PLS method is effective by using a rich class of functions to approximate the conditional mean of R, and is interpretable by including the variable selection procedure. However, l_1-PLS does not directly target on prescriptive variables and the performance guarantee relies on the correctness of model specification. The implicit goal of l_1-PLS focuses on the prediction accuracy of the conditional clinical response which may be suboptimal in optimizing the decision rule.

To avoid modeling R directly, [37] proposed a very interesting OWL method to maximize (2.1) under the weighted classification framework by showing

$$d_0(\mathbf{X}) = \text{argmin}_{d\in\mathcal{D}} \mathbf{E}[\frac{R}{\pi(A,\mathbf{X})}\mathbb{I}(A \neq d(\mathbf{X}))]. \tag{2.3}$$

They use the convex hinge loss in the Support Vector Machine ([5]) to substitute the 0-1 loss function in (2.3). In particular, OWL is to find an optimal ITR by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i(1-A_i f(\mathbf{X}_i))_+}{\pi(A_i,\mathbf{X}_i)} + \lambda||f||^2,$$

where $(\cdot)_+ = \max(\cdot, 0)$ and $||f||$ is some norm of function f. Although OWL directly targets on the prescriptive variables and could possibly make a better decision, it requires positive R to compute and the resulting treatment rule tends to keep treatment assignments that the patients actually received in the randomized trial ([39]).

[39] recently proposed RWL to improve the finite performance of OWL by developing a two-step procedure. The first step is to calculate the residuals as weights r_i by regressing R on X. After calculating the weights which can be negative, they used a non-convex ramp loss function T to compute the optimal ITR under the classification framework. In particular, RWL aims to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i T(A_i f(\mathbf{X}_i))}{\pi(A_i,\mathbf{X}_i)} + \lambda||f||^2.$$

An effective difference of convex algorithm was applied in RWL. However, RWL may require relative large computational costs especially in high dimensional or

nonlinear decision boundary settings. More importantly, RWL highly relies on the correct calculation of the residuals and may fail if the model for residual calculation in the first step is misspecified. Furthermore, additional variability may be introduced by using a two-step procedure. Both OWL and RWL were designed for binary treatment settings. In Section 2.2, we propose an effective method to combine the strengths of both model-based and classification-based methods.

2.2. D-learning

Our goal is to estimate the optimal ITR by a single step and reduce the risk of model misspecification. In particular, model-based methods impose certain regression model assumptions to fit $\mathbf{E}[R|\mathbf{X}, A]$. For the classification based method RWL, calculating the residual r_i is needed before the weighted classification step. One of the main motivations of D-learning is to avoid such model assumptions. As shown in both [24] and [37], we could rewrite the optimal ITR in (2.2) as

$$d_0(\mathbf{X}) = \text{sign}(\mathbf{E}[R|\mathbf{X}, A = 1] - \mathbf{E}[R|\mathbf{X}, A = -1]) := \text{sign}(f_0(\mathbf{X})). \tag{2.4}$$

Note that the optimal decision function $f_0(\mathbf{X})$ could be further written as

$$\begin{aligned} f_0(\mathbf{X}) &= \mathbf{E}[R|\mathbf{X}, A = 1] - \mathbf{E}[R|\mathbf{X}, A = -1] \\ &= \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}, A = 1]\pi(1, \mathbf{X}) + \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}, A = -1]\pi(-1, \mathbf{X}) \\ &= \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}]. \end{aligned} \tag{2.5}$$

We observe that (2.5) gives us a direct way to learn the optimal ITR by estimating the conditional expectation $\mathbf{E}[\frac{RA}{\pi(A,\mathbf{X})}|\mathbf{X}]$. Then the estimated ITR is based on the sign of the estimated $\mathbf{E}[\frac{RA}{\pi(A,\mathbf{X})}|\mathbf{X}]$. In contrast to OWL which requires all rewards to be nonnegative, $f_0(\mathbf{X})$ is invariant to a constant shift for the reward R. If R is replaced by $R + c(\mathbf{X})$ for any random variable $c(\mathbf{X})$ depending only on $\mathbf{X}$, then

$$\begin{aligned} \mathbf{E}[\frac{(R + c(\mathbf{X}))A}{\pi(A, \mathbf{X})}|\mathbf{X}] &= \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}] + \mathbf{E}[\frac{c(\mathbf{X})A}{\pi(A, \mathbf{X})}|\mathbf{X}] \\ &= \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}] + \mathbf{E}[c(\mathbf{X}) - c(\mathbf{X})] \\ &= \mathbf{E}[\frac{RA}{\pi(A, \mathbf{X})}|\mathbf{X}] = f_0(\mathbf{X}). \end{aligned} \tag{2.6}$$

Using the results in (2.5), our D-learning estimates $f_0(\mathbf{X})$ and uses $\text{sign}(f_0(\mathbf{X}))$ as the estimated ITR. To further understand D-learning, we would like to point out an interesting interpretation. Assume that the clinical outcome R could be expressed as

$$R = m(\mathbf{X}) + \delta(\mathbf{X})A + W, \tag{2.7}$$

where W is the mean zero random error term, $m(\mathbf{X})$ is the main effect of covariates $\mathbf{X}$ for both treatments. Then one can see that $f_0(\mathbf{X}) = \mathbf{E}[\frac{RA}{\pi(A,\mathbf{X})}|\mathbf{X}] = 2\delta(\mathbf{X})$, which captures the heterogeneity between the two treatment effects without the need of estimating $m(\mathbf{X})$. The general form of R for a multiple treatment setting can be represented as

$$R = m(\mathbf{X}) + \sum_{k=1}^{K} \delta_k(\mathbf{X})\mathbf{1}(A = k) + W.$$

The next lemma provides us a way to estimate $f_0(\mathbf{X})$.

Lemma 1. *Under the assumption of interchange between differentiation and expectation,* $f_0(\mathbf{X}) \in argmin_f\ \mathbf{E}[\frac{1}{\pi(A,\mathbf{X})}(2RA - f(\mathbf{X}))^2]$.

Lemma 1 offers a simple way to estimate $f_0(\mathbf{X})$. Specifically, there are many existing regression methods we can adopt to estimate $f_0(\mathbf{X})$. In the next two subsections, we consider both linear and nonlinear D-learning to estimate ITR.

2.2.1. D-learning for linear decision rules

We first consider the setting of a two-arm randomized trial. Assume we observe independent identically distributed triplet data $\{(A_i, \mathbf{X}_i, R_i); i = 1, \cdots, n\}$. If we consider $\mathcal{F} := \{f(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathcal{R}^p\}$ to be the class of linear functions to approximate $f_0(\mathbf{X})$, then the ordinary least square (OLS) estimator is given by

$$\hat{\boldsymbol{\beta}}_n^{\text{ols}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\pi(A_i, \mathbf{X}_i)}(2R_iA_i - \mathbf{X}_i^T\boldsymbol{\beta})^2. \tag{2.8}$$

However, in high dimensional settings especially when $p > n$, OLS may fail to estimate the parameter vector $\boldsymbol{\beta}$ with risk of potential over-fitting. Therefore, it is desirable to perform sparse regularization to improve the prediction accuracy and interpretability of the model. In our case, this will not only help us identify the crucial prescriptive variables but also enhance the estimation accuracy of the decision boundary.

There are many linear regression techniques in the literature. The Least Absolute Shrinkage and Selection Operator (LASSO) is one of the most famous variable selection methods in high dimensional regression and the estimator is given by:

$$\hat{\boldsymbol{\beta}}_n^{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\pi(A_i, \mathbf{X}_i)}(2R_iA_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + 2\lambda \sum_{i=1}^{p} |\beta_i|, \tag{2.9}$$

where λ is the tuning parameter for the l_1 penalty. Then the resulting ITR is given by

$$\hat{d}_n(\mathbf{X}) = \text{sign}(\hat{f}_{0(n)}(\mathbf{X})) = \text{sign}(\mathbf{X}^T\hat{\boldsymbol{\beta}}_n^{\text{lasso}}). \tag{2.10}$$

Besides the LASSO, one could also use other convex penalties such as the elastic net ([40]), and non-convex penalties such as SCAD ([10]) and MCP ([36]).

2.2.2. D-learning for nonlinear decision rules

While linear D-learning is simple and interpretable, more complex nonlinear decision rules may be necessary sometimes in practice. Thus we consider two nonlinear methods to estimate $f_0(\mathbf{X})$, Kernel Ridge Regression (KRR) ([6]) and the Component Selection and Smoothing Operator (COSSO) ([19]).

KRR combines the kernel trick in machine learning with ridge regression by optimizing the following objective function to estimate the decision rule

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi(A_i,\mathbf{X}_i)}(2R_iA_i - f(\mathbf{X}_i))^2 + \frac{\lambda}{2}||f||^2_{\mathcal{H}_K}, \tag{2.11}$$

where $\mathcal{H}_K$ denotes the Reproducing Kernel Hilbert Space (RKHS) and $\|\bullet\|_{\mathcal{H}_K}$ is the corresponding norm. By the representer theorem ([15]), we can express the function $f(\mathbf{X}_i) = \sum_{j=1}^{n} c_j K(\mathbf{X}_i, \mathbf{X}_j)$, where $\mathbf{K}$ denotes the $n \times n$ kernel matrix and $K(\mathbf{X}_i, \mathbf{X}_j)$ denotes the (i, j)-entry of $\mathbf{K}$. The RKHS norm could be evaluated as $||f||^2_{\mathcal{H}_K} = \boldsymbol{c}^T\mathbf{K}\boldsymbol{c}$. Then the optimization is equivalent to optimizing over $\boldsymbol{c} \in \mathcal{R}^n$ via

$$\min_{\boldsymbol{c}\in\mathcal{R}^n} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi(A_i,\mathbf{X}_i)}(2R_iA_i - [\mathbf{K}\boldsymbol{c}]_i)^2 + \frac{\lambda}{2}\boldsymbol{c}^T\mathbf{K}\boldsymbol{c}, \tag{2.12}$$

where $[\mathbf{K}\boldsymbol{c}]_i$ denotes the i-th element of the vector $\mathbf{K}\boldsymbol{c}$. Note that KRR maps the covariates into an infinite dimensional space and is computationally efficient due to the kernel trick. However, it does not provide model selection for nonlinear function estimation. In order to perform model selection in nonlinear function estimation, we use the COSSO proposed by [19] based on the smoothing spline analysis of variance (SS-ANOVA) model ([32]). The SS-ANOVA model can be written as

$$f(x) = \alpha + \sum_{i=1}^{p} f_i(x_i) + \sum_{j<k} f_{jk}(x_j, x_k) + \cdots,$$

where α is the intercept, $\mathbf{X}_i = (x_1, \cdots, x_p)^T$, f_i's are baseline functions and f_{jk}'s are two-way interactions, etc. Then the functional space in the SS-ANOVA model could be written as

$$\mathcal{H}_K = \{1\} \oplus \mathcal{H}_1 \quad \text{with} \quad \mathcal{H}_1 = \oplus_{i=1}^{d}\mathcal{H}^i,$$

where $\mathcal{H}^i; i = 1, \cdots, d$, are d orthogonal subspaces of $\mathcal{H}_K$. If the model is only related to baseline functions, then $d = p$. For a two-way interaction model, $d = \frac{p(p+1)}{2}$. With $\mathcal{H}_K$ in place, COSSO estimates f via minimizing

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi(A_i,\mathbf{X}_i)}(2R_iA_i - f(\mathbf{X}_i))^2 + 2\lambda\sum_{j=1}^{d}||P^i f||, \tag{2.13}$$

where $P^i f$ is the orthogonal projection of f on $\mathcal{H}^i$ and $||\bullet||$ is the pseudonorm in RKHS. [19] showed that (2.13) could be solved efficiently via alternative minimization methods.

As a remark, we would like to point out that [30] used a similar formula as (2.9) for estimating interactions between a treatment and covariates while the working model is linear. However, in addition to linear learning, we also propose D-learning for nonlinear decision rules, which can capture a broader class of functions. This is not equivalent to fitting a kernel regression using the formulation in [30]. More importantly, in Section 2.3, we extend our proposed D-learning to handle multiple treatments in ITR problems.

2.3. Multi-category D-learning

Most existing methods for estimating optimal ITRs only focus on binary treatment settings. However, in practice, it often has applications with multiple treatment settings ([1]). To the best of our knowledge, few research attempts settings with more than two treatments. In this section, we consider a K-armed treatment setting, where $A \in \mathcal{A} = \{1, \cdots, K\}$. For simplicity, we focus on the class of linear decision rules to approximate the optimal ITR while the extension to nonlinear setting can be derived similarly. In addition, the notations and assumptions remain the same as before.

Expanding the equation (2.1), for K treatments we can get:

$$V(d) = \mathbf{E}[\sum_{k=1}^{K} \mathbf{E}[R|\mathbf{X}, A = k]\mathbb{I}(d(\mathbf{X}) = k)].$$

Similar to the binary treatment setting, the optimal ITR is given by

$$d_0(\mathbf{X}) = \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} \mathbf{E}[R|\mathbf{X}, A = k]. \tag{2.14}$$

Without changing the order of each element, we can further write the optimal ITR as

$$\begin{aligned} d_0(\mathbf{X}) &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} K\mathbf{E}[R|\mathbf{X}, A = k] - \sum_{i=1}^{K} \mathbf{E}[R|\mathbf{X}, A = i] \\ &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} \sum_{i\neq k}^{K} \{\mathbf{E}[R|\mathbf{X}, A = k] - \mathbf{E}[R|\mathbf{X}, A = i]\} \\ &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} \sum_{i\neq k}^{K} \mathbf{E}[\frac{RA_{ki}}{\pi_{ki}(A_{ki}, \mathbf{X})}|\mathbf{X}, A = k \text{ or } i] \\ &:= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} \sum_{i\neq k}^{K} f_{ki}(\mathbf{X}) := \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} f_k(\mathbf{X}), \end{aligned} \tag{2.15}$$

where A_{ki} is a binary random variable with $\pi_{ki}(1, \mathbf{X}) = \mathbf{P}[A = k|\mathbf{X}, A = k \text{ or } i] = \frac{\pi(k,\mathbf{X})}{\pi(k,\mathbf{X})+\pi(i,\mathbf{X})}$ and $\pi_{ki}(-1, \mathbf{X}) = \mathbf{P}[A = i|\mathbf{X}, A = k \text{ or } i] = \frac{\pi(i,\mathbf{X})}{\pi(i,\mathbf{X})+\pi(k,\mathbf{X})}$. For each $f_{ki}(\mathbf{X})$, where $k, i = 1, \cdots, K, i \neq k$, it is equivalent to a binary setting,

where we can use previously proposed D-learning method for estimation. The function $f_k(\mathbf{X})$, which we name as the *effect measure* of treatment k, is the sum of pairwise decision functions between the k-th treatment over other treatments.

One of the key differences between our method and fitting a regression model with treatment-by-covariate interactions is that D-learning avoids modeling the main effect functions, but directly targets on decision functions. The difference is significant because of the mismatch between minimizing the prediction error and maximizing the value function in finding optimal ITRs. Model-based methods are focused on predicting responses, i.e., minimizing the prediction error. In contrast, although our proposed D-learning also has the regression form, it directly targets on maximizing the value function. This difference can be more substantial for nonlinear learning due to the potential over-fitting.

In practice, among K treatments, one may be interested in learning the relative strengths of all treatments besides the estimated ITR. For example, if two treatments have similar effect measure, the doctor and patient may decide to use the suboptimal one if it costs less or has less potential side effects. Thus, our effect measure provides additional useful information besides the ITR for precision medicine. If there exits multiple treatment options producing the same largest expected treatment effect, other possible outcomes besides the current outcome may be considered for decision making. For example, assume that we also observe $\mathbf{Z}$, the side effect after a patient receives a treatment. Similar to the treatment effect measure $f_k(\mathbf{X})$ in our proposed D-learning, we can estimate the side effect measure of each treatment, such as $g_k(\mathbf{X})$, for each patient by the multi-category D-learning method. Then we can use $f_k(\mathbf{X}) + \tau g_k(\mathbf{X})$ to decide the best treatment for each patient, where τ is used to balance these two outcomes and can be decided by doctors. In theory, if multiple treatments have the same largest expected outcome, these treatments are equivalent in terms of consistency. However, one can pick a treatment among these equivalent ones using other factors, such as costs, etc.

Our proposed multi-category D-learning first estimates $f_{ki}(\mathbf{X})$ separately for $1 \le i < k \le K$. Then the decision rule is based on the maximum of estimated effect measures $\hat{f}_{k(n)}(\mathbf{X})$ for $k = 1, \cdots, K$ according to (2.15). Here the subscript (n) means that the estimated function depends on the sample size n. We would like to point out that it is sufficient to estimate $\frac{K(K-1)}{2}$ decision functions because $f_{ij}(\mathbf{X}) = -f_{ji}(\mathbf{X})$ by definition.

There is a close relationship between our multi-category D-learning and the standard majority vote One-vs-One (OVO) multi-category classification method. For K-category classification, the OVO multi-category classification method needs to build $\frac{K(K-1)}{2}$ classifiers, each of which distinguishes a pair of classes i and j for $i, j = 1, \cdots, K$. If the corresponding classifier g_{ij} for the (i, j) pair is positive, then we assign 1 to class i and -1 to class j, otherwise reverse the assignment. Then the final classifier $g = \text{argmax}_i \sum_{j=1}^{K} \text{sign}(g_{ij})$. The key difference is that this OVO multi-category classification approach uses the majority vote to assign the class label while multi-category D-learning is based on the effect measure as the cumulative pairwise contrast value among all

treatments. Note that if we replace $\sum_{i\neq k}^{K} f_{ki}(\mathbf{X})$ by $\sum_{i\neq k}^{K} \text{sign}(f_{ki}(\mathbf{X}))$ in (2.15), then it becomes similar as the majority vote OVO classification. However, by our derivation, the majority vote is not suitable for D-learning.

As a remark, from (2.15) we observe that $\sum_{k=1}^{K} f_k(\mathbf{X}) = 0$, which implies that one of the effect measures is redundant. This is similar as the sum-of-zero constraint in simultaneous multi-category classification methods, where the constraint is used to guarantee the consistency of classifiers ([34]).

2.4. Tuning parameter selection

For existing methods such as l_1-PLS, OWL and RWL, the tuning parameter is selected via cross-validation, mainly based on maximizing the empirical value function on validation data defined as

$$\hat{V}(d) = \frac{\mathbf{E}_n[R\mathbb{I}(A = d(\mathbf{X}))/\pi(A, \mathbf{X})]}{\mathbf{E}_n[\mathbb{I}(A = d(\mathbf{X}))/\pi(A, \mathbf{X})]}, \tag{2.16}$$

where $\mathbf{E}_n$ denotes the empirical average.

For D-learning, since it directly estimates the decision boundary, we consider an alternative way to select the tuning parameter λ. In particular, we select the choice of λ which minimizes the mean square error (MSE) on the validation data. For example, in the binary treatment setting, it is defined as

$$MSE(\hat{f}) = \mathbf{E}_n[(2RA - \hat{f})^2],$$

where $\hat{f}$ is the estimated decision boundary function. Since it is the same procedure as standard regression techniques such as LASSO, it can be easily implemented via standard software package such as glmnet ([11]) in R software to estimate the optimal ITR. We would like to point out that unlike tuning criterion (2.16) for previous methods, both our model building and tuning match in the sense that they both use the least square loss.

3. Theoretical properties of D-learning

In this section, we establish the estimation consistency of D-learning and obtain a value reduction bound of the estimated ITR from the optimal ITR. Fast convergence rates can be achieved under some reasonable assumptions. We consider linear and nonlinear D-learning in Section 3.1 and 3.2 respectively.

3.1. Consistency and value reduction bounds under linear decision rules

We first state the generalized Margin condition (gMC) used in our theoretical results.

Assumption (gMC)**.** For any $\epsilon > 0$, there exists some constants $C > 0$ and $\alpha > 0$ such that

$$P(|f_i(\mathbf{X}) - f_j(\mathbf{X})| \le 2\epsilon) \le C\epsilon^{\alpha}, \tag{3.1}$$

for $i, j = 1, \cdots, K$.

Theorem 1. *For the estimated effect measures $\hat{f}_{k(n)}(\mathbf{X})$ for $i = 1, \cdots, k$ and the corresponding ITR $\hat{d}_n$ by multi-category D-learning, we have*

$$V(d_0) - V(\hat{d}_n) \le \frac{K-1}{K} \sum_{k=1}^{K} (E||f_k(\mathbf{X}) - \hat{f}_{k(n)}(\mathbf{X})||_2^2)^{\frac{1}{2}}. \tag{3.2}$$

Furthermore, if we assume the gMC holds, then we could improve the rate as

$$V(d_0) - V(\hat{d}_n) \le C'(K)(\sum_{k=1}^{K} \mathbf{E}[||f_k(\mathbf{X}) - \hat{f}_{k(n)}(\mathbf{X})||_2^2])^{\frac{1+\alpha}{2+\alpha}}, \tag{3.3}$$

where C' depends on the constant C, α and K.

Remark 1. *Theorem 1 provides an approach to bound the reduction in the value function by the prediction error $\sum_{k=1}^{K} E||f_k(\mathbf{X}) - \hat{f}_{k(n)}(\mathbf{X})||_2$. The gMC (3.1) characterizes the behavior of the distribution of the decision function near 0. If the gMC holds, then the larger α is, the larger the exponent $\frac{1+\alpha}{2+\alpha}$ is, and the corresponding shaper upper bound in (3.3) can be achieved. In addition, if the data are fully separated in the sense that there exists an $\epsilon > 0$, one can have*

$$P(|f_i(\mathbf{X}) - f_j(\mathbf{X})| \le 2\epsilon) = 0, \tag{3.4}$$

for $i, j = 1, \cdots, K$. Then we could further improve the bound as

$$V(d_0) - V(\hat{d}_n) \le \frac{1}{2\epsilon} \frac{K-1}{K} \sum_{k=1}^{K} E||f_{k(n)}(\mathbf{X}) - \hat{f}_k(\mathbf{X})||_2^2. \tag{3.5}$$

For the special case when $K = 2$, Theorem 1 implies

$$V(d_0) - V(\hat{d}_n) \le (E||f_0(\mathbf{X}) - \hat{f}_{0(n)}(\mathbf{X})||_2^2)^{\frac{1}{2}}, \tag{3.6}$$

where $f_0(\mathbf{X})$ and $\hat{f}_{0(n)}(\mathbf{X})$ are the optimal decision function and corresponding estimation as defined in the previous section. With the gMC assumption, we have

$$V(d_0) - V(\hat{d}_n) \le C'(E[||f_0(\mathbf{X}) - \hat{f}_{0(n)}(\mathbf{X})||_2^2)^{\frac{1+\alpha}{2+\alpha}}, \tag{3.7}$$

where C' depends on the constant C and α.

Theorem 1 indicates that our D-learning in minimizing the empirical prediction error can estimate the optimal ITR. In particular, for linear D-learning, we consider $\mathcal{F} := \{f(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathcal{R}^p\}$ to be the class of linear functions to approximate $f_0(\mathbf{X})$. For the remaining theoretical results, without loss of generality, we assume that $\pi(A, \mathbf{X}) = \frac{1}{2}$ for any $A \in \mathcal{A}, \mathbf{X} \in \mathcal{X}$ and $K = 2$. The resulting prediction error consists of the approximation error and the estimation error since we do not require the true function $f_0(\mathbf{X})$ lies in $\mathcal{F}$.

Before we characterize the prediction error bound, we say a subset $S \subseteq \{1, 2, \cdots, p\}$ satisfies the compatibility condition ([3]), if for some constant $\phi(S) > 0$ and for all $\boldsymbol{\beta} \in \mathcal{R}^p$, with $||\boldsymbol{\beta}_{S^c}||_1 \leq 3||\boldsymbol{\beta}_S||_1$, we have

$$||\boldsymbol{\beta}_S||_1^2 \leq (\boldsymbol{\beta}^T \hat{\Sigma} \boldsymbol{\beta})|S|/\phi^2(S), \tag{3.8}$$

where $|S|$ is the cardinality of S, S^c is the complement of set S, and the j-th element of $\boldsymbol{\beta}_S$ denoted by $\beta_{S,j} = \beta_j \mathbb{I}(j \in S)$. In addition, we use $\mathbf{X}_n$ to denote the design matrix with n samples. Then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}_n^T \mathbf{X}_n$. Furthermore, we let Ω be the collection of index sets S satisfying the compatibility condition. We provide the finite sample bound for prediction errors.

Before establishing the error bound for the proposed D-learning, we state our assumptions below:

(i) Assume $R_i = m(\mathbf{X}_i) + A_i\delta(\mathbf{X}_i) + W_i$ for $i = 1, \cdots, n$. Let $\epsilon_i = 2R_iA_i - \mathbf{E}[\frac{R_iA_i}{\pi(A_i,\mathbf{X}_i)}|\mathbf{X}_i] = 2m(\mathbf{X}_i)A_i + 2W_iA_i$; $i = 1, ..., n$.

(ii) Assume the main effect is the linear function with $m(\mathbf{X}_i) = \mathbf{X}_i^T\boldsymbol{\gamma}_0$, where $||\boldsymbol{\gamma}_0||_2 \leq O(\sqrt{\log(2p)})$.

(iii) For any $\mathbf{X} \in \mathcal{X}$, there exists some constant a, such that $\max_{1\leq i\leq p} |X_i| \leq a$.

(iv) For the design matrix $\mathbf{X}_n$, there exists $\rho > 0$, such that

$$\boldsymbol{\gamma}^T \hat{\Sigma} \boldsymbol{\gamma} \leq \rho||\boldsymbol{\gamma}||_2^2,$$

for any vector $\boldsymbol{\gamma}$.

(vi) The compatibility condition (3.8) holds for all $S \in \Omega$.

(vii) Assume $||\mathbf{X}_n\boldsymbol{\beta}^* - f^0||_2^2/n \leq \lambda||\hat{\boldsymbol{\beta}}_{S_*} - \boldsymbol{\beta}^*_{S_*}||_1$, where $S^* = \{j : \beta_j^* \neq 0\}$, $\hat{\beta}_{j,S_*} = \hat{\beta}_j \mathbb{I}(j \in S_*)$, and λ is the tuning parameter in (2.9).

Define the oracle $\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}:S_{\boldsymbol{\beta}}\in\Omega}\{||\mathbf{X}_n\boldsymbol{\beta} - f^0(\mathbf{X}_n)||_2^2/n + \frac{4\lambda^2 s_{\boldsymbol{\beta}}}{\phi^2(S_{\boldsymbol{\beta}})}\}$. Then we have following theorem.

Theorem 2. *Assume assumptions (i)–(vii) hold. For $t > 0$, let the tuning parameter be*

$$\lambda = 16\sqrt{2}t^2\sqrt{\frac{log^2(2p)}{n}}. \tag{3.9}$$

Then for the constant C_1 which depends on the constants a, ρ and σ^2 and t sufficiently large, with the probability at least $1 - \frac{C_1}{t^2}$, we have

$$\frac{2||\mathbf{X}_n\hat{\boldsymbol{\beta}} - f_0||_2^2}{n} + 3\lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq 6||\mathbf{X}_n\boldsymbol{\beta}^* - f_0||_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2}. \tag{3.10}$$

Remark 2.
(1) The first term in right hand side of (3.10) is the approximation error to the true function f_0. The second term is the estimation error decided by sparsity s_*, the compatibility constant ϕ_* and the tuning parameter λ. Furthermore, if we assume the underlying function is linear, i.e. $f^0 = \mathbf{X}^T\boldsymbol{\beta}^*$, then the right hand side of (3.10) can be reduced to $\mathbf{O}(\frac{\log^2(2p)}{n})$ by setting $\lambda = 16\sqrt{2}t^2\sqrt{\frac{log^2(2p)}{n}}$.

(2) The main difference between our theoretical result and those in the usual LASSO lies in the unequal variance of the error term for each observation. Thus assumptions (ii)–(iv) are needed to bound the tail of the error term. Specifically, we bound the prediction error using the Nemirovski moment inequality ([31]) under some mild assumptions on the baseline function $m(\mathbf{X})$. The price we need to pay for unequal variances among the observations is that the tuning parameter needs to be larger, i.e, $\mathbf{O}(\sqrt{\frac{log^2(p)}{n}})$ compared with $\mathbf{O}(\sqrt{\frac{\log(p)}{n}})$ in the LASSO theory. In addition, the probability bound for (3.10) converges to 1 in a polynomial rate, slower than the exponential rate in the theory of LASSO.

Theorem 3. *Suppose the true decision function is linear, and let the tuning parameter be the one defined in Theorem 2. If the gMC and assumptions in Theorem (2) hold, then with probability at least $1 - \frac{C_1}{t^2}$, with C_1 depending on a, Φ and σ^2, we have*

$$V(d^*) - V(\hat{d}) \leq C_2 (\frac{\log(2p)}{n})^{\frac{1+\alpha}{2+\alpha}}, \tag{3.11}$$

for the constant C_2 determined by the gMC constant, t, ϕ_*, and s_*.

Remark 3. *The inequality (3.11) gives us the value reduction bound between the optimal ITR and the estimated ITR as $\mathbf{O}((\frac{\log(2p)}{n})^{\frac{1+\alpha}{2+\alpha}})$. If the dimension p does not grow exponentially faster than the sample size n, then as $n \to \infty$, $V(\hat{d}) \to V(d^*)$. If we fix the dimension p, the convergence rate is at least of order $n^{-\frac{1}{2}}$, i.e., $\alpha = 0$. If the data are completely separable, then by remark 1, the fast convergence rate of order $\frac{1}{n}$ can be achieved. Our results are consistent with those derived by [24].*

3.2. Value reduction bounds under nonlinear decision rules

In this section, we establish the value reduction bounds for the nonlinear Model (2.11). For Model (2.13), we refer to [19] for the finite sample error bounds under the fixed design setting. We assume the outcome R is bounded, i.e., $|R| \leq C_0$ almost surely for some constant C_0 and recall that we let $\pi(x, a) = \frac{1}{2}$ as discussed in Section 3.1.

Before proceeding our results, we need the following definitions.

Definition 1. *Consider $\mathcal{F}$ to be a class of real value measurable functions $f : \mathcal{Z} \to \mathbb{R}$. The Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}_n(\mathcal{F}) := \mathbf{E}[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i)], \tag{3.12}$$

where $Z_1, \cdots, Z_n$ are drawn i.i.d from some probability distribution P_Z and Rademacher random variables $\sigma_1, \cdots, \sigma_n$ are drawn i.i.d from a uniform distribution over $\{1, -1\}$.

The corresponding empirical Rademacher complexity of $\mathcal{F}$ is defined as

$$\hat{\mathcal{R}}_n(\mathcal{F}) := \mathbf{E}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)|Z_1,\cdots,Z_n], \tag{3.13}$$

where we can see that $\mathbf{E}[\hat{\mathcal{R}}_n(\mathcal{F})] = \mathcal{R}_n(\mathcal{F})$.

Let $Y = 2RA$ and correspondingly $Y_i = 2R_iA_i$, for $i = 1,\cdots,n$. Define $L(f) = \mathbf{E}[D(f)]$, where $D(f) = (Y - f(\mathbf{X}))^2$ and let $f_{\lambda_n} = \text{argmin}_{f\in\mathcal{H}} L(f) + \frac{\lambda_n}{2}||f||^2_{\mathcal{H}}$, where $\mathcal{H}$ is RKHS. Then $\mathcal{A}(\lambda_n) := L(f_{\lambda_n}) + \frac{\lambda_n}{2}||f_{\lambda_n}||^2_{\mathcal{H}} - L(f_0)$ is consider to be the approximation error. The empirical version of $L(f)$ is defined as $L_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2$.

According to Theorem 1, we need to establish bounds for the prediction error under Model (2.11) in order to get the value reduction bound for our estimator. Note that

$$\begin{aligned} L(f) - L(f_0) &= \mathbf{E}[Y - f(\mathbf{X})]^2 - \mathbf{E}[Y - f_0(\mathbf{X})]^2 \\ &= -2\mathbf{E}[Yf(\mathbf{X})] + \mathbf{E}[f^2(\mathbf{X})] + 2\mathbf{E}[Yf_0(\mathbf{X})] - \mathbf{E}[f_0^2(\mathbf{X})] \\ &= -2\mathbf{E}[f_0(\mathbf{X})f(\mathbf{X})] + \mathbf{E}[f^2(\mathbf{X})] + \mathbf{E}[f_0^2(\mathbf{X})] \\ &= \mathbf{E}[f_0(\mathbf{X}) - f(\mathbf{X})]^2, \end{aligned} \tag{3.14}$$

where the third equality is based on the definition of $f_0(\mathbf{X})$ in Lemma 1. Thus bounding the prediction error is equivalent to bounding the excess risk: $L(\hat{f}) - L(f_0)$, where $\hat{f}$ is the estimator under Model (2.11). Based on statistical learning theory, we have the following lemma.

Lemma 2. *For any distribution P over $(\mathbf{X}, A, R)$ with $|R| \le C_0$, if we use bounded kernels in Model (2.11), then with probability at least $1-\epsilon$, we can have*

$$L(\hat{f}) - L(f_0) \le 4M_3\mathcal{R}_n(\Pi) + M_2\sqrt{\frac{8\log(\frac{1}{\epsilon})}{n}} + \mathcal{A}(\lambda_n), \tag{3.15}$$

where $\Pi = \{f|\ f \in \mathcal{H}_1, \frac{\lambda_n}{2}||f||^2_{\mathcal{H}} \le M_1\}$, for some constants M_1, M_2 and M_3.

Remark 4. *This lemma quantifies the excess risk by two parts: the estimation error and the approximation error. The estimation error corresponds to the first two terms of the right hand side in (3.15). The approximation error $\mathcal{A}(\lambda_n)$ is controlled by the tuning parameter λ_n.*

Theorem 4. *Consider $\mathcal{H}_K$ in Model (2.11) to be RKHS with the Gaussian kernel and suppose assumptions in Lemma 2 hold. If $\mathcal{A}(\lambda_n) \le C_1\lambda_n^{\omega}$, where $\omega \in (0,1]$, then by choosing $\lambda_n = \mathbf{O}(n^{-\frac{1}{2\omega+1}})$, we have*

$$L(\hat{f}) - L(f_0) \le c_1 n^{-\frac{\omega}{2\omega+1}},$$

with probability at least $1-\epsilon$, for some constant c_1 that is independent of n and decreasing in ϵ.

Remark 5. *This theorem shows that the excess risk converges to* 0 *in probability under some conditions. The upper bound assumption on the approximation error* $\mathcal{A}(\lambda_n)$ *is standard in the statistical learning literature such as [29] to derive the convergence rate. Replacing this upper bound assumption by the assumption that* $\mathbf{X}$ *belongs to a compact metric space, convergence of the excess risk to 0 can still be established by the universal consistency property of the Gaussian kernel ([28]).*

Corollary 4.1. *Let the tuning parameter be the one defined in Theorem 4. If all the assumptions in Lemma 2 and Theorem 4 hold, and also gMC holds, then with probability at least* $1-\epsilon$*, we have*

$$V(d^*) - V(\hat{d}) \leq C_3 n^{-\frac{\omega(1+\alpha)}{(2\omega+1)(2+\alpha)}} \tag{3.16}$$

for some constant C_3*, which is determined by the gMC constant and* c_1 *in Theorem 4.*

Remark 6. *Corollary 4.1 gives us the convergence rate of the value reduction bound for Model (2.11). The result does not require strong model assumptions, which can further demonstrate the robustness of our nonlinear D-learning method.*

4. Simulation study

In this section, we conduct extensive simulation studies to investigate D-learning's finite sample performance. The outcome is generated following Model (2.7) with $W \sim \mathcal{N}(0, \sigma^2)$, where σ is set to be 1 or 4. Each covariate is independently generated by a uniform distribution between -1 and 1. The randomized treatment A follows a uniform distribution among treatments. The number of replications is 100 times. We evaluate D-learning under both linear and nonlinear settings in Sections 4.1 and 4.2 for binary treatments, and in Section 4.3 we consider the multi-category setting.

4.1. Linear decision boundary study

We consider the sample size $n = 100, 400$ and the dimension p from 30 to 1920 increased by a factor of 4. We compare D-learning with the following three methods:

(1) l_1-PLS by [24] with basis function $(1, \mathbf{X}, A, \mathbf{X}A)$;
(2) OWL by [37] with linear kernel;
(3) RWL by [39] with linear kernel.

The following four linear boundary scenarios are considered:

(1) $m(\mathbf{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4,\ \delta(\mathbf{X}) = 1.8(0.3 - x_1 - x_2)$;
(2) $m(\mathbf{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4,\ \delta(\mathbf{X}) = 0.442|x_3|(1 - x_1 - x_2)$;
(3) $m(\mathbf{X}) = 1 + 2x_1 + x_2 + 0.5x_3,\ \delta(\mathbf{X}) = 0.3(0.9 - x_1)$;
(4) $m(\mathbf{X}) = 1 + x_1^2 + x_2^2,\ \delta(\mathbf{X}) = 1.8(0.3 - x_1 - x_2)$.

The first scenario was considered in [39]. The second scenario corresponds to the situation that x_3 interacts with the treatment but does not affect the treatment selection strategy. The third scenario represents the situation where prescriptive variables have weak power to predict the outcome but are vital for decision making, in terms of the effective size (ES) defined in [4]

$$ES = \frac{|\mathbf{E}[R|A=1] - \mathbf{E}[R|A=-1]|}{\sqrt{Var[R|A=1] + Var[R|A=-1]}}.$$

We can check that the effective size in the third scenario is small (less than 0.2). Thus it is hard to estimate the optimal ITR correctly, especially for the l_1-PLS method. The fourth scenario considers the nonlinear main effect $m(\mathbf{X})$ to evaluate the robustness of D-learning. In all four scenarios, optimal ITRs depend on the first two covariates. We use 10-fold cross-validation to select the tuning parameter based on the empirical value function over the validation dataset. We evaluate all the methods based on two criteria. The first criterion is the misclassification rate of the estimated ITR from the optimal ITR based on the independently generated test data. The second criterion is the empirical value function of the estimated ITR on the test data via (2.16). Specifically, we generate $10,000$ independent test data to assess the performance based on these two criteria.

Figure 1 shows the misclassification rates for all four scenarios when $n = 100$ and $\sigma = 1$. Additional simulation results are provided in the appendix. Based on the results, we can conclude that for all situations, D-learning has competitive low misclassification error rates than the other three methods, especially when the noise is large. Specifically, for the first linear scenario, D-learning has comparable performance with l_1-PLS but performs better than OWL and RWL in low dimensions. When the dimension gets higher, D-learning performs the best among all methods. One potential reason of the competitive performance of our proposed D-learning is the effective prescriptive variable selection in high dimensional settings. For the second scenario, our method performs better than l_1-PLS because our proposed method can effectively identify prescriptive variables that have qualitative interactions with the treatment. For the third scenario, l_1-PLS performs worse than D-learning and RWL due to the mismatch between prediction and ITR estimation in l_1-PLS. The interaction effect term in this example has little power in prediction. For the last scenario, since the main effect is non-linear, RWL performs worse than D-learning and l_1-PLS because of the improper residual calculation due to the model misspecification. Our proposed D-learning does not need to modify the weights. Figure 2 corresponds to the empirical value functions of estimated ITRs on test data for all four scenarios. D-learning performs the best among most scenarios. Finally, we compare D-learning with l_1-PLS in selecting true prescriptive variables using the average False Positive (FP) and False Negative (FN). Table 1 shows that D-learning has comparable small average FNs as l_1-PLS but much smaller average FPs than l_1-PLS for both low and high dimensional problems. We conclude that D-learning can better identify true variables of the optimal ITR for these examples.

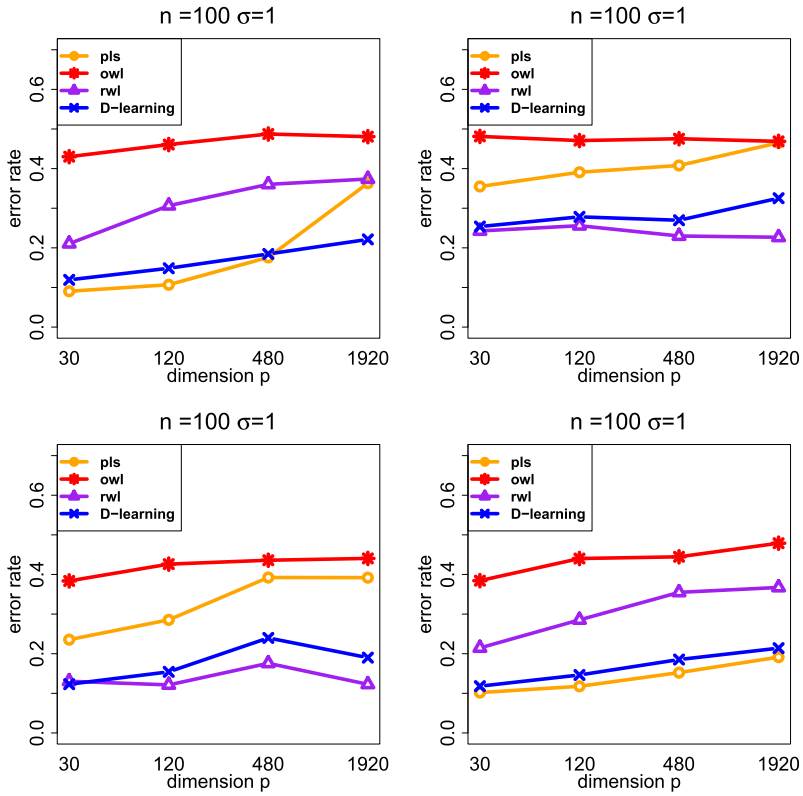

FIG 1. *Comparison of misclassification error rates for simulated examples* $n = 100$ *and* $\sigma = 1$ *on the test data. From left to right, each represents scenarios (1)–(4) respectively. The y-axis denotes the misclassification error rates for four methods and the x-axis is the dimension p from* 30 *to* 1920 *increased by a factor 4. Overall, D-learning performs the best compared to other three methods.*

4.2. Nonlinear decision boundary study

In this subsection, we evaluate D-learning when the true decision boundary is nonlinear. We consider the sample size to be $n = 100, 400$ and dimension of covariates $p = 5, 50$. We compare our methods including linear D-learning, Gaussian kernel KKR D-learning, and Gaussian Kernel COSSO D-learning with the following three alternative methods:

(1) l_1-PLS with basis function $(1, \mathbf{X}, A, \mathbf{X}A)$;
(2) RWL with linear kernel;
(3) RWL with Gaussian kernel.

We consider the following two nonlinear boundary scenarios:

(1) $m(\mathbf{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4,\ \delta(\mathbf{X}) = 3.8(0.8 - x_1^2 - x_2^2)$;
(2) $m(\mathbf{X}) = 1 + x_1^2 + x_2^2 + x_3^2 + 0.5x_4^2,\ \delta(\mathbf{X}) = 1.3(x_2 - 2x_1^2 + 0.3)$.

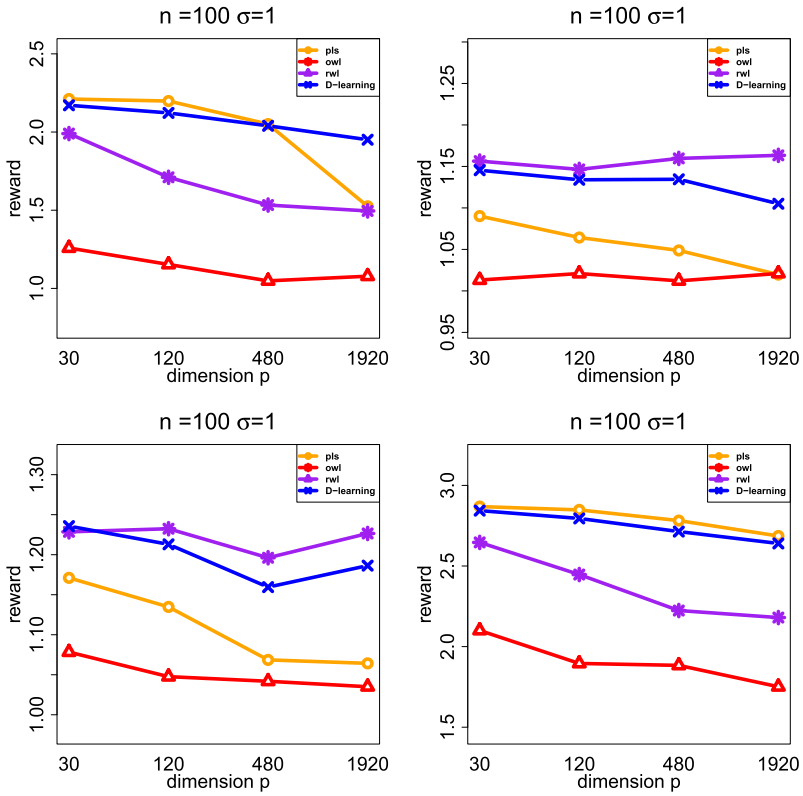

FIG 2. *Comparison of empirical value functions for simulated examples with* $n = 100$ *and* $\sigma = 1$ *on the test data. From left and right, each represents scenarios (1)–(4) respectively. The y-axis denotes the empirical value functions for four methods and the x-axis is the dimension p from* 30 *to* 1920 *increased by a factor 4. Overall, D-learning performs the best compared to other three methods.*

These examples were considered in [39]. The first scenario decision boundary is a parabola while the second scenario corresponds to a circle boundary. From Tables 2 and 3, we can see that, in general, nonlinear D-learning methods perform better than the other three methods, especially when the sample size is large. In addition, our linear D-learning performs competitively even the decision function $\delta(\mathbf{X})$ is misspecified. In practice, the choice of kernel functions can be viewed as part of the tuning parameter selection. One can use various kernel functions such as linear, polynomial and Gaussian kernels, and then select the best one using cross-validation. The optimal value is calculated via $\mathbf{E}^d[R]$ since we know the optimal decision rule in simulation studies.

4.3. Multi-category linear decision boundary study

In this subsection, we evaluate our proposed multi-category D-learning when the true decision boundary is linear. For comparison, we extend OWL and RWL

TABLE 1

Variable selection performance for four linear scenarios based on average (std) FNs (%) and FPs (%). The noise level σ is 1. The best average FNs and FPs are in bold.

	$n = 100, p = 30$			
	FN (l_1-PLS)	FN (D-learning)	FP (l_1-PLS)	FP (D-learning)
Scenario 1	0.01 (0.009)	**0.00 (0.00)**	10.79 (0.90)	**6.12 (0.45)**
Scenario 2	**0.85 (0.09)**	1.40 (0.07)	11.19 (0.86)	**2.34 (0.32)**
Scenario 3	**0.00 (0.00)**	**0.00 (0.00)**	8.32 (0.86)	**5.50 (0.38)**
Scenario 4	**0.51 (0.05)**	0.84 (0.04)	12.40 (0.98)	**1.97 (0.24)**
	$n = 100, p = 240$			
	FN (l_1-PLS)	FN (D-learning)	FP (l_1-PLS)	FP (D-learning)
Scenario 1	0.05 (0.03)	**0.03 (0.02)**	14.30 (0.79)	**9.81 (0.86)**
Scenario 2	1.80 (0.04)	**1.65 (0.06)**	12.72 (0.90)	**6.67 (0.98)**
Scenario 3	**0.03 (0.02)**	0.04 (0.02)	**8.32 (0.94)**	12.45 (0.89)
Scenario 4	0.97 (0.02)	**0.92 (0.03)**	12.58 (0.94)	**5.11 (0.96)**

TABLE 2

Results of average means (std) of empirical value functions and misclassification rates for two simulations of nonlinear scenarios with 5 covariates on the test data. The best value functions and minimal misclassification rates are in bold.

	$n = 100$		$n = 400$	
	Value	Misclassification	Value	Misclassification
Scenario 1 (Optimal value 1.96)				
l_1-PLS	1.69 (0.04)	0.23 (0.02)	1.72 (0.03)	0.22 (0.01)
KKR	1.48 (0.18)	0.31 (0.06)	1.60 (0.12)	0.26 (0.05)
RWL-Linear	1.64 (0.08)	0.25 (0.04)	1.73 (0.03)	0.24 (0.02)
RWL-Gaussian	1.73 (0.09)	0.21 (0.05)	1.88 (0.04)	0.11 (0.03)
D-learning	1.62 (0.1)	0.27 (0.05)	1.71 (0.05)	0.24 (0.02)
D-learning KKR	1.63 (0.1)	0.26 (0.05)	1.73 (0.07)	0.22 (0.04)
D-learning COSSO	**1.79 (0.12)**	**0.17 (0.08)**	**1.92 (0.04)**	**0.07 (0.03)**
Scenario 2 (Optimal value 3.88)				
l_1-PLS	2.96 (0.15)	0.38 (0.04)	3.02 (0.04)	0.37 (0.01)
KKR	3.01 (0.03)	0.37 (0.01)	3.07 (0.02)	0.37 (0.01)
RWL-Linear	3.12 (0.16)	0.35 (0.05)	3.27 (0.04)	0.31 (0.01)
RWL-Gaussian	**3.62 (0.13)**	**0.19 (0.05)**	3.81 (0.04)	0.11 (0.02)
D-learning linear	2.92 (0.17)	0.39 (0.07)	3 (0.06)	0.37 (0.02)
D-learning KKR	3.23 (0.1)	0.31 (0.03)	3.24 (0.1)	0.31 (0.03)
D-learning COSSO	3.61 (0.13)	0.30 (0.08)	**3.85 (0.07)**	**0.09 (0.03)**

methods to multi-category scenario by using the OVO multi-category classification scheme. Here we consider two versions. One is to use the majority vote OVO multi-category classification method, and the other is to use the procedure similar to the effect measure used for D-learning as discussed in Section 2.3. We name the first version as the *hard* OVO and the second one as the *soft* OVO. We consider the sample size to be $n = 800, 1600$ and dimension of covariates the same as the binary linear decision boundary study. Here we set σ to be 1

TABLE 3

Results of average mean (std) of empirical value functions and misclassification rates for two simulation of nonlinear scenarios with 50 covariates on the test data. The best value functions and minimal misclassification rates are in bold.

	$n = 100$		$n = 400$	
	Value	Misclassification	Value	Misclassification
Scenario 1 (Optimal value 1.96)				
l_1-PLS	**1.66 (0.06)**	**0.25 (0.03)**	1.71 (0.03)	0.23 (0.01)
KKR	0.98 (0.19)	0.5 (0.06)	1.02 (0.11)	0.48 (0.04)
RWL-Linear	1.40 (0.13)	0.37 (0.04)	1.60 (0.06)	0.29 (0.03)
RWL-Gaussian	1.38 (0.14)	0.38 (0.04)	1.62 (0.06)	0.28 (0.03)
D-learning linear	1.56 (0.09)	0.31 (0.04)	1.68 (0.07)	0.25 (0.03)
D-learning KKR	1.46 (0.08)	0.35 (0.02)	1.48 (0.03)	0.35 (0.01)
D-learning COSSO	1.48 (0.2)	0.32 (0.08)	**1.71 (0.09)**	**0.22 (0.05)**
Scenario 2 (Optimal value 3.88)				
l_1-PLS	2.84 (0.19)	0.42 (0.05)	3 (0.04)	0.37 (0.01)
KKR	3.01 (0.02)	0.37 (0.01)	3.07 (0.02)	0.37 (0.00)
RWL-Linear	2.88 (0.17)	0.40 (0.05)	2.99 (0.05)	0.38 (0.01)
RWL-Gaussian	2.86 (0.19)	0.42 (0.05)	2.95 (0.09)	0.40 (0.02)
D-learning linear	2.91 (0.2)	0.4 (0.05)	2.99 (0.07)	0.38 (0.02)
D-learning KKR	**2.93 (0.16)**	**0.39 (0.04)**	3.01 (0.03)	0.37 (0.01)
D-learning COSSO	2.78 (0.22)	0.43 (0.05)	**3.28 (0.32)**	**0.3 (0.09)**

and simulation results with large σ are in the appendix. We compare our multi-category D-learning with the following methods:

(1) l_1-PLS with basis function $(1, \mathbf{X}, A, \mathbf{X}A)$;
(2) hard OVO-OWL, soft OVO-OWL;
(3) hard OVO-RWL, soft OVO-RWL.

We consider the main model to be

$$R = m(\mathbf{X}) + \delta_1(\mathbf{X})\mathbf{1}(A = 1) + \delta_1(\mathbf{X})\mathbf{1}(A = 2) + \delta_1(\mathbf{X})\mathbf{1}(A = 3), \tag{4.1}$$

with the following two linear boundary scenarios

(1) $m(\mathbf{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$,
$\delta_1(\mathbf{X}) = (0.3 - x_1 - x_2),\ \delta_2(\mathbf{X}) = 0.2x_1 - x_2,\ \delta_3(\mathbf{X}) = 0$;
(2) $m(\mathbf{X}) = 1 + x_1^2 + x_2^2$,
$\delta_1(\mathbf{X}) = (0.3 - x_1 - x_2),\ \delta_2(\mathbf{X}) = 0.2x_1 - x_2,\ \delta_3(\mathbf{X}) = 0$.

The only difference between these two scenarios is the functional form of the main effect. Figure 3 shows the misclassification rates for these two scenarios with $n = 1600$ and $\sigma = 1$. Since it is often to have ties for the hard OVO-OWL and OVO-RWL, according to our simulation study, the corresponding results of the hard version are significantly worse than the soft version. We include the comparison between hard and soft methods and additional simulation results in the appendix. Based on the results, we can conclude that for both scenarios, multi-category D-learning has the lowest error rates than the other three methods. Figure 4 corresponds to the empirical value functions of the estimated ITR on the test data for these two scenarios. Our proposed multi-category D-learning

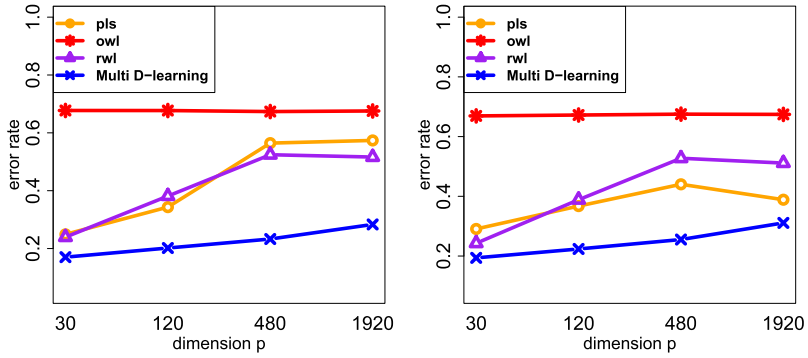

FIG 3. *Comparison of misclassification error rates for simulated examples with* $n = 1600$ *and* $\sigma = 1$ *on the test data. From left to right, each represents multi-category scenarios (1)–(2) respectively. The y-axis denotes the misclassification error rates for four methods and the x-axis is the dimension p from* 30 *to* 1920 *increased by a factor 4.*

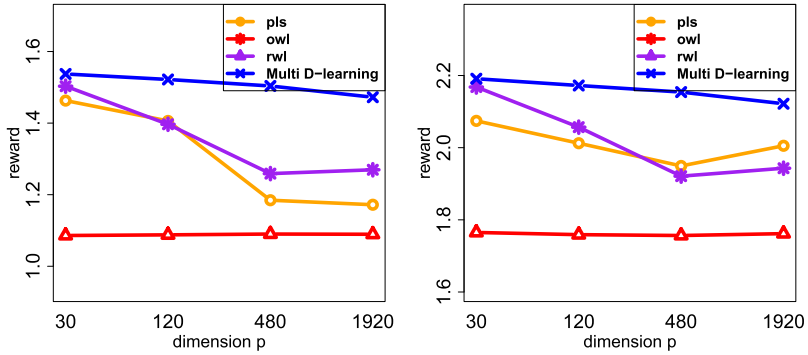

FIG 4. *Comparison of empirical value functions for simulated examples* $n = 1600$ *and* $\sigma = 1$ *on the test data. From left to right, each represents scenarios (1)–(2) respectively. The y-axis denotes the empirical value functions for four methods and the x-axis is the dimension p from* 30 *to* 1920 *increased by a factor 4.*

has the largest value functions among most scenarios. Finally, we compare the computational time of D-learning with all other three methods in Table 4. Due to the simplicity of D-learning, our proposed multi-category D-learning has the lowest computational cost in most settings. As the dimension gets large, our proposed multi-category D-learning is even faster than l_1-PLS since l_1-PLS needs to fit LASSO regression with $Kp + 2$ variables and multi-category D-learning only needs to fit $\frac{K(K-1)}{2}$ separate LASSO regression problems, each with $p + 1$ variables.

5. Applications to AIDS clinical data

In this section, we apply D-learning with linear kernel to the data from AIDS Clinical Trials Group (ACTG) 175 ([13]). This trial was designed to evaluate

TABLE 4
Results of the average means (std) of CPU computational time for two multi-category treatment settings. The lowest time costs are in bold.

	$n = 800$		$n = 1600$	
	$p = 30$	$p = 1920$	$p = 30$	$p = 1920$n
		Scenario 1		
l_1-PLS	**0.41(0)**	42.7(0.23)	0.45(0)	83.43(0.56)
soft OVO-OWL	35.84(0.27)	643.6(12.36)	102.21(0.59)	2493.02(47.6)
soft OVO-RWL	55.48(0.56)	340.76(7.02)	43.11(0.43)	556.96(10.23)
Multi D-learning	0.44(0)	**10.56(0.05)**	**0.11(0)**	**32.71(0.22)**
		Scenario 2		
l_1-PLS	**0.4(0)**	41.24(0.26)	0.43(0)	79.48(0.42)
soft OVO-OWL	51.32(0.35)	1021.66(19.85)	152.19(0.83)	3786.63(62.21)
soft OVO-RWL	59.5(0.65)	398.88(8.31)	45.86(0.42)	667.63(8.82)
Multi D-learning	0.44(0)	**10.82(0.07)**	**0.11(0)**	**33.5(0.25)**

whether a single treatment of HIV infection is worse than the combination treatments based on the counts of CD4+ T cells of patients. In this trial, 2139 patients with HIV infection were randomly assigned to four treatment groups with the same probability: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine (ZAL), and ddI monotherapy.

We choose the difference between early stage CD4+ T(cells/mm^3) cell amount and the baseline CD4+ T prior to trial as the clinical outcome. The larger this change is, in general the better condition the patient is. In addition to the treatment, we consider 12 clinical covariates in our model as done in [21] and [9]. Five of these covariates are continuous: weight (kg), CD4+ T cells amount at baseline, CD8 amount at baseline(cells/mm^3), age (year) and Karnofsky score (scale at 0-100). The remaining seven covariates are binary: gender (0=female, 1=male), race (0=white, 1=non-white), homosexual activity (0 = no, 1 = yes), history of intravenous drug use (0=no, 1=yes), symptomatic status (0=asymptomatic, 1=symptomatic), antiretroviral history (0=naive, 1=experienced) and hemophilia (0=no, 1=yes). For the interpretablity of the decision rule, we consider to use linear D-learning to estimate ITR.

5.1. Pairwise comparison

Our first goal is to estimate the optimal ITR for the following four scenarios:

- Scenario 1: $A = 1$ for ZDV alone vesus $A = -1$ for the other three treatments, with $\pi(A = 1) = 0.25$;
- Scenario 2: $A = 1$ for ZDV + ddI vesus $A = -1$ for ZDV + Zal, with $\pi(A = 1) = 0.5$;
- Scenario 3: $A = 1$ for ZDV + ddI vesus $A = -1$ for ddI, with $\pi(A = 1) = 0.5$;

- Scenario 4: $A = 1$ for ZDV + Zal vesus $A = -1$ for ddI, with $\pi(A = 1) = 0.5$.

For scenario 1, the only non-zero coefficient of the estimated ITR by D-learning is the intercept with a negative number, implying that the other three treatments are better than ZDV.

For scenario 2, linear D-learning identifies three important prescriptive variables: age, homosexual activity, and baseline CD4+ T cell amount. The estimated ITR is sign(41.38+14.03×age−13.45×baseline CD4+ cell−9.55×homo), which is similar to the result in [21] but identifies one more variable: baseline CD4+ T cell amount. For young patients having homosexual activity experience with high baseline CD4 + T cell amount, the estimated ITR assigns them to the treatment ZDV + Zal. For old patients not having homosexual activity experience with low baseline CD4 + T cell amount, the estimated ITR assigns them to the treatment ZDV + ddI.

For scenario 3, D-learning identifies four important prescriptive variables: age, homosexual activity, baseline CD4+ T cell and the history of intravenous drug use. The estimated ITR is sign(37.19 + 7.94 × age − 25.20 × baseline CD4+ cell − 25.19 × homo + 33.10 × drug), which is similar to the results in [21] and [9] but identifies one more variable: history of intravenous drug use. For young patients having homosexual activity experience with the high baseline CD4 + T cell amount but without using intravenous drug before, the estimated ITR assigns them to the treatment ddI. For old patients not having homosexual activity experience with low baseline CD4 + T cell amount but did use intravenous drug before, the estimated ITR assigns them to the treatment ZDV + ddI.

For scenario 4, D-learning identifies one important prescriptive variable: history of intravenous drug use. The estimated ITR is sign(−10.43 + 9.45 × drug). However, the estimated ITR is always −1, which means the treatment ddI is always preferable to the treatment ZDV + Zal. The efficacy of these two treatments is similar when the patients have the history of using intravenous drug.

To compare D-learning with l_1-PLS, linear OWL and linear RWL, we report the empirical value functions in Table 5. We split the data into three folds and use two of them as training data and the remaining as test data. The procedure is repeated 1200 times.

From Table 5, we could see that D-learning has the highest empirical value function in the first two scenarios. In the last two scenarios, while D-learning is not the highest, it still performs well (second place). Overall, D-learning performs better than the other three methods in finding the optimal ITR.

5.2. Overall comparison

So far, we have considered several binary problems to find the ITRs for each pair of treatments separately. Compared with several alternative methods, the results show that D-learning gives us better decision boundaries for choosing ITRs between two treatments in order to maximize the outcome. Our second goal is to estimate the optimal ITR for considering four treatments simultaneously.

TABLE 5
Results of empirical value functions on test data for four scenarios of AIDS data. The best empirical value function in each scenario is in bold.

	l_1-PLS	OWL	RWL	D-learning
Scenario 1	33.32 (0.12)	24.93 (0.16)	34.61 (0.13)	**36.15 (0.13)**
Scenario 2	53.05 (0.27)	41.52 (0.26)	52.38 (0.26)	**53.66 (0.25)**
Scenario 3	54.82 (0.26)	42.03 (0.26)	**57.83 (0.26)**	56.04 (0.26)
Scenario 4	24.99 (0.30)	**28.97 (0.19)**	22.79 (0.30)	25.7 (0.21)

TABLE 6
Results of empirical value functions on the test data. The best empirical value function is in bold.

l_1-PLS	51.82 (0.33)	hard OVO-RWL	21.32(1.03)
hard OVO-OWL	20.36(0.26)	soft OVO-RWL	52.55 (0.29)
soft OVO-OWL	24.28 (0.37)	Multi D-learning	**55.95 (0.26)**

TABLE 7
Results of estimated coefficients for comparison functions.

Variable Name (1-7)	ZDV ($f_1(\mathbf{X})$)	ZDV+ddI ($f_2(\mathbf{X})$)	ZDV+Zal($f_3(\mathbf{X})$)	ddI ($f_4(\mathbf{X})$)
Intercept	-165.26	163.63	-17.28	18.91
Age	-8.24	30.89	-14.38	-8.27
Weight	0	0	0	0
Karnofsky Score	-3.66	3.66	0	0
CD4 baseline	19.78	-57.71	11.33	26.59
Days pre-anti-retroviral therapy	0	0	0	0
Hemophilia	0	0	0	0
Homosexual activity	4.92	-41.00	10.44	25.64
History of drug use	-19.48	53.03	19.09	-52.64
Race	16.53	-16.53	0	0
Gender	10.01	-10.01	0	0
Antiretroviral history	0	0	0	0
Symptomatic indicator	0	0	0	0

We first compare our proposed multi-category D-learning with l_1-PLS, linear OVO-OWL and linear OVO-RWL based on the empirical value function. Following the same procedure as before, we random split the data into three folds and use two of them as the training data and the remaining as the test data. The procedure is repeated 1200 times. The results are shown in Table 6. We can see that multi-category D-learning has an significant advantage over other three methods based on the value function. For further investigations, we report coefficient estimation of linear comparison functions $\hat{f}_k$, where $k = 1, \cdots, 4$ in Table 7. Besides the important prescriptive variables identified in Section 5.1, multi-category D-learning identifies three more variables: Karnofsky score (scale at 0-100), gender and race, which play a moderate role in deciding the decision rule because the absolute values of such coefficients are not large. Compared with the coefficient estimation of prescriptive variables identified in Section 5.1, the sign of multi-category D-learning coefficient estimation for those prescriptive variables is consistent with the previous study. According to the previous study by [13], treatment with ZDV + ddI, ZDV + Zal or ddI alone slows the pro-

gression of HIV disease and is superior to the treatment with ZDV alone. This is consistent with our findings. As found in [14], Zal has most serious adverse event among other two treatments so it is general not recommended. Our finding also supports this suggestion because the absolute value of coefficient estimation of $f_3(\mathbf{X})$ is relatively small compared with others. In addition, according to VIDEX (didanosine) released by Federal Drug Administration (FDA) http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/020154s50,20155s39,20156s40,21183s16lbl.pdf, ddI has a significant interaction with allopurinol, which is sold by IV drug and tablet. Thus one should avoid using them together. The large negative coefficient estimation of IV drug in $f_4(\mathbf{X})$ by D-learning also supports this recommendation.

6. Discussion

In this article, we propose a direct learning (D-learning) method to estimate the optimal ITR by reformulating the optimal decision function. D-learning is very simple and flexible with highly competitive performance.

The goal of D-learning is to estimate the optimal ITR by combining the advantage of model-based methods such as l_1-PLS and classification-based methods such as RWL and OWL. Because of its direct formulation, D-learning is robust to potential model misspecification, and can unify the goal of prediction and identification of the optimal ITR. Moreover, D-learning can be applied to multiple treatment settings. The proposed D-learning methods indeed can achieve better performance compared with several existing methods based on the simulation and real data studies.

In this paper, we consider randomized clinical studies where propensity score $\pi(A, \mathbf{X})$ is assumed to be known. Although it can be estimated by statistical models such as multinomial logistic regression in observational studies, it may suffer from potential model misspecification in estimating propensity scores. This limitation may be addressed by the results in [35]. In their work, by using the augmented inverse probability weighting strategy, their methods have doubly robust properties for observational studies. In particular, either the correct postulated regression model or correct propensity score model gives the consistency results. Another possible approach to relieve the risk of misspecification to propensity score estimation in our proposed methods is to use nonparametric methods, such as regression trees, to estimate $\pi(A, \mathbf{X})$.

Several possible extensions can be explored for future study. The current framework of D-learning focuses on estimating the single stage optimal ITR. Extensions of D-learning to multiple-stage ITR problems can be obtained. Several methods have been developed to explore the dynamic ITR such as [22], [38] and [20]. We refer [17] for a review. It would be interesting to compare the performance of D-learning in multiple stages with existing methods. Another possible extension is to develop D-learning for other types of outcomes such as binary data and survival time.

Acknowledgements

The authors would like to thank the editor, the associate editor, and reviewers, whose helpful comments and suggestions led to a much improved presentation. Zhengling Qi and Yufeng Liu's research was partially supported by NSF grants IIS1632951, DMS-1821231 and NIH grant R01GM126550.

7. Appendix

7.1. More simulation results

For D-learning in the linear boundary case, we consider more situations where σ can be 1 or 4 and sample size is 400. Figure 5 shows misclassification rates for various settings. We can find the similar results as the case of $n = 100$ and $\sigma = 1$. When the σ becomes large, our proposed D-learning has the lowest error rate than the other three methods. The corresponding value functions as shown in Figure 6 is consistent with misclassification results. For multi-category treatment settings, we present the simulation results for the remaining settings such as $n = 800, \sigma = 1$, $n = 800, \sigma = 4$ and $n = 1600, \sigma = 4$. Our proposed multi-category D-learning has the lowest error rate and highest empirical value functions among all the methods as shown in Figures 7 and 8 respectively. In addition, according to our results in Figures 9 and 10, we conclude that the performance of soft OVO-OWL and OVO-RWL is better than hard OVO-OWL and OVO-RWL.

7.2. Proof

Proof of Lemma 1

Proof. If the differential operator and expectation could exchange, let $g(f) = \mathbf{E}\frac{1}{\pi(A,\mathbf{X})}(2RA - f)^2$, and we have

$$\begin{aligned}\frac{\partial g(f)}{\partial f} &= \mathbf{E}_{\mathbf{X}}\{\mathbf{E}[-\frac{2}{\pi(A,\mathbf{X})}(2RA - f(\mathbf{X}))|\mathbf{X}]]\} \\ &= \mathbf{E}_{\mathbf{X}}\{f\mathbf{E}[\frac{2}{\pi(A,\mathbf{X})}|\mathbf{X}] - 4\mathbf{E}[\frac{RA}{\pi(A,\mathbf{X})}|\mathbf{X}]\} \\ &= 4\mathbf{E}_{\mathbf{X}}\{f(\mathbf{X}) - \mathbf{E}[\frac{RA}{\pi(A,\mathbf{X})}|\mathbf{X}]\} \\ &= 4\mathbf{E}_{\mathbf{X}}\{f(\mathbf{X}) - f_0(\mathbf{X})\}.\end{aligned}$$

Then $\frac{\partial g(f_0)}{\partial f_0} = 0$ and by convexity, we have

$$f_0(\mathbf{X}) \in \text{argmin}\mathbf{E}\frac{1}{\pi(A,\mathbf{X})}(2RA - f)^2. \qquad \square$$

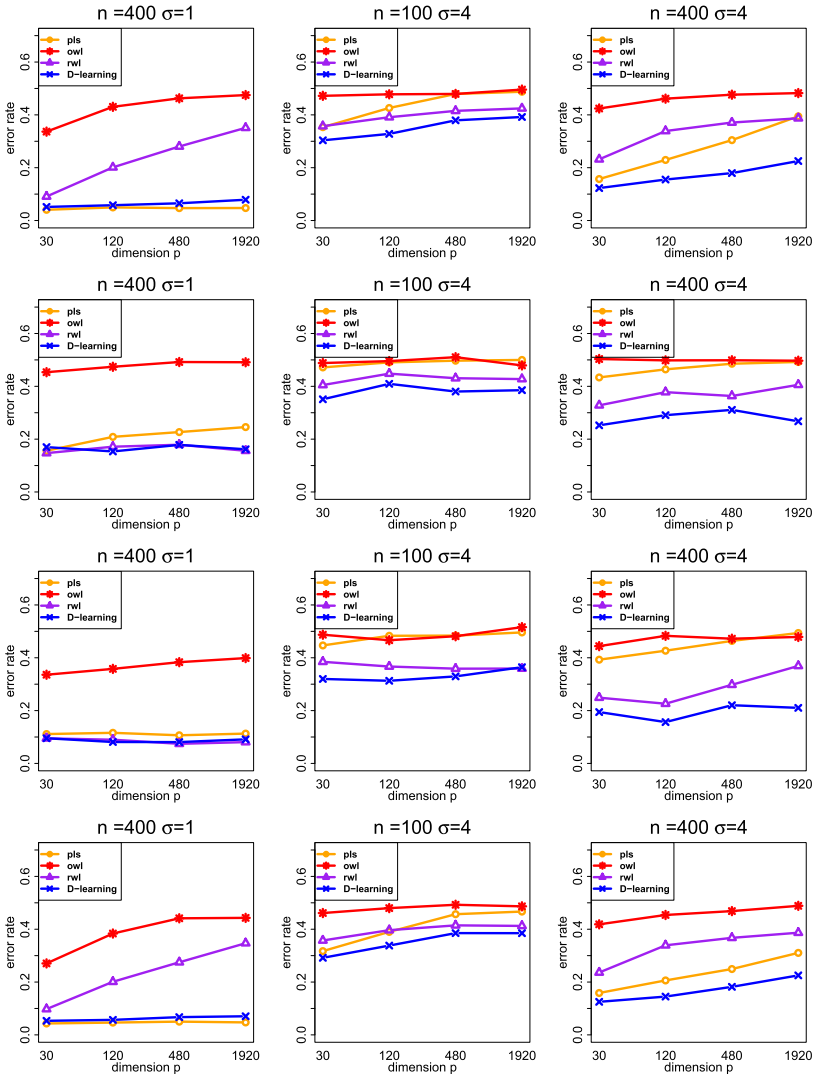

FIG 5. *Comparison of misclassification error rates on test data for simulated examples. From top to bottom, each row represents scenarios (1)–(4) respectively. The y-axis denotes the misclassification error rates for four methods and the x-axis is the dimension p from* 30 *to* 1920*, increased by a factor 4.*

Proof of Theorem 1

Proof. For any decision rule d, we have

$$V(d) = \mathbf{E}[\frac{R}{\pi(A, \mathbf{X})}\mathbb{I}(A = d(\mathbf{X}))]$$

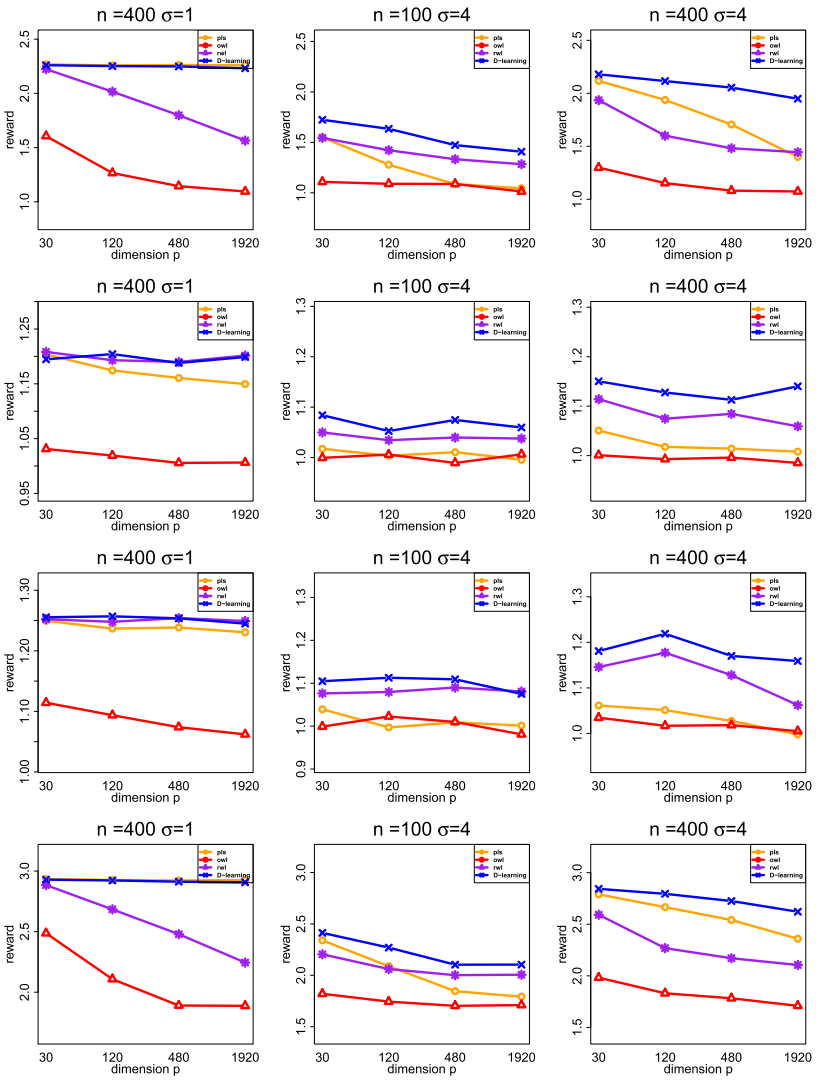

FIG 6. *Comparison of empirical value functions on test data for simulated examples. From top to bottom, each row represents scenarios (1)–(4) respectively. The y-axis denotes the empirical value functions for four methods and the x-axis is the dimension p from* 30 *to* 1920*, increased by a factor 4.*

$$= \mathbf{E}\{\sum_{k=1}^{K} \mathbf{E}[R|\mathbf{X}, A = k]\mathbb{I}(d(\mathbf{X}) = k)\}$$

$$= \mathbf{E}\{\mathbf{E}[R|A = i, \mathbf{X}] + \sum_{k\neq i}^{K} (\mathbf{E}[R|\mathbf{X}, A = i] - \mathbf{E}[R|\mathbf{X}, A = k])\mathbb{I}(d(\mathbf{X}) = k)\}$$

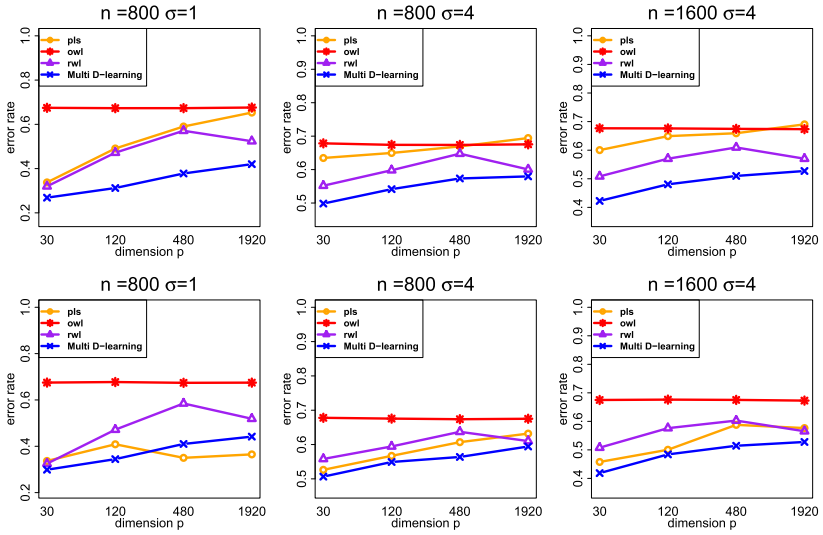

FIG 7. *Comparison of misclassification error rates on test data for simulated examples. From top to bottom, each row represents multi-category scenarios (1)–(2) respectively. The y-axis denotes the misclassification error rates for four methods and the x-axis is the dimension p from* 30 *to* 1920*, increased by a factor 4. Overall, multi-category D-learning performs the best compared to other three methods.*

$$= \mathbf{E}\{\mathbf{E}[R|A=i,\mathbf{X}] + \sum_{k\neq i}^{K} f_{ki}(\mathbf{X})\mathbb{I}(d(\mathbf{X})=k)\}.$$

By repeating this for K times with different i in the above, we can get

$$\begin{aligned} V(d) &= \frac{1}{K}\sum_{i=1}^{K}\{\mathbf{E}\{\mathbf{E}[R|A=i,\mathbf{X}] + \sum_{k\neq i}^{K} f_{ki}(\mathbf{X})\mathbb{I}(d(\mathbf{X})=k)\}\} \\ &= \mathbf{E}\{\frac{1}{K}\sum_{i=1}^{K}\mathbf{E}[R|A=i,\mathbf{X}]\} + \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{X})\mathbb{I}(d(\mathbf{X})=k)\}. \end{aligned}$$

Then, we have the value difference to be

$$\begin{aligned} & V(d_0) - V(\hat{d}_n) \\ &= E[\frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{X})\mathbb{I}(d_0(\mathbf{X})=k)] - E[\frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{X})\mathbb{I}(\hat{d}(\mathbf{X})=k))] \\ &= \frac{1}{K}E[\sum_{k=1}^{K} f_k(\mathbf{X})(\mathbb{I}(d_0(\mathbf{X})=k) - \mathbb{I}(\hat{d}(\mathbf{X})=k))] \\ &\leq \frac{1}{K}E[\sum_{i\neq j}^{K} |f_i(\mathbf{X}) - f_j(\mathbf{X})|(\mathbb{I}(d_0(\mathbf{X})=i, \hat{d}(\mathbf{X})=j)] \end{aligned}$$

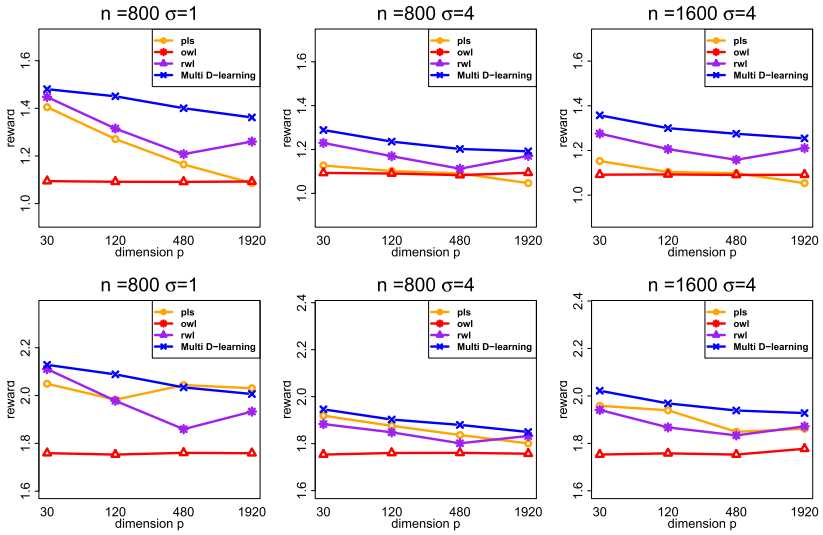

FIG 8. *Comparison of empirical value functions on test data for simulated examples. From top to bottom, each row represents scenarios (1)–(2) respectively. The y-axis denotes the empirical value functions for four methods and the x-axis is the dimension p from* 30 *to* 1920*, increased by a factor 4. Overall, multi-category D-learning performs the best compared to other three methods.*

$$
\begin{aligned}
&\le \frac{1}{K} E[\sum_{i\neq j}^{K} |f_i(\mathbf{X}) - f_j(\mathbf{X})| \mathbb{I}((f_i(\mathbf{X}) - f_j(\mathbf{X}))(\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X})) < 0)] \\
&\le \frac{1}{K} E[\sum_{i\neq j}^{K} |f_i(\mathbf{X}) - f_j(\mathbf{X}) - (\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X}))| \\
&\qquad\qquad \times \mathbb{I}((f_i(\mathbf{X}) - f_j(\mathbf{X}))(\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X})) < 0)] \\
&\le \frac{K-1}{K} \sum_{k=1}^{K} (\mathbf{E}[||f_k(\mathbf{X}) - \hat{f}_k(\mathbf{X})||_2^2])^{\frac{1}{2}},
\end{aligned}
$$

where the Hölder and Minkowski inequality are used in the last inequality. If Condition (3.1) is assumed as well, then we have

$$
\begin{aligned}
&V(d_0) - V(\hat{d}_n) \\
&\quad \le \frac{1}{K} E[\sum_{i\neq j}^{K} |f_i(\mathbf{X}) - f_j(\mathbf{X})| \mathbb{I}((f_i(\mathbf{X}) - f_j(\mathbf{X}))(\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X})) < 0)] \\
&\quad \le \frac{1}{K} \sum_{i\neq j}^{K} \{2\epsilon E[\mathbb{I}(|f_i(\mathbf{X}) - f_j(\mathbf{X})| \le 2\epsilon) \mathbb{I}((f_i(\mathbf{X}) - f_j(\mathbf{X}))(\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X})) < 0)] \\
&\qquad + \frac{1}{2\epsilon} E[(f_i(\mathbf{X}) - f_j(\mathbf{X}))^2 \mathbb{I}((f_i(\mathbf{X}) - f_j(\mathbf{X}))(\hat{f}_i(\mathbf{X}) - \hat{f}_j(\mathbf{X})) < 0)]\}
\end{aligned}
$$

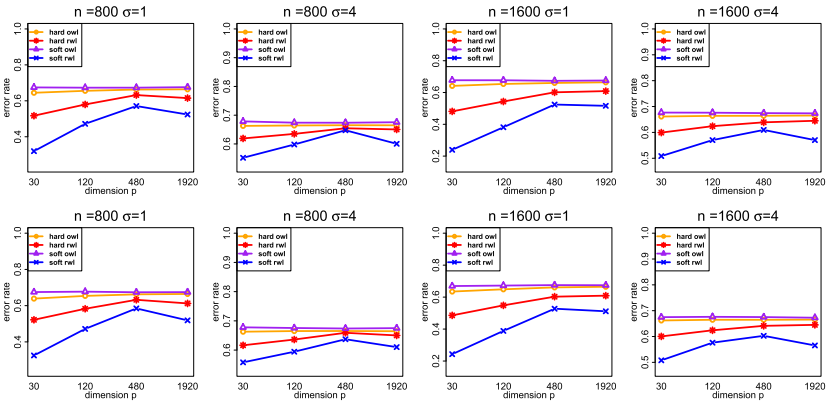

FIG 9. *Comparison of misclassification rates on test data for simulated examples. From top to bottom, each row represents scenarios (1)–(2) respectively. The y-axis denotes the misclassification rates for soft and hard versions of OVO-OWL and OVO-RWL and the x-axis is the dimension p from* 30 *to* 1920*, increased by a factor 4. Overall, the soft version performs better than the hard version.*

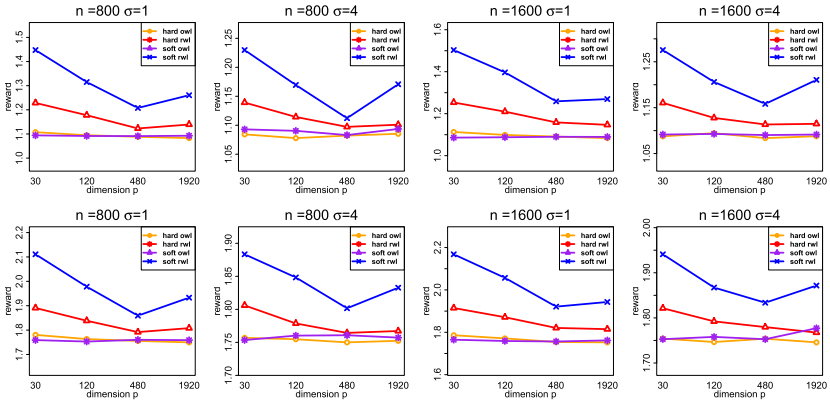

FIG 10. *Comparison of empirical value functions on test data for simulated examples. From top to bottom, each row represents scenarios (1)–(2) respectively. The y-axis denotes the empirical value functions for soft and hard versions of OVO-OWL and OVO-RWL and the x-axis is the dimension p from* 30 *to* 1920 *increased by a factor 4. Overall, the soft version performs better than the hard version.*

$$\leq \frac{1}{K}\sum_{i\neq j}^{K}\{\epsilon P(|(f_i(\mathbf{X})-f_j(\mathbf{X}))|\leq\epsilon)$$
$$+\frac{1}{\epsilon}E||f_i(\mathbf{X})-f_j(\mathbf{X})-(\hat{f}_i(\mathbf{X})-\hat{f}_j(\mathbf{X}))||_2^2\}$$
$$\leq \frac{1}{K}\sum_{i\neq j}^{K}\{2C\epsilon^{\alpha+1}+\frac{1}{2\epsilon}\mathbf{E}[||f_i(\mathbf{X})-\hat{f}_i(\mathbf{X})||]_2^2+\frac{1}{2\epsilon}\mathbf{E}[||f_j(\mathbf{X})-\hat{f}_j(\mathbf{X}))||]_2^2\}$$

$$\leq 2(K-1)C\epsilon^{\alpha+1} + \frac{1}{2\epsilon}\frac{K-1}{K}\sum_{k=1}^{K}(\mathbf{E}[||f_k(\mathbf{X}) - \hat{f}_k(\mathbf{X})||_2^2])^{\frac{1}{2}}.$$

Choosing $\epsilon = (\frac{\sum_{k=1}^{K}\mathbf{E}[||f_k(\mathbf{X})-\hat{f}_k(\mathbf{X})||_2]}{KC})^{\frac{1}{\alpha+2}}$ to minimize the above upper bound yields

$$V(d_0) - V(\hat{d_n}) \leq C'(K)(\sum_{k=1}^{K}\mathbf{E}[||f_k(\mathbf{X}) - \hat{f}_k(\mathbf{X})||_2^2])^{\frac{1+\alpha}{2+\alpha}}. \tag{7.1}$$

When $K = 2$, the result is similar. □

Proof of Theorem 2

Lemma 3 (Nemirovski moment inequality)**.** *For $m \geq 1$ and $p \geq e^{m-1}$, we have the following inequality*

$$\mathbf{E}\max_{1\leq j\leq p}|\sum_{i=1}^{n}(\gamma_j(Z_i) - \mathbf{E}\gamma_j(Z_i)|^m \leq [8\log(2p)]^{\frac{m}{2}}\mathbf{E}[\max_{1\leq j\leq p}\sum_{i=1}^{n}\gamma_j^2(Z_i)]^{\frac{m}{2}}. \tag{7.2}$$

Next we prove Theorem 2.

Proof. We start from the following basic inequality

$$||X(\hat{\beta} - f^0)||_2^2/n + \lambda||\hat{\beta}||_1 \leq 2\epsilon^T X(\hat{\beta} - \beta^*)/n + \lambda||\beta^*||_1 + \frac{||X\beta^* - f^0||_2^2}{n}. \tag{7.3}$$

For the first term in the left hand, by the Hölder's inequality we can have

$$2|\epsilon^T X(\hat{\beta} - \beta^0)/n| \leq (\max_{1\leq j\leq p} 2|\epsilon^T X^{(j)}|/n)||\hat{\beta} - \beta^0||_1.$$

Let $\lambda_0 = 16\sqrt{2}t^2\sqrt{\frac{log^2(2p)}{n}}$ and define a set

$$\Lambda := \{2\max_{1\leq j\leq p}|\epsilon^T X^{(j)}|/n \leq \lambda_0\}. \tag{7.4}$$

By applying Lemma 3 with $m = 2$, we can bound $P(\Lambda^c)$ as follow:

$$\begin{aligned}
P(\Lambda^c) &= P(\{2\max_{1\leq j\leq p}|\epsilon^T X^{(j)}|/n \geq \lambda_0\}) \\
&\leq \frac{\mathbf{E}\max_{1\leq j\leq p} 4|\epsilon^T X^{(j)}|^2/n^2}{\lambda_0^2} \\
&\leq \frac{4[8\log(2p)/n]\mathbf{E}[\max_{1\leq j\leq p}\sum_{i=1}^{n}\epsilon_i^2 X_{ij}^2/n]}{16\times 32t^2\frac{\log^2(2p)}{n}} \\
&\leq \frac{a^2\sum_{i=1}^{n}\mathbf{E}\epsilon_i^2/n}{16t^2\log(2p)} = \frac{a^2(4||X\gamma^0||_2^2/n + 4\sigma^2)}{t^2\log(2p)} \\
&\leq \frac{a^2(\Phi||\gamma^0||_2^2 + \sigma^2)}{4t^2\log(2p)}
\end{aligned}$$

$$\leq \frac{a^2(\Phi O(\log(2p)) + \sigma^2)}{4t^2 \log(2p)}$$
$$\leq \frac{C}{t^2}.$$

Thus we have

$$P(\Lambda) \geq 1 - \frac{C}{t^2} \quad \text{for any } t > 0.$$

With probability at least $1-\alpha$, where $\alpha = \frac{C}{t^2}$ for sufficiently large t, we have

$$\begin{aligned} & ||X(\hat{\beta} - f^0)||_2^2/n + \lambda||\hat{\beta}||_1 \\ \leq & \ (\max_{1\leq j\leq p} 2|\epsilon^T X^{(j)}|/n)||\hat{\beta} - \beta^*||_1 + \lambda||\beta^*||_1 + \frac{||X\beta^* - f^0||_2^2}{n} \\ \leq & \ \lambda_0||\hat{\beta} - \beta^*||_1 + \lambda||\beta^*||_1 + \frac{||X\beta^* - f^0||_2^2}{n}. \end{aligned}$$

Let $\lambda \geq 4\lambda_0 = 16\sqrt{2}t^2\sqrt{\frac{log^2(2p)}{n}}$, then we have the following lemma.

Lemma 4. *On the set Λ, with $\lambda \geq 2\lambda_0$,*

$$4||X(\hat{\beta} - f^0)||_2^2/n + 3\lambda||\hat{\beta}_{S_*^c}||_1 \leq 5\lambda||\hat{\beta}_{S_*} - \beta^*_{S_*}||_1 + 4||X\beta^* - f^0||_2^2/n, \quad (7.5)$$

where $S_* = \{j : \beta_j^* \neq 0\}$.

Proof. Focusing on the event Λ, by (7.3), and $\lambda = 4\lambda_0$, we have

$$4||X(\hat{\beta} - \beta^0)||_2^2/n + 4\lambda||\hat{\beta}||_1 \leq \lambda||\hat{\beta} - \beta^0||_1 + 4\lambda||\beta^0||_1.$$

By the triangle inequality

$$\begin{aligned} ||\hat{\beta}||_1 &= ||\hat{\beta}_{S_*}||_1 + ||\hat{\beta}_{S_*^c}||_1 \\ &= ||\beta_{S_*} - (\beta^*_{S_*} - \hat{\beta}_{S_*})||_1 + ||\hat{\beta}_{S_*^c}||_1 \\ &\geq ||\beta_{S_*}||_1 + ||\beta^*_{S_*} - \hat{\beta}_{S_*}||_1 + ||\hat{\beta}_{S_*^c}||_1, \end{aligned} \quad (7.6)$$

and $||\hat{\beta} - \beta^*||_1 = ||\beta^*_{S_*} - \hat{\beta}_{S_*}||_1 + ||\hat{\beta}_{S_*^c}||_1$. Then the desired inequality holds. □

By using this lemma, we have

$$\begin{aligned} & 4||X(\hat{\beta} - f^0)||_2^2/n + 3\lambda||\hat{\beta} - \beta^*||_1 \\ =& 4||X(\hat{\beta} - \beta^*)||_2^2/n + 3\lambda||\hat{\beta}_{S_*} - \beta^*_{S_*}||_1 + 3\lambda||\hat{\beta}_{S_*^c}||_1 \\ \leq& 12\lambda||\hat{\beta}_{S_*} - \beta^*_{S_*}||_1 \leq 12\sqrt{s_*}\lambda||X(\hat{\beta} - \beta^*)_{S_*}||_2/(\sqrt{n}\phi_*) \\ \leq& \sqrt{s_*}||X\hat{\beta} - f^0 + f^0 - X\beta_*||_2/(\sqrt{n}\phi_*) \\ \leq& 12\sqrt{s_*}||X\hat{\beta} - f^0||_2/(\sqrt{n}\phi_*) + 12\sqrt{s_*}||X\beta_* - f^0||_2/(\sqrt{n}\phi_*) \\ \leq& 6||X\beta^* - f^0||_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2} + 2||X\hat{\beta} - f^0||_2. \end{aligned} \quad (7.7)$$

Then $\frac{2||X\hat{\beta} - f^0||_2^2}{n} + 3\lambda||\hat{\beta} - \beta^*||_1 \leq 6||X\beta^* - f^0||_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2}$. In particular, if the underlying function f_0 is linear, i.e. there exists a β^* such that $f^0 = x^T\beta^*$, then

the first term of the right hand is 0, and $\frac{2||X\hat{\beta}-f^0||_2^2}{n} + 3\lambda||\hat{\beta}-\beta^*||_1 \leq \frac{24\lambda^2 s_*}{\phi_*^2} = O(\frac{\log(2p)}{n})$. □

Proof of Theorem 3

The results follow directly by combining Theorems 1 and 2.

Proof of Theorem 2

Proof. We first decompose the excess risk into estimation error and approximation error as follows:

$$\begin{aligned}
&L(\hat{f}) - L(f_0)\\
\leq &L(\hat{f}) - L_n(\hat{f}) + L_n(\hat{f}) + \frac{\lambda_n}{2}||\hat{f}||_{\mathcal{H}}^2\\
&-(L_n(f_{\lambda_n}) + \frac{\lambda_n}{2}||f_{\lambda_n}||_{\mathcal{H}}^2) + L_n(f_{\lambda_n}) - L(f_{\lambda_n}) + \mathcal{A}(\lambda_n)\\
\leq &(L(\hat{f}) - L_n(\hat{f})) + (L_n(f_{\lambda_n}) - L(f_{\lambda_n})) + \mathcal{A}(\lambda_n)\\
= &(I) + (II) + \mathcal{A}(\lambda_n),
\end{aligned}$$

where the first two terms (I) and (II) are estimation errors. The last inequality is because of the definition of $\hat{f}$.

In order to bound the estimation error, we first obtain the bounds for $||\hat{f}||_{\mathcal{H}}$ and $||f_{\lambda_n}||_{\mathcal{H}}$ correspondingly. By the definition of $\hat{f}$ in Model (2.11), we have

$$L_n(\hat{f}) + \frac{\lambda_n}{2}||\hat{f}||_{\mathcal{H}}^2 \leq L_n(0),$$

where $L_n(0,0) = \frac{1}{n}\sum_{i=1}^n R_i^2 \leq M_1$, since R is bounded by C_0 by assumption. Since $L_n \geq 0$, we can have $\frac{\lambda_n}{2}||\hat{f}||_{\mathcal{H}}^2 \leq M_1$ and $|L_n(\hat{f})| \leq M_1$, or equivalently $D(\hat{f}(\mathbf{X}_i)) = (R_i - \hat{f}(\mathbf{X}_i))^2 \leq M_2$ since $|\hat{f}(\mathbf{X}_i)| \leq ||\hat{f}||_{\mathcal{H}} \sup_{\mathbf{X}\in\mathcal{X}} K(\mathbf{X},\mathbf{X})$. By the similar argument, we can also obtain $\frac{\lambda_n}{2}||f_{\lambda_n}||_{\mathcal{H}}^2 \leq M_1$ and $|D(f_{\lambda_n})| \leq M_2$.

Define the following functional class

$$\Xi := \{D(f) \mid f \in \mathcal{H}, \frac{\lambda_n}{2}||f||_{\mathcal{H}}^2 \leq M_1, |D(f)| \leq M_2\}.$$

Let $\{Z_i\}_{i=1}^n = \{\mathbf{X}_i, A_i, R_i\}_{i=1}^n$ and P_n be the corresponding empirical measure on Z_n. We first derive the bound for the estimation errors (I) and (II). For the term (I), note that $(I) \leq \sup_\Xi PD(f) - P_nD(f)$, where P is probability measure of $(\mathbf{X}, A, R)$. When any $(\mathbf{X}_i, A_i, R_i)$ changes, by the definition of Ξ, $\sup_\Xi PD(f) - P_nD(f)$ is changed no more than $\frac{M_2}{n}$. Then by the McDiarmid's inequality, with probability at least $1 - \frac{\epsilon}{2}$, we can get

$$\sup_\Xi PD(f) - P_nD(f) \leq \mathbf{E}[\sup_\Xi PD(f) - P_nD(f)] + M_2\sqrt{\frac{2\log(\frac{1}{\epsilon})}{n}}. \tag{7.8}$$

Using the idea of symmetrization by introducing data $\{Z_i'\}_{i=1}^n$ and Rademacher variables $\{\sigma_i\}_{i=1}^n$, where σ_i is uniform over $\{1-1\}$, we can obtain

$$\begin{aligned}\mathbf{E}[\sup_{\Xi} PD(f) - P_nD(f)] &\le \mathbf{E}[\sup_{\Xi} \mathbf{E}[P_n'L_S(f,\alpha) - P_nL_S(f,\alpha)]] \\ &\le \mathbf{E}[\sup_{\Xi} P_n'D(f) - P_nD(f)] \\ &= \mathbf{E}[\sup_{\Xi} P_n\sigma(D(f) - D'(f)] \\ &\le \mathbf{E}[\sup_{\Xi} P_n\sigma D(f)] + \mathbf{E}[\sup_{\Xi} -P_n\sigma D(f)] \\ &= 2\mathbf{E}[\sup_{\Xi} P_n\sigma D(f)] \\ &= 2\mathcal{R}_n(\Xi).\end{aligned}$$

For the term (II), by the similar argument, we can show with probability at least $1-\frac{\epsilon}{2}$ that

$$\begin{aligned}(II) &\le \sup_{\Xi} P_nD(f) - PD(f) \\ &\le \mathbf{E}[\sup_{\Xi} P_nD(f) - PD(f)] + M_2\sqrt{\frac{2\log(\frac{1}{\epsilon})}{n}} \\ &\le 2\mathcal{R}_n(\Xi) + M_2\sqrt{\frac{2\log(\frac{1}{\epsilon})}{n}}.\end{aligned} \tag{7.9}$$

Then combining bounds of (I) and (II) together gives that with probability at least $1-\epsilon$,

$$(I) + (II) \le 4\mathcal{R}_n(\Xi) + M_2\sqrt{\frac{8\log(\frac{1}{\epsilon})}{n}}. \tag{7.10}$$

Define a class of functions as

$$\Pi := \{f \mid f \in \mathcal{H}, \frac{\lambda_n}{2}||f||_{\mathcal{H}}^2 \le M_1\}.$$

Note that $D(f)$ is Lipschitz with constant M_3 over Ξ. By Corollary 3.17 in [18], we have

$$\mathcal{R}_n(\Xi) \le M_3\mathcal{R}_n(\Pi). \tag{7.11}$$

Thus, combining together, with probability $1-\epsilon$,

$$L(\hat{f}) - L(f_0) \le 4M_3\mathcal{R}_n(\Pi) + M_2\sqrt{\frac{8\log(\frac{1}{\epsilon})}{n}} + \mathcal{A}(\lambda_n). \qquad \square$$

Proof of Theorem 4

Proof. Based on the assumptions and the definition of Π, by Lemma 22 in [2], we have

$$\mathcal{R}_n(\Pi) \le \sqrt{\frac{2M_1}{n\lambda_n}}.$$

According to Theorem 2 and assumptions on the approximation error, with probability $1-\epsilon$,

$$\begin{aligned} & L(\hat{f}) - L(f_0) \\ \leq & 4M_1\mathcal{R}_n(\Pi) + M_2\sqrt{\frac{8\log(\frac{1}{\epsilon})}{n}} + \mathcal{A}(\lambda_n) \\ \leq & \max(\sqrt{32M_1^2}, M_2\sqrt{8\log(\frac{1}{\epsilon})})\sqrt{\frac{1}{n\lambda_n}} + C_1\lambda_n^{\omega}. \end{aligned}$$

Then optimizing the right hand side with respect to λ_n, we can let $\lambda_n = \mathbf{O}((n^{-\frac{1}{2\omega+1}})$ and obtain the final result in the sense that with probability at least $1-\epsilon$,

$$L(\hat{f}) - L(f_0) \leq c_1 n^{-\frac{\omega}{2\omega+1}},$$

for some constant c_1 decreasing in ϵ. □

Proof of Corollary 4.1

Proof. The results follow directly by combining Theorems 1 and 4. □

References

[1] G. Baron, E. Perrodeau, I. Boutron, and P. Ravaud. Reporting of analyses from randomized controlled trials with multiple arms: a systematic review. *BMC medicine*, 11(1):84, 2013.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. MR1984026

[3] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. MR2807761

[4] J. Cohen. Statistical power analysis for the behavior science. *Lawrance Eribaum Association*, 1988.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[6] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[7] Y. Cui, R. Zhu, and M. Kosorok. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics*, 11(2):3927–3953, 2017. MR3714303

[8] A. Fan, W. Lu, and R. Song. Sequential advantage selection for optimal treatment regime. *The annals of applied statistics*, 10(1):32, 2016. MR3480486

[9] C. Fan, W. Lu, R. Song, and Y. Zhou. Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1565–1582, 2017. MR3731676

[10] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001. MR1946581

[11] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

[12] L. Gunter, J. Zhu, and S. Murphy. Variable selection for qualitative interactions. *Statistical methodology*, 8(1):42–55, 2011. MR2741508

[13] S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

[14] T. N. Kakuda. Pharmacology of nucleoside and nucleotide reverse transcriptase inhibitor-induced mitochondrial toxicity. *Clinical therapeutics*, 22(6):685–708, 2000.

[15] G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970. MR0254999

[16] E. Laber and Y. Zhao. Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514, 2015. MR3394271

[17] E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225, 2014. MR3263118

[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013. MR2814399

[19] Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006. MR2291500

[20] Y. Liu, Y. Wang, M. R. Kosorok, Y. Zhao, and D. Zeng. Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in medicine*, 2018.

[21] W. Lu, H. H. Zhang, and D. Zeng. Variable selection for optimal treatment decision. *Statistical methods in medical research*, page 0962280211428383, 2011. MR3190671

[22] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003. MR1983752

[23] S. A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6(Jul):1073–1097, 2005. MR2249849

[24] M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011. MR2816351

[25] J. M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004. MR2129402

[26] P. J. Schulte, A. A. Tsiatis, E. B. Laber, and M. Davidian. Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640, 2014. MR3300363

[27] R. Song, M. Kosorok, D. Zeng, Y. Zhao, E. Laber, and M. Yuan. On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4(1):59–68, 2015. MR3405390

[28] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008. MR2450103

[29] I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007. MR2336860

[30] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. MR3293607

[31] S. A. van de Geer, M. C. Veraar, J. A. Wellner, et al. Nemirovski's inequalities revisited. *American Mathematical Monthly*, 117(2):138–160, 2010. MR2590193

[32] G. Wahba. An introduction to smoothing spline anova models in rkhs, with examples in geographical data, medicine, atmospheric sciences and machine learning. *IFAC Proceedings Volumes*, 36(16):531–536, 2003.

[33] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3–4):279–292, 1992.

[34] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007. MR2411659

[35] B. Zhang, A. A. Tsiatis, E. B. Laber, and M. Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. MR3040007

[36] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010. MR2604701

[37] Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. MR3010898

[38] Y.-Q. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015. MR3367249

[39] X. Zhou, N. Mayer-Hamblett, U. Khan, and M. R. Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017. MR3646564

[40] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. MR2137327