Multi-armed Angle-based Direct Learning for Estimating Optimal Individualized Treatment Rules with Various Outcomes

Zhengling Qi[1], Dacheng Liu[2], Haoda Fu[3], and Yufeng Liu[1*]

[1]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, Chapel Hill

[2]Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA

[3]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA

[*]E-mail: yfliu@email.unc.edu

Abstract

Estimating an optimal individualized treatment rule (ITR) based on patients' information is an important problem in precision medicine. An optimal ITR is a decision function that optimizes patients' expected clinical outcomes. Many existing methods in the literature are designed for binary treatment settings with the interest of a continuous outcome. Much less work has been done on estimating optimal ITRs in multiple treatment settings with good interpretations. In this paper, we propose angle-based direct learning (AD-learning) to efficiently estimate optimal ITRs with multiple treatments. Our proposed method can be applied to various types of outcomes, such as continuous, survival or binary outcomes. Moreover, it has an interesting geometric interpretation on the effect of different treatments for each individual patient, which can help doctors and patients make better decisions. Finite sample error bounds have been established to provide a theoretical guarantee for AD-learning. Finally, we demonstrate the superior performance of our method via an extensive simulation study and real data applications.

Keywords: Modified Matrix; Multivariate responses regression; Multi-armed treatments; Personalized medicine

1 Introduction

Precision medicine, which recommends different treatments for individual patients, has been a popular research area in the scientific community. Compared with traditional "one-size-fits-all" medical procedures, precision medicine provides an individualized decision for each patient based on their information, such as clinical covariates, genetics, in order to maximize the outcome of each patient. There are different types of outcomes such as time to event, health index or the disease indicator.

There are a number of existing statistical methods for estimating optimal ITRs in the literature. These methods can be roughly characterized into two types. The first type includes value-based methods such as Q-learning (Watkins and Dayan (1992), Watkins (1989), Murphy (2005), Qian and Murphy (2011) and A-learning (Murphy (2003), Robins (2004)). Q-learning estimates optimal ITRs via modeling the conditional outcome function based on covariates while A-learning models the contrast between rewards of two treatments. The second type of methods directly targets the decision rules. One major approach of this type is to recast the estimating ITRs problem as weighted classification problems and use machine learning techniques to estimate optimal ITRs (Zhang et al. (2012), Zhao et al. (2012), Zhou et al. (2017), Zhao et al. (2015a), Tao and Wang (2016)). In order to enhance interpretability of decision rules, tree based methods were also proposed ((Zhang et al., 2015; Foster et al., 2011; Laber and Zhao, 2015)). Other direct-search methods include Tian et al. (2014) and Direct Learning (D-learning) (Qi and Liu (2017)), which directly estimate the decision function that leads to optimal ITRs by regression techniques. Recently, a general statistical framework to estimate optimal ITRs was proposed by Chen et al. (2017).

Censored data are commonly seen in practice. Thus, it is also important to develop methods to estimate optimal ITRs for the survival outcome. Various methods have been proposed in the literature to estimate optimal ITRs for survival outcomes, such as Goldberg and Kosorok (2012), Zhao et al. (2015b) and Cui et al. (2017). Recently, Bai et al. (2016) and Jiang et al. (2016) proposed several methods to estimate the optimal ITR that can maximize the survival probability of patients. However, for general ITR problems, most of these existing methods are designed for binary treatment settings only. There are many multi-armed ITR problems in pratice (Baron et al. (2013)). To the best of our knowledge, not much has been done for estimating the optimal

ITR for the multi-armed treatment settings with various outcomes, such binary and survival outcomes. Thus it is essential to develop methods to take multiple treatments into consideration simultaneously and estimate optimal ITRs for various outcomes, which can help to improve the estimating efficiency and the classification accuracy.

Besides the accurate estimation of ITRs, good interpretations are also important for multi-armed treatment settings. For binary treatment settings, value-based methods can report a single value difference function between two treatments to illustrate the relative effectiveness. For classification based methods such as O-learning (Zhao et al. (2012)), interpretation of the decision rule for binary treatment settings may not be as clear. Meanwhile for K-armed treatment settings, at least $\frac{K(K-1)}{2}$ pairwise value difference functions need to be estimated to illustrate the relative performance of treatments for each patient. Although such an extension can be simple to implement, it does not use the data simultaneously and consequently may yield suboptimal rules.

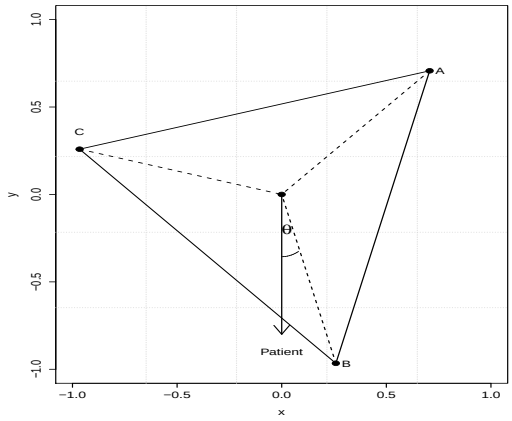


Figure 1: Graphical illustration of the estimated ITR for a given patient in a three-treatment setting. Vertices A, B and C represent 3 treatments. The estimated ITR of the patient has the least angle with treatment B which is thus more preferable than the other two treatments.

To get accurate estimation of optimal ITRs and obtain a good interpretation jointly under the multi-armed setting, we consider a K-vertex simplex structure in an Euclidean space, where each vertex represents one treatment. The simplex lies in a $K-1$ dimensional space with the origin as the center and has equal inner products among vertices. Using the expression of the optimal ITR, we transform the problem of finding the optimal ITR maximizing the value function into maximizing the inner product between the decision function vector and the corresponding vertex

in the simplex space. Such a transformation allows us to estimate the optimal ITR using multiple response regression methods. In particular, for each patient, our estimated decision function vector maps the covariates into this $K-1$ dimensional space. The angle between each treatment vertex and the estimation function vector can be interpreted as a measure of preference to this treatment. We recommend a patient to take the corresponding treatment having the least angle with our estimated decision function vector. Figure 1 shows an example with our estimated ITR for a given patient. In this case, we recommend treatment B as the best option for this patient. In addition, we can see treatment C is more preferable than treatment A for this patient based on their angles.

We call our method angle-based direct learning (AD-learning) which can directly estimate optimal ITRs under multi-armed treatment settings using multiple response regression techniques. Furthermore, our proposed AD-learning can be extended to various types of outcome such as binary and survival responses. Compared with existing methods, our proposed AD-learning enjoys several advantages. In particular, our method is robust in the sense that it is not necessary to model the main effect function of the conditional outcome. Due to direct learning scheme, our method does not suffer from the mismatch problem between minimizing prediction errors and maximizing value functions in model based methods such as l_1-PLS (Qian and Murphy (2011)) and can perform better in high dimensional settings. Moreover, by representing each treatment as a vertex of a standard simplex in the Euclidean space, our proposed method provides an attractive geometric interpretation of the relative effectiveness of all treatments for a given patient. The resulting relative effectiveness of different treatments can be interpreted as the angles between the decision function vector for the patient and each vertex corresponding to the treatment. These angles can be scaled between $[0, \pi]$. In addition, flexible structures such as group and low rank sparsity can be also incorporated to further improve the model interpretation and simplicity, which can be applied in high dimensional settings. Finally, our proposed method is easy to implement with efficient algorithms.

The remainder of this paper is organized as follows. In Section 2, we introduce our AD-learning to estimate optimal ITRs in multiple treatment settings. In Section 3, we discuss how to extend our proposed method to binary and survival outcomes. In Section 4, we provide a theoretical guarantee for our AD-learning under some mild assumptions. In Section 5, we conduct an extensive simulation study to evaluate the finite sample performance of our method

with implementation details including algorithms. Furthermore, we illustrate our method using the AIDS data in Section 6. We conclude our paper with some discussions and possible future extensions in Section 7.

2 Angle Based Direct Learning

For notation of the paper, we use boldface capital and lowercase symbols to denote matrices and vectors respectively. For a matrix $\mathbf{B}$, we define a mixed l_1 and l_2 norm as $||\mathbf{B}||_{2,1} = \sum ||\mathbf{B}_j||_2$, where $\mathbf{B}_j$ is the j-th row vector of $\mathbf{B}$. We use $\mathrm{Tr}(\mathbf{B})$ to denote the sum of the diagonal value of the matrix $\mathbf{B}$.

We consider a randomized treatment framework for estimating optimal ITRs. For each patient, we observe a triplet random vector $(\mathbf{x}, A, R)$. In particular, $\mathbf{x} = (1, X_1, \cdots, X_p) \in \mathcal{X}$ denotes patients' p-dimensional covariates with an intercept. The random variable A represents the randomized treatment that a patient receives. Here we consider the K-treatment-armed setting where $A \in \{1, 2, \cdots, K\}$ with a known prior probability distribution $\pi(A, \mathbf{x})$, which is the conditional probability depending on $\mathbf{x}$. In a general setting other than the randomized trial study, $\pi(A, \mathbf{x})$ denotes the propensity score and can be estimated by the generalized linear models such as multinomial logistic regression. The variable R is a patient's outcome after receiving the treatment A. Without loss of generality, we assume that the outcome R is bounded and the larger R is, the better the treatment works for this patient.

One of the most important goals of our problem is to estimate the optimal ITR that can maximize the expected clinical outcome of each patient under this ITR. Mathematically speaking, an ITR is a decision function $d(\mathbf{x}) : \mathcal{X} \to \mathcal{A}$, mapping from the covariate space into the treatment space. According to Qian and Murphy (2011) and Zhao et al. (2012), the value function under the ITR d can be expressed as

$$V(d) =: \mathbf{E}[R|d(\mathbf{x}) = A] = \mathbf{E}[\frac{R\mathbb{I}(A = d(\mathbf{x})}{\pi(A, \mathbf{x})}], \tag{1}$$

where $\mathbb{I}(\bullet)$ is the indicator function. Then the optimal ITR is defined as

$$d_0(\mathbf{x}) = \mathrm{argmax}_{d \in D} V(d) \tag{2}$$

within a pre-specified class of treatment rules D. Before introducing our proposed AD-learning,

we first discuss the direct learning framework.

2.1 The Direct Learning Framework

Consider a binary problem with $K = 2$. We encode treatment A to be 1 or -1. Then from the value function and optimal ITRs defined in (1) and (2) respectively, we can further represent the optimal ITR as

$$\begin{aligned} d_0(\mathbf{x}) &= \text{sign}(\mathbf{E}[R|\mathbf{x}, A = 1] - E[R|\mathbf{x}, A = -1]) \\ &= \text{sign}(\mathbf{E}[\frac{RA}{\pi(A|\mathbf{x})}|\mathbf{x}]) := \text{sign}(f_0(\mathbf{x})). \end{aligned} \tag{3}$$

Using Equation (3), similarly in Tian et al. (2014), the ITR problem becomes to estimate the optimal decision function $f_0(\mathbf{x}) = \mathbf{E}[\frac{RA}{\pi(A|\mathbf{x})}|\mathbf{x}]$ via various regression methods such as l_1 penalized regression (LASSO). The final decision rule is determined by the sign of the estimator.

Binary D-learning directly estimates the decision rule. It is very different from the outcome weighted learning (OWL) proposed by Zhao et al. (2012) because binary D-learning uses regression methods to estimate the optimal ITR directly. Note that binary D-learning can be simply extended to the K-treatment-arm setting by rewriting the optimal ITR as

$$\begin{aligned} d_0(\mathbf{x}) =& \underset{k\in\{1,\cdots,K\}}{\text{argmax}} \mathbf{E}[R|\mathbf{x}, A = k] \\ =& \underset{k\in\{1,\cdots,K\}}{\text{argmax}} K\mathbf{E}[R|\mathbf{x}, A = k] - \sum_{i=1}^{K} \mathbf{E}[R|\mathbf{x}, A = i] \\ =& \underset{k\in\{1,\cdots,K\}}{\text{argmax}} \sum_{i\neq k}^{K} \{\mathbf{E}[R|\mathbf{x}, A = k] - \mathbf{E}[R|\mathbf{x}A = i]\} \\ =& \underset{k\in\{1,\cdots,K\}}{\text{argmax}} \sum_{i\neq k}^{K} \mathbf{E}[\frac{RA_{ki}}{\pi_{ki}(A_{ki}, \mathbf{x})}|\mathbf{x}, A = k \text{ or } i] \\ :=& \underset{k\in\{1,\cdots,K\}}{\text{argmax}} \sum_{i\neq k}^{K} f_{ki}(\mathbf{x}) := \underset{k\in\{1,\cdots,K\}}{\text{argmax}} f_k(\mathbf{x}), \end{aligned} \tag{4}$$

where $A_{ki} \in \{-1, 1\}$ represents treatments k and i, and $f_{ki}(\mathbf{x})$ is defined as the optimal decision function between treatment k and i. Each pairwise decision function can be estimated by a binary D-learning method. The final treatment decision rule is to compare the cumulative sum of pairwise decison functions $f_k(\mathbf{x})$ for $k = 1, \cdots, K$, and select the largest one. We refer this pairwise method as pairwise D-learning.

Binary D-learning gives us a directed way to estimate optimal ITRs. However, pairwise D-

learning, which is based on binary D-learning, focuses only on pairwise comparisons between treatments without considering all treatments simultaneously. Although the proposed effect measure $f_k(\mathbf{x})$ can capture the relative strength of a treatment for a given patient, it may be suboptimal.

To handle multi-armed ITR problems with various outcomes, we propose AD-learning that considers all treatments together to estimate the optimal ITR. Moreover, the AD-learning can provide a more effective measure of treatments for patients with a good interpretation.

2.2 Angle Based D-learning for Continuous Outcomes

For a K-armed ITR problem, one natural approach is to estimate K functions for all treatments. Since only one function is needed for the binary ITR problem, one indeed only needs $K-1$ functions for a K-armed problem. Instead of using K functions with a constraint on these functions, we aim to directly estimate $K-1$ functions. To that end, we project the treatment A into K simplex vertices defined on $\mathcal{R}^{K-1}$. Specifically, we encode the j-th treatment as a vector $\mathbf{w}_j \in \mathcal{R}^{K-1}$ with

$$\mathbf{w}_j = \begin{cases} (K-1)^{-1/2}\mathbf{1}_{K-1}, & \text{if } A = 1 \\ -(1+\sqrt{K})/(K-1)^{3/2}\mathbf{1}_{K-1} + (\frac{K}{K-1})^{1/2} e_{A-1}, & \text{if } 2 \le A \le K, \end{cases} \tag{5}$$

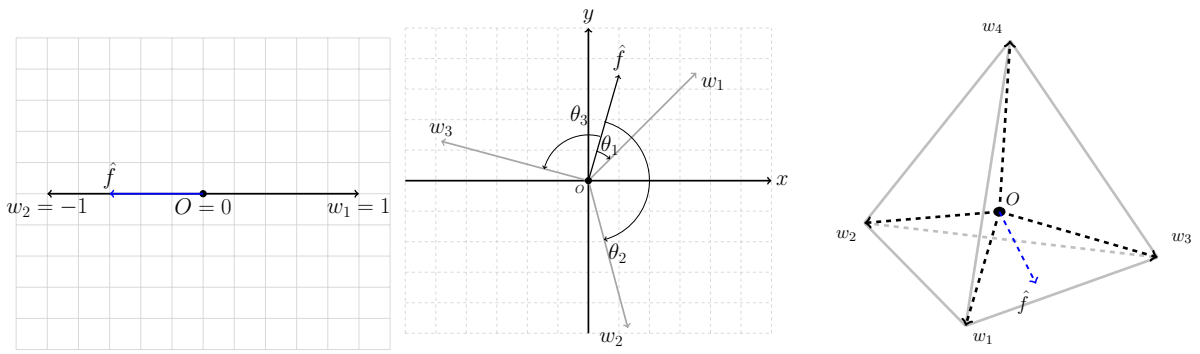
where e_i is a $K-1$ dimensional vector with every element being 0, except the i-th location being 1. Define $\mathbf{w}$ as a random vector with $\mathbf{P}[\mathbf{w} = \mathbf{w}_j|\mathbf{x}] = \mathbf{P}[A = j|\mathbf{x}]$. This simplex encoding scheme has several properties. In particular, the center of these vertices is the origin of the space, that is $\sum_{j=1}^{K} \mathbf{w}_j = 0$ with $||\mathbf{w}_j||_2 = 1$ for $j = 1, \cdots, K$. The angle between each pair of vertices is equal, that is $\mathbf{w}_i^T \mathbf{w}_j = C(K) < 1$ for $i \neq j$, where the constant C only depends on K. Interestingly, we

can then express the optimal ITR as

$$
\begin{aligned}
d_0(\mathbf{x}) &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \mathbf{E}[R|\mathbf{x}, A=k] \\
&= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ (1-c(K))\mathbf{E}[R|\mathbf{x}, A=k] \\
&= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \{(1-c(K))\mathbf{E}[R|\mathbf{x}, A=k] + c(K)\sum_{j=1}^{K}\mathbf{E}[R|\mathbf{x}, A=j]\} \\
&= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \{\mathbf{E}[R|\mathbf{x}, A=k] + c(K)\sum_{j\neq k}^{K}\mathbf{E}[R|\mathbf{x}, A=j]\} \\
&= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \{\mathbf{w}_k^T\mathbf{E}[R\mathbf{w}|\mathbf{x}, A=k] + \mathbf{w}_k^T\sum_{j\neq k}^{K}\mathbf{E}[R\mathbf{w}|\mathbf{x}, A=j]\} \\
&= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \mathbf{w}_k^T\mathbf{E}[\frac{R\mathbf{w}}{\pi(A,\mathbf{x})}|\mathbf{x}] := \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\ \mathbf{w}_k^T\mathbf{f}_0(\mathbf{x}),
\end{aligned}
\tag{6}
$$

where $\mathbf{f}_0(\mathbf{x})$ is a function vector from $\mathcal{R}^{p+1}$ to $\mathcal{R}^{K-1}$ with some abuse of notation. Then the optimal ITR is given by comparing the inner product between $\mathbf{w}_k$ and $\mathbf{f}_0(\mathbf{x})$ for each treatment k. We define the angle between each pair of vertices in $[0, \pi]$. Then $\mathbf{w}_k^T\mathbf{f}_0(\mathbf{x})$ is the largest if and only if the angle between $\mathbf{w}_k$ and $\mathbf{f}_0(\mathbf{x})$ is the least, for $k = 1, \cdots, K$. Thus we call our proposed method as Angle based D-learning (AD-learning). Note that the simplex coding is unique up to the orthogonal rotation.

Our proposed AD-learning has an attractive geometric interpretation. In particular, this least angle decision rule can be understood through newly defined treatment regions for each patient. For example, when $K = 3$, as shown in Figure 2 (b), vectors $\mathbf{w}_k$; $k = 1, \cdots, K$ form an equilateral triangle in the $\mathcal{R}^2$ space, where each divided region represents a treatment region. The decision function vector $\mathbf{f}_0(\mathbf{x})$ maps from the covariate space into the treatment region. One can observe that the angles between vertices are the same, and consequently each treatment is treated equally. Such a simplex coding scheme does not require a balanced group size for each treatment since treatment proportions are taken into account by the term $\pi(A, \mathbf{x})$ in Equation (6). We name the angle between each $\mathbf{w}_k$ and $\mathbf{f}_0(\mathbf{x})$ as the *treatment score* which lies in a bounded interval $[0, \pi]$. If a patient has the angle of 0 with the i-th treatment, then the i-th treatment is a perfect fit for this patient compared with other treatments. Figure 2 gives a geometric explanation of our AD-learning.



(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

Figure 2: Geometric interpretation of our least angle decision rule. When $K = 3$ or $K = 4$, the estimate $\hat{f}$ has the smallest angle with treatment 1 so we recommend treatment 1 as the optimal treatment. When $K = 2$, we can see $\hat{f}$ has the smallest angle with vector w_2 and the optimal rule for this patient is treatment 2.

To further illustrate our AD-learning, we propose the following alternative interpretation. Suppose the clincal outcome R can be modeled as

$$R = \mu(\mathbf{x}) + \sum_{i=1}^{K} \delta_i(\mathbf{x}) \mathbb{I}(A = i) + \epsilon, \tag{7}$$

where $\mu(\mathbf{x})$ is main effect function, $\delta_i(\mathbf{x})$ is the interaction effect between covariates and the i-th treatment, and ϵ is mean zero random error. Then we can get

$$\begin{aligned} E[\frac{R\mathbf{w}}{\pi(A,\mathbf{x})}|\mathbf{x}] &= \mu(\mathbf{x}) E[\frac{\mathbf{w}}{\pi(A,\mathbf{x})}|\mathbf{x}] + \sum_{i=1}^{K} \delta_i(\mathbf{x})_i \mathbf{E}[\frac{\mathbf{w}\mathbb{I}(A=i)}{\pi(A,\mathbf{x})}|\mathbf{x}] + E[\frac{\mathbf{w}}{\pi(A,\mathbf{x})}|\mathbf{x}]\mathbf{E}[\epsilon|\mathbf{x}] \\ &= \sum_{i=1}^{K} \delta_i(\mathbf{x}) \mathbf{w}_i. \end{aligned} \tag{8}$$

Furthermore, the optimal ITR is

$$\begin{aligned} d_0(\mathbf{x}) &= \text{argmax}_{k \in \{1,\cdots,K\}} \mathbf{w}_k^T \mathbf{E}[\frac{R\mathbf{w}}{\pi(A|\mathbf{x})}|\mathbf{x}] \\ &= \text{argmax}_{k \in \{1,\cdots,K\}} \mathbf{w}_k^T \sum_{i=1}^{K} \delta_i(\mathbf{x}) \mathbf{w}_i \\ &= \text{argmax}_{k \in \{1,\cdots,K\}} C(K) \sum_{i=1}^{K} \delta_i(\mathbf{x}) + (1 - C(K)) \delta_k(\mathbf{x}) \\ &= \text{argmax}_{k \in \{1,\cdots,K\}} \delta_k(\mathbf{x}), \end{aligned} \tag{9}$$

which is exactly to compare each treatment interaction effect with the covariates.

As a remark, we note that extensions of methods for binary treatment settings to multiple treatment settings using all treatments jointly can be nontrivial since we need to account for multiple treatment effect comparisons without sacrificing too much efficiency. Our proposed AD-learning achieves this by first projecting treatments into a $K-1$ dimensional space. A simplex with K vertices is used to represent the K treatments. Then Equation (6) provides an innovative but direct way to efficiently estimate the decision function vector and considers all the data simultaneously. Inherited from the simplex structure, our proposed method has an attractive geometric interpretation to show the relative effectiveness of different treatments for a patient. Thus it provides an informative comparison of all treatments for patients and doctors to make decisions.

Note that the simplex coding scheme was previously used by Wu and Lange (2010) and Zhang and Liu (2014) for classification problems. However, our proposed AD-learning is very different because it is not a classification method. Consequently, our method is not an extension of O-learning proposed by Zhao et al. (2012). Instead, by transforming the problem (2) into (6), our goal is to estimate the decision function $\mathbf{f}_0(\mathbf{x})$ directly, using multiple response regression introduced in Section 2.3.

2.3 Estimation Procedures of AD-learning

In order to estimate the optimal ITR, it is equivalent to estimating $\mathbf{f}_0(\mathbf{x})$ from Section 2.2. The next lemma provides us a way for estimation of $\mathbf{f}_0(\mathbf{x})$.

Lemma 1. *Under the exchange of differential and expectation condition,* $\mathbf{f}_0(\mathbf{x})$ *is an optimal solution to*

$$\underset{\mathbf{f}\in\mathcal{R}^{K-1}}{argmin} \ \mathbf{E}[\frac{1}{\pi(A,\mathbf{x})}(KR\mathbf{w}-\mathbf{f}(\mathbf{x}))^T\Sigma(KR\mathbf{w}-\mathbf{f}(\mathbf{x}))], \tag{10}$$

where Σ *can be any invertible matrix that characterizes the dependency among responses. Without knowing any prior knowledge, one could simply let* $\Sigma = I_{K-1}$.

Assume we observe independent identically distributed data $\{(\mathbf{x}_i, A_i, R_i), i = 1, \cdots, n\}$. Then we can estimate $\mathbf{f}_0(\mathbf{x})$ via empirical average approximation

$$\underset{\mathbf{f}\in\mathcal{F}}{\text{argmin}} \quad \frac{1}{n(K-1)}\sum_{i=1}^{n}\frac{1}{\pi(A_i,\mathbf{x}_i)}(KR_i\mathbf{w}_i-\mathbf{f}(\mathbf{x}_i))^T(KR_i\mathbf{w}_i-\mathbf{f}(\mathbf{x}_i)), \tag{11}$$

where $\mathcal{F}$ is a pre-specified class of decision functions. For simplicity, we first consider the class of linear decision rules, that is, $\mathcal{F} := \{\mathbf{f}(\mathbf{x}) = \mathbf{B}^T\mathbf{x}, \mathbf{B} \in \mathbb{R}^{p\times(K-1)}\}$. By observing $KR_i\mathbf{w}_i$ as multivariate responses, one can apply ordinary least square estimates for each of the responses separately. However, since the responses share the same clinical outcome R_i for the i-th sample, it is clear that pooling multivariate responses together can efficiently improve the estimation of $\mathbf{f}_0(\mathbf{x})$ (Breiman and Friedman (1997)). This motivates us to incorporate shrinkage and selection strategies that explore the correlations among different responses by

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \quad \frac{1}{n(K-1)}\sum_{i=1}^{n}\frac{1}{\pi(A_i,\mathbf{x}_i)}(KR_i\mathbf{w}_i - \mathbf{B}^T\mathbf{x}_i)^T(KR_i\mathbf{w}_i - \mathbf{B}^T\mathbf{x}_i) + \lambda J(\mathbf{B}), \tag{12}$$

where λ is a positive tuning parameter. Then our final least angle decision rule becomes $d_0(\mathbf{x}) = \operatorname{argmax}_{k\in\{1,\cdots,K\}}\mathbf{w}_k^T\mathbf{B}^T\mathbf{x}$. In this decision rule, the corresponding coefficient for the j-th variable of $\mathbf{x}$ is $\mathbf{w}_k^T\mathbf{B}_j$, for $j = 1, \cdots, p$, where $\mathbf{B}_j$ is the j-th row vector of $\mathbf{B}$. Note that for any orthogonal matrix $\mathbf{\Gamma}$,

$$\begin{aligned} ||\mathbf{B\Gamma}||_{2,1} = \sum_{j=1}^{p}||\mathbf{B}_j^T\mathbf{\Gamma}||_2 \quad &= \sum_{j=1}^{p}\sqrt{\mathbf{B}_j^T\mathbf{\Gamma\Gamma}^T\mathbf{B}_j} \\ = \sum_{j=1}^{p}||\mathbf{B}_j||_2 \quad &= ||\mathbf{B}||_{2,1}, \end{aligned} \tag{13}$$

which implies that $||\mathbf{B}||_{2,1}$ remains to be the same under any orthogonal transformation of $\mathbf{w}$. This is essential since our simplex coding is unique up to the orthogonal rotation. In addition, $\mathbf{B}_j = \mathbf{0}_{K-1}$ implies the j-th variable has no effect on our least angle decision rule. These motivate us to use the group sparsity penalty, i.e., the mixed l_1/l_2 norm as follows

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \quad \frac{1}{n(K-1)}\sum_{i=1}^{n}\frac{1}{\pi(A_i,\mathbf{x}_i)}(KR_i\mathbf{w}_i - \mathbf{B}^T\mathbf{x}_i)^T(KR_i\mathbf{w}_i - \mathbf{B}^T\mathbf{x}_i) + \lambda||\mathbf{B}||_{2,1}. \tag{14}$$

Model (14) is best suited for the case that all treatments share the common interaction covariates. The group sparsity structure of $\mathbf{B}$ will not change under any orthogonal transformation of $\mathbf{w}$.

In the literature, it is known that group sparsity of a matrix is a special case of a low rank matrix. If $\mathbf{B} = \mathbf{UV}^T$ such that $\mathbf{U} \in \mathbb{R}^{p\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times(K-1)}$ with $r < \min(p, K-1)$. Then $\mathbf{B}^T\mathbf{x} = \mathbf{V}(\mathbf{U}^T\mathbf{x})$ implies potential r orthogonal latent factors in the covariates. Hence we can also use the nuclear norm penalty to control the complexity of coefficient matrix $\mathbf{B}$ if there is a

low rank structure or exists latent factors in the covariates by

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \quad \frac{1}{n(K-1)}\sum_{i=1}^{n}\frac{1}{\pi(A_i,\mathbf{x}_i)}(KR_i\mathbf{w}_i-\mathbf{B}^T\mathbf{x}_i)^T(KR_i\mathbf{w}_i-\mathbf{B}^T\mathbf{x}_i)+\lambda||\mathbf{B}||_*, \tag{15}$$

where the $||\mathbf{B}||_*$ is the sum of all singular values of coefficient matrix $\mathbf{B}$. The nuclear norm penalty, unlike the rank constraint, provides soft and stable shrinkage on the singular values. Similar to the penalty $||\mathbf{B}||_{2,1}$, other penalties including $||\mathbf{B}||_*$ that are invariant to any orthogonal rotation of $\mathbf{w}$ can be applied for our methods.

So far, we have only focused on linear decision rules. If $\mathbf{f}_0(\mathbf{x})$ belongs to some classes of nonlinear functions, we can adapt our method to nonlinear learning via kernel learning or basis function expansions. For kernel learning, we can apply kernel ridge regression for each response separately, using Equation (11). However, it may lose some efficiency since it does not consider the dependence among the responses. How to perform kernel learning with multiple responses in our setting is an interesting future research direction. For basis function expansions, depending on the problem, we can use spline basis functions, interaction functions, wavelet functions, etc. to approximate the nonlinear decision function.

To summarize, Models (14) and (15) are proposed to control the complexity of coefficient matrix $\mathbf{B}$ and consequently enhance the estimation and prediction. As our proposed AD-learning directly targets on the decision function $f_0(\mathbf{x})$, it does not suffer the mismatch problem between minimizing prediction errors and maximizing value functions happened for model-based methods such as l_1-PLS. Thus our proposed method tends to perform better in high dimensional settings. If there are group signals in the covariates for optimal ITRs, we recommend to use Model (14). If there are latent factors in the covariates for optimal ITRs, we recommend to use Model (15). One can also use the cross-validation procedure to choose Model (14) or (15) that maximizes the empirical value function on the validation dataset. The computation of these models involves convex optimization and thus can be solved efficiently.

3 Extensions to Other Types of Outcomes

In Sections 2, we proposed AD-learning for continuous outcomes. In practice, especially in clinical studies, other types of outcomes such as binary, count responses, or survival time can also be used. In this section, we extend our AD-learning to more general types of outcomes motivated by the following lemma.

Lemma 2. *Under the exchange of differential and expectation condition,* $\mathbf{f}_0(\mathbf{x})$ *is an optimal solution to*

$$\underset{\mathbf{f}\in\mathcal{F}}{argmin} \quad \mathbf{E}[\frac{1}{\pi(A,\mathbf{x})}(\frac{K}{K-1}R - \mathbf{w}^T\mathbf{f}(\mathbf{x}))^2]. \tag{16}$$

Based on the optimization problem (16), one can write a corresponding working model as

$$\frac{K}{K-1}R = \mathbf{w}^T f(\mathbf{x}) + \epsilon, \tag{17}$$

where ϵ is the random error. Note that when $\mathbf{f} \in \mathcal{F}$, $\mathbf{w}^T\mathbf{f}(\mathbf{x}) = \mathbf{w}^T\mathbf{B}^T\mathbf{x} = \mathrm{Tr}(\mathbf{B}^T(\mathbf{x}\mathbf{w}^T))$. Then $\mathbf{x}\mathbf{w}^T$ can be regarded as modified covariates. Then the multiple response regression model in (11) can be extended to a more general model, namely trace regression model (Rohde et al. (2011)).

Motivated by the optimization problem (16) and the corresponding working model, we can extend our proposed AD-learning to more general settings. In particular, instead of the least squared loss for continuous outcome in (16), we can use other loss functions for corresponding outcomes.

3.1 Binary Outcomes

When R is binary, motivated by Lemma 2 and the connection between (16) and working model (17), we consider to replace the least squared loss in (16) by the deviance loss of logistic regression models. Then we have the following lemma.

Lemma 3. *Under the exchange of differential and expectation condition, an optimal solution to*

$$\underset{\mathbf{f}\in\mathcal{F}}{argmin} \quad \mathbf{E}[-\frac{R\mathbf{w}^T\mathbf{f}}{\pi(A,\mathbf{x})} + \frac{\log(1+\exp(\mathbf{w}^T\mathbf{f}))}{\pi(A,\mathbf{x})}] \tag{18}$$

is the function $\mathbf{f}_0(\mathbf{x})$ *satisfying*

$$\mathbf{P}[R=1|\mathbf{x}, A=i] = \frac{\exp(\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}))}{1+\exp(\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}))}. \tag{19}$$

Analogous to (17), solving (18) is equivalent to fitting a logistic regression working model (19). Based on Lemma 3, we can derive the optimal decision rule for the binary outcome as

$$\begin{aligned} d_0(\mathbf{x}) &= \mathrm{argmax}_{k\in\{1,\cdots,K\}}\mathbf{P}[R=1|\mathbf{x}, A=i] \\ &= \mathrm{argmax}_{k\in\{1,\cdots,K\}}\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}), \end{aligned} \tag{20}$$

which can be also interpreted as the least angle decision rule. Then we can fit a weighted logistic regression with modified covariates $\mathbf{x}^* = \mathbf{x}\mathbf{w}^T$ by modeling

$$\mathbf{P}[R = 1|\mathbf{x}, A] = \frac{\exp(\mathrm{Tr}(\mathbf{B}^T\mathbf{x}^*))}{1 + \exp(\mathrm{Tr}(\mathbf{B}^T\mathbf{x}^*))}, \tag{21}$$

and estimate the coefficient matrix $\mathbf{B}$ by maximum likelihood estimation

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times(K-1)}}{\mathrm{argmin}} \quad l(\mathbf{B}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{R_i\mathrm{Tr}(\mathbf{B}^T\mathbf{x}_i^*)}{\pi(A_i,\mathbf{x}_i)} + \frac{1}{n}\sum_{i=1}^{n}\frac{\log(1+\exp(\mathrm{Tr}(\mathbf{B}^T\mathbf{x}_i^*)))}{\pi(A_i,\mathbf{x}_i)} + \lambda J(\mathbf{B}), \tag{22}$$

where $J(\mathbf{B})$ is either the mixed l_1/l_2 penalty or the nuclear norm penalty under different model assumptions. We can use the accelerated proximal gradient method to solve this problem (Beck and Teboulle (2009)). However, the gradient of the exponential loss function for this model may need relatively large computational time. Efficient group coordinate descent proposed by Breheny and Huang (2015) can be an alternative to solve Model (22) with the mixed l_1/l_2 penalty by vectorizing the modified covariates.

3.2 Survival Outcomes

When R is the survival outcome, due to the potential censoring of observations, we do not always observe the exact outcomes of patients in clinical studies. Thus R becomes a pair of random variables defined as $R = (Y, \delta) = (\tilde{Y} \wedge C, \delta)$, where $\tilde{Y}$ is the patient's survival time, C is the censoring time, and δ is an indicator about whether this patient is censored or not. Motivated by Lemma 2 and a similar derivation as in Section 3.1, we can replace squared error loss in (16) for continuous outcomes by the negative log-likelihood of the Cox model for survival outcomes. Then we have the following lemma for survival outcomes.

Lemma 4. *Under the exchange of differential and expectation condition, an optimal solution to*

$$\underset{\mathbf{f}\in\mathcal{F}}{argmin} \quad \mathbf{E}[\int_0^\tau \frac{\log \mathbf{E}[e^{\mathbf{f}^T\mathbf{w}}\mathbb{I}(Y \geq u)]}{\pi(A,\mathbf{x})} - \frac{\mathbf{f}^T\mathbf{w}}{\pi(A,\mathbf{x})}dN(u)] \tag{23}$$

is the function $\mathbf{f}^*$ *satisfying*

$$\exp(\mathbf{w}_i^T\mathbf{f}^*)\mathbf{E}[\Lambda^*(Y^{(i)})|\mathbf{x}, A = i] = \mathbf{P}[\delta = 1|\mathbf{x}, A = i] \tag{24}$$

for a monotone nondecreasing function $\Lambda^*(u)$*, where* $N(u) = \mathbb{I}(\tilde{Y} \leq u)\delta$*, and* τ *is a fixed time*

point with $\mathbf{P}[\tilde{Y} \geq \tau] > 0$. *If the censoring time is non-informative and the censoring rate for each treatment group is the same, then*

$$argmax_{i\in\{1,\cdots,K\}} - \mathbf{w}_i^T f^* = argmax_{i\in\{1,\cdots,K\}} \mathbf{E}[\Lambda(Y)|\mathbf{x}, A = i]. \tag{25}$$

Using Lemma 4, the optimal decision rule for the survival outcome can be written as

$$d_0(\mathbf{x}) = \text{argmax}_{k\in\{1,\cdots,K\}} \mathbf{w}_i^T(-\mathbf{f}^*). \tag{26}$$

This is equivalent to fitting a weighted Cox Proportional Hazard (CPH) model with modified covariates $\mathbf{x}^* = \mathbf{x}\mathbf{w}^T$, by defining the hazard function as

$$\lambda(t|\mathbf{x}, A) = \lambda_0(t)e^{\text{Tr}(\mathbf{B}^T\mathbf{x}^*)}, \tag{27}$$

where $\lambda_0(t)$ is a baseline hazard function. Then we can estimate the coefficient matrix $\mathbf{B}$ by maximum likelihood estimation such as

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times(K-1)}}{\text{argmin}} \quad l(\mathbf{B}) = \frac{1}{n}\sum_{i:\delta_i=1}\{-\frac{Y_i\text{Tr}(\mathbf{B}^T\mathbf{x}_i^*)}{\pi(A_i,\mathbf{x}_i)} + \frac{1}{\pi(A_i,\mathbf{x}_i)}\log\sum_{j:Y_j\geq Y_i}\exp(\text{Tr}(\mathbf{B}^T\mathbf{x}_i^*))\} + \lambda J(\mathbf{B}), \tag{28}$$

where $J(\mathbf{B})$ is either the mixed l_1/l_2 penalty or the nuclear norm penalty under different model assumptions. As the gradient of the Cox loss function for this model requires heavy computation, similar to Section 3.1, efficient group coordinate descent (Breheny and Huang (2015)) can be used to optimize (28) with the mixed l_1/l_2 penalty through vectorizing the modified covariates.

Note that the modified covariates $\mathbf{x}^*$ in Equation (27) contain the treatment information that can be incorporated into the baseline hazard function. Thus baseline hazard functions can be different for different treatments. For Lemma 4, we assume the censoring rate to be equal for all treatment groups so that our proposed method can be directly extended to the survival outcome. This assumption can possibly be removed by estimating the censoring rate for each group and then adjusting Equation (24).

4 Theoretical Properties of AD-learning

In this section, we show our proposed AD-learning is consistent under some mild conditions and establish finite value reduction bounds for our method. We first state the generalized margin

condition used in our theory.

Assumption 1. *For any $\epsilon > 0$, there exists some constants $C > 0$ and $\alpha > 0$ such that*

$$\mathbf{P}[|(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{f}_0(\mathbf{x})| \leq \epsilon] \leq C\epsilon^{\alpha} \tag{29}$$

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

Assumption 1 is an extension of margin condition used in binary classification problems to obtain sharper bounds on the excess 0-1 risk (Audibert et al. (2007)). For our ITR problems, this generalized margin condition characterizes the behavior of the decision function vector $\mathbf{f}_0(\mathbf{x})$ around the boundary among different treatment regions, thus the level of difficulty in finding the optimal ITR. In the literature, Zhao et al. (2012) used a similar assumption in the binary ITR problem. Using Assumption 1, we have the following theorem for the value reduction bound.

Theorem 1. *For the estimator $\hat{\mathbf{f}}_n$ by our proposed AD-learning and the corresponding ITR $\hat{d}_n$, we have*

$$V(d_0) - V(\hat{d}_n) \leq \frac{2K(K-1)}{1 - C(K)} (E||\mathbf{f}_0 - \hat{\mathbf{f}}_n||_2^2)^{\frac{1}{2}}. \tag{30}$$

Furthermore, if Assumption 1 holds, we can improve the bound by

$$V(d_0) - V(\hat{d}_n) \leq C_1(K, \alpha) (E||\mathbf{f}_0 - \hat{\mathbf{f}}_n||_2^2)^{\frac{1+\alpha}{2+\alpha}}, \tag{31}$$

where $C_1(K, \alpha)$ is the constant that only depends on K and α.

Remark 1. *Based on* (31)*, we can see that when $\alpha = 0$ and $C = 1$, Assumption* (1) *always holds for any $\epsilon > 0$. In this case,* (31) *reduces to* (30)*. Based on* (29)*, if α increases, the outcomes corresponding to various treatments become more different. As a result, the corresponding exponent $\frac{1+\alpha}{2+\alpha}$ becomes larger, and consequently a sharper bound in* (31) *can be obtained.*

Theorem 1 gives an upper bound for the value function reduction in terms of the prediction error. For simplicity, we first consider Model (14) with equal $\pi(A_i, \mathbf{x}_i)$ for each treatment. Then we can use the main idea from Lounici et al. (2009). We first vectorize the multiple responses and the coefficient $\mathbf{B}$ so that the model becomes

$$\underset{\beta \in \mathbb{R}^{p(K-1)}}{\operatorname{argmin}} \quad \frac{1}{n(K-1)} \sum_{k=1}^{K-1} (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda ||\beta||_{2,1}, \tag{32}$$

where vector $\mathbf{y}_k = KR\mathbf{w}_k \in \mathbb{R}^n$ for $k = 1, \cdots, K-1$ and $\mathbf{X}$ is a design matrix with the i-th row being the i-th patient covariates $\mathbf{x}_i$. Denote each column of the coefficients $\mathbf{B}$ as $\boldsymbol{\beta}_k$, for $k = 1, \cdots, K-1$. Then $\boldsymbol{\beta} \in \mathbb{R}^{p(K-1)}$ is formed by stacking the coefficient $\boldsymbol{\beta}_k$, for $k = 1, \cdots, K-1$. We further define the $(K-1)n \times p(K-1)$ block diagonal matrix $\mathbf{Z}$ with its k-th block formed by the design matrix X.

We assume the underlying true $\mathbf{f}_0$ is linear with coefficient $\boldsymbol{\beta}_0$. Define $S(\boldsymbol{\beta}) = \{j : \beta_{kj} \neq 0, k = 1, \cdots, K-1\}$ and the cardinality of $S(\boldsymbol{\beta})$ as $||S(\boldsymbol{\beta})||_0$. We make the following two assumptions as in Lounici et al. (2009). The first one is the Restricted Eigenvalue (RE) assumption considered by Bickel et al. (2009) with an extension to the mixed l_1/l_2 norm.

Assumption 2. *[RE(s)] For any nonzero $\boldsymbol{\beta}$ with $||S||_0 \leq s$ and $||\boldsymbol{\beta}_{S^c}||_{2,1} \leq 3||\boldsymbol{\beta}_S||_{2,1}$, there exists a positive real number $\rho(s)$ such that*

$$\sqrt{\boldsymbol{\beta}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}} \geq \rho(s)||\boldsymbol{\beta}_S||, \tag{33}$$

where S denotes the short notation of $S(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$.

The next assumption is to control the stochastic error term in Model (14) with the bounded variance assumption.

Assumption 3. *(1) Assume that the random error $e_{ki} = (y_{ki} - \mathbf{x}_i^T\boldsymbol{\beta}_k)$; $i = 1, \cdots, n$, $k = 1, \cdots, K-1$, are independent among different i with mean zero and finite variance $\mathbf{E}[e_{ki}^2] \leq \sigma^2$.*

(2) There exists a constant c such that $\max_{1\leq i\leq n}\max_{1\leq j\leq p}|x_{ij}| \leq c$.

With the assumptions in place, we have the following theorem.

Theorem 2. *Consider Model (14), for $p \geq 3$ and $K, n \geq 1$. Assume $S(\boldsymbol{\beta}_0) \leq s$, Assumptions 2 and 3 and the RE(2s) assumption hold. Let*

$$\lambda = \sigma\sqrt{\frac{(\log p)^{1+\delta}}{n(K-1)}},$$

for any $\delta > 0$. Then with probability at least $1 - \frac{(2e\log p - e)c^2}{(\log p)^{1+\delta}}$, for the solution $\hat{\mathbf{B}}$ to the Model (14), we have

$$V(d_0) - V(\hat{d}_n) \leq \frac{\sqrt{K-1}K(K-1)}{1-C(K)}\frac{4\sqrt{10}c}{\rho^2(2s)}\sigma\sqrt{\frac{s(\log p)^{1+\delta}}{n}}. \tag{34}$$

Furthermore, if Assumption 1 is satisfied, we can improve the bound by

$$V(d_0) - V(\hat{d}_n) \leq C(K,\alpha)\frac{32}{\rho^2(s)}\sigma^2 s(\frac{(\log p)^{1+\delta}}{n})^{\frac{1+\alpha}{2+\alpha}}, \tag{35}$$

where $C(K,\alpha)$ only depends on K and the margin condition constant α.

Theorem 2 gives us the value reduction bound of order nearly $\frac{1}{n}$ as long as α is large enough. This value bound is consistent with l_1-PLS proposed by Qian and Murphy (2011) if we assume the underlying true function is linear. For a general function approximation, an additional approximation error to $\mathbf{f}_0(\mathbf{x})$ needs to be considered.

For Model (15), Rohde et al. (2011) has obtained the same rate $\mathbf{O}(\frac{1}{n})$ for the prediction error and thus the order of value reduction bound for Model (15) is the same as Theorem 2. For Model (22), it can be regarded as usual logistic regression with modified covariates. If we consider the mixed l_1/l_2 penalty, error bounds of the same order were developed in Meier et al. (2008). These results are applicable to our proposed AD-learning. However, to the best of our knowledge, the finite sample properties of other settings such as CPH models with the mixed l_1/l_2 penalty or low rank penalty require further developments and we leave it as the future work.

5 Simulation Study

In this section, we perform an extensive simulation study to investigate the finite sample performance of AD-learning for various types of outcomes. For all simulation settings, we consider four-armed ($K = 4$) randomized trials with equal probabilities of patients being assigned to each treatment group. For the low dimensional simulation setting, we set the sample size n to be 200, 400, and 800. The number of covariates p is set to be 20 and 40. For high dimensional simulation settings, we let the sample size be 400 and p be 1000. Each simulation is repeated for 120 times. Additional simulation results are in the supplementary material, such as settings with $n = 200$, low rank decision function simulation studies, etc.

For the implementation details of AD-learning, two types of algorithms can be applied. The first one is the accelerated proximal gradient method. In particular, Models (14) and (15) can be represented as

$$\min F(\mathbf{B}) := L(\mathbf{B}) + \lambda J(\mathbf{B}), \tag{36}$$

where $L(\mathbf{B})$ is a smooth convex function with its gradient being Lipschitz continuous and $J(\mathbf{B})$

is a non-smooth convex function, of which the proximal operator can be computed efficiently. Then we can use the accelerated proximal gradient method to solve it with low computational complexity. It achieves the optimal converge rate $\mathbf{O}(\frac{1}{m^2})$ for gradient methods, where m is the number of iterations for the algorithm. More details can be found in Nesterov (2013) and Toh and Yun (2010).

In binary and survival outcome settings, the gradient of function $L(\mathbf{B})$ may need large computational cost to calculate. To address the problem, the stochastic block coordinate decent algorithm can be applied instead when $J(\mathbf{B})$ is the mixed l_1/l_2 penalty. By using this algorithm, each gradient decent iteration can be efficiently computed. Thus the stochastic block coordinate decent algorithm may cost less time than the accelerated proximal gradient method.

The tuning parameter λ is selected based on the cross-validation procedure. The criterion is to select λ that maximizes the average of estimated value functions on the validation data set defined as

$$\hat{V}(d) = \frac{\mathbf{E}_n[R\mathbb{I}(A = d(\mathbf{x}))/\pi(A, \mathbf{x})]}{\mathbf{E}_n[\mathbb{I}(A = d(\mathbf{x}))/\pi(A, \mathbf{x})]}, \tag{37}$$

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

5.1 Study of Continuous Outcomes

When the clinical outcome R is continuous, we generate our data from Model (7). Specifically, for $i = 1, \cdots, n$, let

$$R_i = \mu(\mathbf{x}_i) + \delta(\mathbf{x}_i) + \epsilon_i,$$

where $\delta(\mathbf{x}_i) = \sum_{k=1}^{K}(\mathbf{x}_i^T\boldsymbol{\beta}_k)\mathbb{I}(A = k)$, each covariate is generated by the uniform distribution from -1 to 1, and ϵ_i follows from the standard normal distribution. For each simulation scenario, we consider $\mu(\mathbf{x}) = 1+X_1+X_2$ and consider other types of main effect functions in the supplementary material. We design the following three interaction functions similar to those in Zhou et al. (2017) and Zhang et al. (2015):

1. $\delta(\mathbf{x}) = (1 + X_1 + X_2 + X_3 + X_4)\mathbb{I}(A = 1) + (1 + X_1 - X_2 - X_3 + X_4)\mathbb{I}(A = 2) + (1 + X_1 - X_2 + X_3 - X_4)\mathbb{I}(A = 3) + (1 - X_1 - X_2 + X_3 + X_4)\mathbb{I}(A = 4)$;

2. $\delta(\mathbf{x}) = (3\mathbb{I}(X_1 \leq 0.5)(\mathbb{I}(X_2 > -0.6) - 1))\mathbb{I}(A = 1) + ((\mathbb{I}(X_3 \leq 1))(2\mathbb{I}(X_4 \leq -0.3) - 1)\mathbb{I}(A = 2) + (4\mathbb{I}(X_5 \leq 0) - 2)\mathbb{I}(A = 3) + (4\mathbb{I}(X_6 \leq 0) - 2)\mathbb{I}(A = 4)$;

3. $\delta(\mathbf{x}) = (0.2 + X_1^2 + X_2^2 - X_3^2 - X_4^2)\mathbb{I}(A = 1) + (0.2 + X_2^2 + X_3^2 - X_2^2 - X_4^2)\mathbb{I}(A = 2) + (0.2 +$

$$X_1^2 + X_4^2 - X_2^2 - X_3^2)\mathbb{I}(A = 3) + (0.2 + X_2^2 + X_3^2 - X_1^2 - X_4^2)\mathbb{I}(A = 4).$$

The first scenario corresponds to linear interaction effects. For the second scenario, we consider tree-type interaction effects. The last scenario includes polynomial interaction effects and we use degree 2 polynomials as basis functions for all methods. For each simulation scenario, we compare our proposed AD-learning using the group sparsity penalty with the following methods:

(1) l_1-PLS proposed by Qian and Murphy (2011) with basis $(1, \mathbf{x}, \mathbf{x}A)$;

(2) pairwise D-learning;

(3) the decision list (DL) method proposed by Zhang et al. (2015);

(4) adaptive contrast weighted learning (ACWL-1 and ACWL-2) methods proposed by Tao and Wang (2016);

(5) the method of virtual twins (VT) proposed by Foster et al. (2011),

where we use degree 2 polynomials as basis functions for all methods in the last scenario. Additional simulation study results on AD-learning using the low rank sparsity penalty are included in the supplementary material. In addition, we also perform the comparison between group l_1-PLS and l_1-PLS in the supplementary material, which shows little differences between l_1-PLS and group l_1-PLS in our simulation studies. This confirms our appropriate use of l_1-PLS instead of group l_1-PLS unless there are some prior information about strong group sparsity structures.

All the tuning parameters are selected via 10-fold cross-validation. We report the value functions and misclassification errors for $p = 40$ on 10000 independently generated test data in Table 1. From Table 1, we can see that our AD-learning has competitive performance among all methods. When we consider linear interaction effect, it is expected that our proposed AD-learning and l_1-PLS perform the best compared with other methods. In particular, our method will potentially be better than l_1-PLS because l_1-PLS suffers the mismatch problem discussed previously. For the second simulation scenario that corresponds to simple tree type interaction effect, while those tree based methods such as VT, DL and ACWL perform well, our method is still competitive. Similar results for $p = 20$ are included in the supplementary material. An interesting observation for this scenario is that although VT has the largest empirical value function among all methods, its misclassification rate is similar to that of our proposed method

when $n = 400$. One potential reason is that VT is focused on model fitting while our method directly targets on decision rules. For the last scenario, since the basis functions we used correctly identify the interaction effect, our proposed AD-learning and l_1-PLS enjoy some advantages over other methods.

Table 1: Results of average means (standard deviations) of empirical value functions and misclassification rates for four continuous-outcome simulation scenarios with 40 covariates. The best value functions and misclassification rates are in bold.

	$n = 400$		$n = 800$	
	Value	Misclassification	Value	Misclassification
		Scenario 1		
Pair-D	2.67(0.06)	0.49(0.02)	3.01(0.02)	0.32(0.02)
l_1-PLS	3.05(0.04)	0.24(0.01)	**3.15**(0.01)	0.16(0.01)
DL	2.6(0.04)	0.54(0.01)	2.78(0.02)	0.47(0.01)
ACWL-1	2.69(0.05)	0.46(0.01)	2.9(0.02)	0.37(0.01)
ACWL-2	2.77(0.05)	0.43(0.01)	3.02(0.01)	0.31(0.01)
VT	2.66(0.03)	0.5(0.01)	2.81(0.02)	0.45(0.01)
Group-AD	**3.06**(0.05)	**0.22**(0.02)	3.14(0.03)	**0.15**(0.02)
		Scenario 2		
Pair-D	2.84(0.12)	0.32(0.04)	2.93(0.1)	0.3(0.03)
l_1-PLS	2.93(0.11)	0.36(0.04)	3.01(0.1)	0.32(0.04)
DL	2.89(0.12)	0.34(0.04)	3.04(0.11)	0.28(0.04)
ACWL-1	2.76(0.11)	0.38(0.02)	2.96(0.11)	0.32(0.02)
ACWL-2	2.81(0.11)	0.38(0.02)	3.03(0.1)	0.29(0.03)
VT	**3.07**(0.09)	**0.31**(0.02)	**3.12**(0.1)	**0.27**(0.02)
Group-AD	2.97(0.1)	**0.31**(0.03)	2.97(0.1)	0.3(0.03)
		Scenario 3		
Pair-D	1.2(0.03)	0.75(0.03)	1.2(0.03)	0.75(0.03)
l_1-PLS	1.42(0.18)	0.61(0.13)	1.58(0.22)	0.47(0.18)
DL	1.38(0.08)	0.64(0.06)	1.5(0.08)	0.57(0.06)
ACWL-1	1.29(0.08)	0.7(0.04)	1.49(0.07)	0.56(0.05)
ACWL-2	1.3(0.07)	0.69(0.04)	1.57(0.06)	0.51(0.05)
VT	1.39(0.05)	0.64(0.03)	1.44(0.04)	0.6(0.03)
Group-D	**1.57**(0.14)	**0.5**(0.11)	**1.76**(0.04)	**0.3**(0.05)

5.2 Study of Binary and Survival Outcomes

For the binary outcome R, the dataset is independently generated by the logistic regression model

$$\text{logit}(\mathbf{P}[R_i = 1]) = \mu(\mathbf{x}_i) + \sum_{k=1}^{K} (\mathbf{x}_i^T \boldsymbol{\beta}_k) \mathbb{I}(A = k),$$

where the link function $\text{logit}(x) = \log \frac{x}{1-x}$. We consider same interaction effects as the first two scenarios of the continuous outcome simulation study.

Since pairwise D-learning and ACWL are not intended for the binary outcome, after modifying the l_1-PLS by using l_1 penalized logistic regression (l_1-PLR), we compare l_1-PLR, DL and VT with our AD-learning. Table 2 shows the value functions and misclassification rates for $p = 40$ and $n = 400, 800$. We can see that our proposed AD-learning has largest value functions and

lowest misclassification rates in both scenarios. Moreover, there are some mismatches in model based methods such as l_1-PLS, where the misclassification rates and the value functions are both high. One potential reason is the mismatch between the optimization criterion and the tuning procedure in l_1-PLS. The other potential reason is the mismatch between minimizing prediction error and maximizing value function in model based methods.

Table 2: Results of average means (standard deviations) of empirical value functions and misclassification rates for two binary-outcome simulation scenarios with 40 covariates. The best value functions and misclassification rates are in bold.

	$n = 400$		$n = 800$	
	Value	Misclassification	Value	Misclassification
		Scenario 1		
l_1-PLR	0.88(0.01)	0.58(0.02)	0.91(0)	0.45(0.02)
DL	0.85(0.01)	0.67(0.01)	0.87(0.01)	0.61(0)
VT	0.84(0.01)	0.68(0.01)	0.84(0)	0.69(0)
Binary-AD	**0.9**(0.01)	**0.44**(0.02)	**0.92**(0)	**0.32**(0.02)
		Scenario 2		
l_1-PLR	0.83(0.01)	0.66(0.05)	0.86(0)	0.61(0.05)
DL	0.81(0.01)	0.53(0.01)	0.85(0.01)	0.44(0.01)
VT	0.83(0.01)	0.43(0.01)	0.83(0.01)	0.51(0)
Binary-AD	**0.86**(0.01)	**0.43**(0.04)	**0.87**(0.01)	**0.4**(0.04)

Next we consider R to be the outcome of time to event. The simulated data are generated by the following model with the exponential distribution

$$R_i = \exp(\lambda_i),$$

where exp denotes the exponential distribution and $\lambda_i = \mu(\mathbf{x}_i) + \sum_{k=1}^{K}(\mathbf{x}_i^T\boldsymbol{\beta}_k)\mathbb{I}(A = k)$ for $i = 1, \cdots, n$. The censoring time $C_i; i = 1, \cdots, n$, are generated from an exponential distribution with mean θ to induce around 25% censoring rate. We consider the same settings as those in the binary case. For comparisons, we apply the l_1 penalized CPH models and compare it with AD-learning, since other methods we use previously are not designed for the survival outcome. From Table 3 with $p = 40$, we can see that our proposed AD-learning has clear advantages over l_1-CPH. In addition, we also observe the mismatch phenomena of l_1-CPH in Scenario 2 of Table 3.

Table 3: Results of average means (standard deviations) of empirical value functions and misclassification rates for two survival-outcome simulation scenarios with 40 covariates. The best value functions and misclassification rates are in bold.

	$n = 400$		$n = 800$	
	Value	Misclassification	Value	Misclassification
		Scenario 1		
l_1-CPH	41.35(2.2)	0.33(0.04)	45.05(1.1)	0.21(0.02)
Surv-AD	**43.91**(1.3)	**0.25**(0.02)	**45.56**(1.06)	**0.18**(0.01)
		Scenario 2		
l_1-CPH	21.95(0.63)	0.57(0.04)	**23.21**(0.59)	0.5(0.04)
Surv-AD	**22.1**(0.62)	**0.46**(0.02)	22.78(0.53)	**0.44**(0.02)

5.3 Study of High Dimensional Problems

We evaluate our AD-learning performance for high dimensional settings. We consider the sample size $n = 400$ so that each treatment group has roughly 100 patients and number of covariates $p = 800$. Scenarios 1-2, 3-4, 5-6 correspond to continuous, binary, and survival outcomes respectively. The interaction effects considered here are the same as the first two scenarios in the continuous setting in Section 5.1.

From Table 4, we can find that our proposed AD-learning performs better than l_1-PLS. One of the possible reasons is that our proposed method tends to select right covariates for the interaction effect function due to the direct learning of the decision rule. An interesting observation is that although pairwise D-learning has the lowest misclassification rate in Scenario 2, its corresponding value function is the lowest. This mismatch comes from the potential sub-optimality of pairwise comparisons.

Table 4: Results of average means (standard deviations) of empirical value functions and misclassification rates for six high dimensional simulation scenarios. The best value functions and misclassification rates are in bold.

	Method	Value	Misclassification
Scenario 1	l_1-PLS	5.3(0.02)	0.17(0.01)
	Pair-D	4.51(0.14)	0.47(0.03)
	Group-AD	**5.31**(0.04)	**0.15**(0.02)
Scenario 2	l_1-PLS	5.64(0.03)	0.22(0.01)
	Pair-D	5.51(0.02)	**0.2**(0.01)
	Group-AD	**5.65(0.04)**	0.21(0.01)
Scenario 3	l_1-PLR	0.88(0.02)	0.64(0.04)
	Binary-AD	**0.92**(0.02)	**0.46**(0.06)
Scenario 4	l_1-PLR	0.84(0.01)	0.7(0.02)
	Binary-AD	**0.87**(0.01)	**0.45**(0.03)
Scenario 5	l_1-CPH	771.35(126.2)	0.41(0.09)
	Surv-AD	**1004.57**(40.19)	**0.2**(0.02)
Scenario 6	l_1-CPH	150.87(7.71)	0.63(0.02)
	Surv-AD	**158.92**(4.73)	**0.45**(0.02)

6 Real Data Applications

In this section, we perform a real data analysis to further evaluate our proposed AD-learning. We consider a clinical trial dataset from "AIDS Clinical Trials Group (ACTG) 175" in Hammer et al. (1996) to study whether there is a subgroup of patients suitable for different combination treatments of AIDS. In this study, with equal probabilities, a total number of 2139 patients with HIV infection were randomly assigned into four treatment groups: zidovudine (ZDV) monotherapy, ZDV combined with didanosine (ddI), ZDV combined with zalcitabine (ZAL), and ddI monotherapy.

We choose 12 baseline covariates in our model: age (year), weight(kg), CD4+T cells amount at baseline, CD8 amount at baseline, Karnofsky score (scale at 0-100), gender (1 = male, 0 = female), race (1 = non white, 0 = white), homosexual activity (1 = yes, 0 = no), history of intravenous drug use (1 = yes, 0 = no), symptomatic status (1=symptomatic, 0=asymptomatic), antiretroviral history (1=experienced, 0=naive) and hemophilia (1=yes, 0=no). The first five covariates are continuous and have been scaled before estimation. The remaining seven covariates are binary categorical variables.

We consider two outcomes for our analysis. The first outcome is the difference between the early stage (around 25 weeks) CD4+ T (cells/mm^3) cell amount and the baseline CD4+ T cells prior to the trial. This was also studied in Lu et al. (2013) and Fan et al. (2016). Using this short term outcome, our goal is to use AD-learning to find the short term optimal ITR for each patient with AIDS among four treatment groups. We report the estimator of the coefficient $\mathbf{w}_i^T\mathbf{B}^T$ for each treatment in Table 5.

Table 5: Results of coefficients estimation for comparison functions.

Variable Name (1-7)	ZDV	ZDV+ddI	ZDV+Zal	ddI
Intercept	−49.86	44.66	−3.53	8.73
Age	−0.47	4.33	−3.34	−0.52
Weight	0	0	0	0
Karnofsky Score	0	0	0	0
CD4 baseline	3.58	−14.79	−14.78	9.46
Days pre-anti-retroviral therapy	0	0	0	0
Hemophilia	0	0	0	0
Homosexual activity	−0.28	−3.96	0.65	3.60
History of drug use	−2.50	8.20	4.03	−9.74
Race	0	0	0	0
Gender	0	0	0	0
Antiretroviral history	0	0	0	0
Symptomatic indicator	0	0	0	0

In Table 5, we can see that four covariates including Age, CD4 baseline, homosexual activity and history of drug use, are identified to play an important role in our estimated optimal ITRs. These variables were also identified in the previous literature such as Lu et al. (2013) and Fan et al. (2016). According to the analysis in Hammer et al. (1996), ZDV alone is inferior to the other treatments, which is also confirmed in our estimated ITR. Based on the CD4 change in the early stage, Zal treatment is generally not recommended in our finding with one possible reason that Zal has the most serious adverse event compared with ZDV and ddI (Kakuda (2000)). According to our estimated ITRs, those old patients with small amount of CD4 T cell baseline and having history of drug use but not homosexual activity, are recommended to take ZDV + ddI. The patients with large amount of CD4 T cell baseline and history of homosexual activity but not drug use history, are more advisable to take ddI alone.

To evaluate the performance of our proposed AD-learning, we randomly split the data into five folds and use four folds to train the model. We evaluate our method on the remaining one fold of data based on the empirical value function. We repeat this procedure for 1000 times. From Table 6, we can see our AD-learning has the largest value.

Table 6: Results of empirical value functions on one fold of testing data. The best empirical value function is in bold.

l_1-PLS	Pair-D	DL	ACWL-1	ACWL-2	VT	AD low rank	AD group
53.73 (0.33)	57.17 (0.40)	53.25 (0.47)	52.74 (0.45)	54.04 (0.45)	54.84 (0.45)	50.48 (0.38)	**59.69(0.39)**

The second outcome is patients' time to event. Using this long term outcome, our second goal is to find the long term optimal ITR for patients among four treatment groups. The AIDS data consist of 2139 patient time to event responses with around 75% censor rate during the four-year long trial study. We use our proposed Model (23) to estimate the optimal ITR. We report the estimates of the coefficient $\mathbf{w}_i^T\mathbf{B}^T$ for each treatment of 12 covariates in Table 7. We can see that all covariates, except the indicator of homosexual activity and symptomatic, play an important role in the estimated optimal ITR. It may not be surprising because it is a long term study and thus more complicated. Since we model via the hazard function, the smaller the coefficient is, the longer the survival time is.

Compared with the previous finding based on the short term CD4 T cells amount, covariates including age, CD4 baseline and history of drug use have the similar effect on the ZDV + ddI and ddI alone treatments. In addition, we also find that ZDV + Zal treatment may not be good

to take for the female patients with hemophilia, but may be suitable for the male patients with high Karnosky score and history of drug use. The estimated optimal ITR for other treatments can be interpreted in the similar way. In general, ZDV alone is always the least preferable among other treatments for patients and ZDV+ddI is always preferable for patients. Based on time to event outcome, ZDV + Zal is relatively more preferable than ddI alone. In addition, we evaluate our AD-learning with l_1-CPH using the same scheme based on value functions. Our AD-learning has an average value of 911.20, compared with the average value 905.02 for l_1-CPH.

Table 7: Results of coefficient estimation for survival time of failure.

Variable Name (1-7)	ZDV	ZDV+ddI	ZDV+Zal	ddI
Age	0.04	−0.11	0.04	0.03
Weight	0.11	0.02	0.02	−0.14
Karnofsky Score	0.06	0.03	−0.09	0.01
CD4 baseline	−0.04	0.04	−0.00	0.00
Days pre-anti-retroviral therapy	0.09	−0.07	-0.04	0.02
Hemophilia	0.05	−0.06	0.16	−0.15
Homosexual activity	0.00	0.00	0.00	0.00
History of drug use	0.04	−0.11	−0.12	0.18
Race	0.03	−0.04	0.01	0.01
Gender	0.31	−0.08	−0.16	−0.07
Antiretroviral history	0.17	−0.15	0.04	−0.06
Symptomatic Indicator	0.00	0.00	0.00	0.00

7 Conclusion

In this article, we propose a AD-learning method to estimate the optimal ITRs in multiple treatment settings for various types of outcomes. Our proposed method provides a clear geometric interpretation about the relative effectiveness of treatments for patients, which is quantified by angles in the Euclidean space. Our proposed AD-learning is robust to model misspecification. By incorporating group or low rank sparsity, our AD-learning can further improve the estimation of decision rules and interpretation, especially for high dimensional settings. The competitive performance of our method has been demonstrated via the simulation studies and data applications.

Several possible extensions can be explored for future study. Our proposed method for the survival outcome is based on the non-informative censoring and Cox proportional hazard assumption. It will be interesting to develop methods for more complex settings. In order to use nonlinear functions to approximate $\mathbf{f}_0(\mathbf{x})$, we can use different types of basis functions such polynomials or wavelet functions. It will be also interesting to develop kernel methods for our AD-learning, such as multiple kernel learning (Bach et al. (2004)). Finally, the current AD-learning focuses on

a single decision point. It will be worthwhile to develop the corresponding methods for multiple decision points (Zhao et al., 2015a; Liu et al., 2016).

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.

Appendix

Proof of Lemma 1

Let $g(f) = \mathbf{E}[\frac{1}{\pi(A,\mathbf{x})}(KR\mathbf{w} - \mathbf{f}(\mathbf{x}))^T\Sigma(KR\mathbf{w} - \mathbf{f}(\mathbf{x}))]$. Taking the derivative over f and setting it to zero, we get

$$\begin{aligned}\frac{\partial g(f)}{\partial f} &= 2\Sigma\mathbf{E}_{\mathbf{x}}\{\mathbf{E}[(\frac{KRW}{\pi(A,\mathbf{x})} - \frac{f(\mathbf{x})}{\pi(A,\mathbf{x})})|\mathbf{x}]\} \\ &= 2\Sigma\mathbf{E}_{\mathbf{x}}\{K\mathbf{E}[\frac{RW}{\pi(A,\mathbf{x})}|\mathbf{x}] - f(\mathbf{x})|\mathbf{x}]\} = 0.\end{aligned}$$

Proof of Lemma 2

Let $g(f) = \mathbf{E}[\frac{1}{\pi(A,\mathbf{x})}(\frac{K}{K-1}R - \mathbf{w}^T\mathbf{f}(\mathbf{x}))^T(\frac{K}{K-1}R - \mathbf{w}^T\mathbf{f}(\mathbf{x}))]$. Taking the derivative over f and setting it to zero, we get

$$\begin{aligned}\frac{\partial g(f)}{\partial f} &= \mathbf{E}_{\mathbf{x}}\{\mathbf{E}[W(\frac{KR}{(K-1)\pi(A,\mathbf{x})} - \frac{W^T f(\mathbf{x})}{\pi(A,\mathbf{x})})|\mathbf{x}]\} \\ &= \mathbf{E}_{\mathbf{x}}\{\frac{K}{K-1}\mathbf{E}[\frac{RW}{\pi(A,\mathbf{x})}|\mathbf{x}] - \frac{K}{K-1}f(\mathbf{x})|\mathbf{x}]\} = 0,\end{aligned}$$

where the second equality holds because $\mathbf{E}[\frac{WW^T}{\pi(A,\mathbf{x})}|\mathbf{x}] = \frac{K}{K-1}I_{K-1}$ by definition. Thus $\mathbf{f}_0(\mathbf{x})$ is an optimal solution.

Proof of Lemma 3

Let $g(f) = \mathbf{E}[-\frac{R\mathbf{w}^T f}{\pi(A,\mathbf{x})} + \frac{\log(1+\exp(\mathbf{w}^T f))}{\pi(A,\mathbf{x})}]$. Taking the derivative over f and setting it to zero,

we get

$$\begin{aligned}\frac{\partial g(f)}{\partial f} &= 2\mathbf{E}_{\mathbf{x}}\{\mathbf{E}[(\frac{RW}{\pi(A,\mathbf{x})} - \frac{W\exp(\mathbf{w}^T f)}{1+\exp(\mathbf{w}^T f))\pi(A,\mathbf{x})})|\mathbf{x}]\} \\ &= 2\mathbf{E}_{\mathbf{x}}\{\sum_{i=1}^{K}\mathbf{w}_i\mathbf{P}[R=1|\mathbf{x},A=i] - \sum_{i=1}^{K}\mathbf{w}_i\frac{\exp(\mathbf{w}_i^T f)}{1+\exp(\mathbf{w}_i^T f)}\} \\ &= 0.\end{aligned}$$

If $\mathbf{P}[R=1|\mathbf{x},A=i] = \frac{\exp(\mathbf{w}_i^T f^*)}{1+\exp(\mathbf{w}_i^T f^*)}$, then f^* is an optimal solution to (18).

Proof of Lemma 4

Let $g(f) = \mathbf{E}[\int_0^\tau \frac{\log \mathbf{E}[e^{f^T\mathbf{w}}\mathbb{I}(Y\geq u)]}{\pi(A,\mathbf{x})} - \frac{f^T\mathbf{w}}{\pi(A,\mathbf{x})}dN(u)]$. Taking the derivative over f and setting it to zero, we get

$$\begin{aligned}\frac{\partial g(f)}{\partial f} &= \mathbf{E}_{\mathbf{x}}\{\int_0^\tau \sum_{i=1}^{K}\mathbf{w}_i\mathbf{E}[\mathbb{I}(Y\geq u)\lambda_i(u,\mathbf{x})|\mathbf{x},A=i] - \frac{\mathbf{w}_i\exp(\mathbf{w}_i^T f)\mathbb{I}(Y^{(i)}\geq u)\mathbf{E}[\mathbb{I}(Y\geq u)\lambda(u,\mathbf{x})|\mathbf{x}]}{\mathbf{E}[\exp(W^T f)\mathbb{I}(Y\geq u)]}du\} \\ &= \mathbf{E}_{\mathbf{x}}\{\int_0^\tau \sum_{i=1}^{K}\mathbf{w}_i(\mathbf{E}[\mathbb{I}(Y\geq u)\lambda_i(u,\mathbf{x})|\mathbf{x},A=i] - \exp(\mathbf{w}_i^T f)\Lambda^*(Y^{(i)}))du\} \\ &= 0,\end{aligned}$$

where $\lambda_i(u,\mathbf{x})$ is the hazard function for the i-th treatment and $\Lambda^*(Y)$ is the cumulative hazard function. Then we get a sufficient condition that if $\exp(\mathbf{w}_i^T f)\Lambda^*(Y^{(i)}) = \mathbf{P}[\delta=1|\mathbf{x},A=i]$, then f^* is an optimal solution. If the censoring time in each treatment group is the same, then we get (25).

Proof of Theorem 1

For any ITR d, we have

$$
\begin{aligned}
V(d) &= \mathbf{E}[\sum_{k=1}^{K} \mathbf{E}[R|\mathbf{x}, A=k]\mathbb{I}(d(\mathbf{x})=k)] \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K}(1-C(K))\mathbf{E}[R|\mathbf{x}, A=k]\mathbb{I}(d(\mathbf{x})=k) \\
&+ \sum_{j=1}^{K} C(K)\mathbf{E}[R|\mathbf{x}, A=j]\} - \frac{C(K)}{1-C(K)}\sum_{j=1}^{K}\mathbf{E}[R|\mathbf{x}, A=j]] \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K} \mathbf{E}[R|\mathbf{x}, A=k]\mathbb{I}(d(\mathbf{x})=k) \\
&+ \sum_{j=1}^{K} C(K)\mathbf{E}[R|\mathbf{x}, A=j]\sum_{i\neq j}^{K}\mathbb{I}(d(\mathbf{x})=i)\}] - \Delta \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K} \mathbf{E}[R|\mathbf{x}, A=k]\mathbb{I}(d(\mathbf{x})=k) \\
&+ \sum_{i=1}^{K}\sum_{j\neq i}^{K} C(K)\mathbf{E}[R|\mathbf{x}, A=j]\mathbb{I}(d(\mathbf{x})=i)\}] - \Delta \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K}(\mathbf{E}[R|\mathbf{x}, A=k] \\
&+ \sum_{j\neq k}^{K} C(K)\mathbf{E}[R|\mathbf{x}, A=k])\mathbb{I}(d(\mathbf{x})=k)\}] - \Delta \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K}\mathbf{w}_k^T\mathbf{E}[\frac{RW}{\pi(A,\mathbf{x})}|\mathbf{x}]\mathbb{I}(d(\mathbf{x})=k)\}] - \Delta \\
&= \mathbf{E}[\frac{1}{1-C(K)}\{\sum_{k=1}^{K}\mathbf{w}_k^T\mathbf{f}_0(\mathbf{x})\mathbb{I}(d(\mathbf{x})=k)\}] - \Delta,
\end{aligned}
\tag{38}
$$

where $\Delta = \mathbf{E}[C(K)\sum_{j=1}^{K}\mathbf{E}[R|\mathbf{x}, A=j]]$ that does not depend on the ITR d. Then we can obtain

the value reduction bound between the optimal ITR d_0 and our estimated ITR $\hat{d}$ by using (38):

$$
\begin{aligned}
&V(d_0) - V(\hat{d}) \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{k=1}^{K}\mathbf{w}_k^T\mathbf{f}_0(\mathbf{x})(\mathbb{I}(d(\mathbf{x})=k) - \mathbb{I}(\hat{d}(\mathbf{x})=k)\}] \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{i\neq j}|\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x})|\mathbb{I}(d(\mathbf{x})=i,\hat{d}(\mathbf{x})=j)\}] \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{i\neq j}|\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x})|\mathbb{I}(\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x}))(\mathbf{w}_i^T\hat{f}(\mathbf{x}) - w_j^T\hat{f}(\mathbf{x})<0)\}] \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{i\neq j}|\mathbf{w}_i^T(\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})) - w_j^T(\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x}))| \\
&\mathbb{I}(\mathbf{w}_i^T(\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x}))w_j^T(\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x}))<0)\}] \\
\leq &\frac{1}{1-C(K)}\sum_{i\neq j}(\mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2 + \mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2) \\
\leq &\frac{2K(K-1)}{1-C(K)}(\mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2^2)^{\frac{1}{2}},
\end{aligned}
\tag{39}
$$

where the second to last inequaltiy holds by using the Hölder and Minkowski inquality together with $||\mathbf{w}_i|| = 1$ for $i = 1, \cdots, K$. Furthermore, if we assume Assumption 1 holds, then we can further bound the value reduction by

$$
\begin{aligned}
&V(d_0) - V(\hat{d}) \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{i\neq j}|\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x})|\mathbb{I}(\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x})(\mathbf{w}_i^T\hat{f}(\mathbf{x}) - w_j^T\hat{f}(\mathbf{x})<0)\}] \\
\leq &\frac{1}{1-C(K)}\mathbf{E}[\{\sum_{i\neq j}\epsilon\mathbb{I}(|(\mathbf{w}_i - w_j)^T\mathbf{f}_0(\mathbf{x})| < \epsilon)\mathbb{I}((\mathbf{w}_i - w_j)^T\mathbf{f}_0(\mathbf{x}))((\mathbf{w}_i - w_j)^T\hat{f}(\mathbf{x})) < 0)\}] \\
&+ \frac{1}{1-C(K)\epsilon}\mathbf{E}[\{\sum_{i\neq j}(\mathbf{w}_i^T\mathbf{f}_0(\mathbf{x}) - w_j^T\mathbf{f}_0(\mathbf{x}))^2\mathbb{I}((\mathbf{w}_i - w_j)^T\mathbf{f}_0(\mathbf{x}))((\mathbf{w}_i - w_j)^T\hat{f}(\mathbf{x})) < 0)\}] \\
\leq &\frac{1}{1-C(K)}\sum_{i\neq j}\epsilon\mathbf{P}[|(\mathbf{w}_i - w_j)^T\mathbf{f}_0(\mathbf{x})| < \epsilon] + \frac{2}{\epsilon}(\mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2^2 + \mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2^2) \\
\leq &\frac{1}{1-C(K)}\sum_{i\neq j}C\epsilon^{\alpha+1} + \frac{4}{\epsilon}\mathbf{E}||\mathbf{f}_0(\mathbf{x}) - \hat{f}(\mathbf{x})||_2^2),
\end{aligned}
\tag{40}
$$

for any $\epsilon > 0$. We can then minimize right hand side above over ϵ and get the desired bound

$$V(d_0) - V(\hat{d}_n) \leq C_1(K,\alpha)(E||\mathbf{f}_0 - \hat{\mathbf{f}}_n||_2^2)^{\frac{1+\alpha}{2+\alpha}}.$$

Proof of Theorem 2

Define $\boldsymbol{\beta}^j = (\boldsymbol{\beta}_{kj}, k = 1, \cdots, (K-1))^T$, and let $\lambda = \sigma\sqrt{\frac{(\log p)^{1+\delta}}{n(K-1)}}$. With probability at least $1 - \frac{(2e\log p - e)c}{(\log p)^{1+\delta}}$, we have the following inequality

$$\begin{aligned}\frac{1}{n(K-1)}||\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||^2 + \lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{2,1} \leq \\ \leq \frac{1}{n(K-1)}||\mathbf{Z}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||^2 + 4\lambda \sum_{j \in S(\boldsymbol{\beta})} ||\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j||,\end{aligned} \tag{41}$$

for any $\boldsymbol{\beta}$. This was previously shown in Theorem 5.2 by Lounici et al. (2009). Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Then with probability at least $1 - \frac{(2e\log p - e)c}{(\log p)^{1+\delta}}$, we have

$$\begin{aligned}\frac{1}{n(K-1)}||\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||^2 &\leq 4\lambda \sum_{j \in S(\boldsymbol{\beta})} ||\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j|| \\ &\leq 4\lambda\sqrt{s}||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S||\end{aligned}$$

and

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{2,1} \leq 4||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S||,$$

which implies $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{S^c} \leq 3||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S||$. Then by the RE(s) assumption, with probability at least $1 - \frac{(2e\log p - e)c}{(\log p)^{1+\delta}}$, we have

$$\begin{aligned}\frac{1}{n(K-1)}||\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||^2 &\leq 4\lambda\sqrt{s}||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_S|| \\ &\leq 4\lambda\sqrt{s}\frac{||\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||}{\rho(s)\sqrt{n}},\end{aligned}$$

such that we can bound the empirical error by

$$\frac{1}{n}||\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||^2 \leq \frac{16(K-1)}{\rho(s)}\sigma^2 s\frac{(\log p)^{1+\delta}}{n}.$$

With the RE(2s) assumption, we can further show that with the same probability

$$\frac{1}{\sqrt{K-1}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| \leq \frac{4\sqrt{10}}{\rho^2(2s)}\sigma\sqrt{\frac{s(\log p)^{1+\delta}}{n}}.$$

Combining with Theorem 1, we get the value reduction bound

$$V(d_0) - V(\hat{d}_n) \leq \frac{\sqrt{K-1}K(K-1)}{1-C(K)} \frac{4\sqrt{10}c}{\rho^2(2s)} \sigma \sqrt{\frac{s(\log p)^{1+\delta}}{n}}.$$

Together with our margin condition, we can directly get the corresponding improved bound (31).

References

J.-Y. Audibert, A. B. Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

X. Bai, A. A. Tsiatis, W. Lu, and R. Song. Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Analysis*, pages 1–20, 2016.

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.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187, 2015.

L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

S. Chen, L. Tian, T. Cai, and M. Yu. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 2017.

Y. Cui, R. Zhu, M. Kosorok, et al. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic Journal of Statistics*, 11(2):3927–3953, 2017.

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)*, 2016.

J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

Y. Goldberg and M. R. Kosorok. Q-learning with censored data. *Annals of Statistics*, 40(1):529, 2012.

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.

R. Jiang, W. Lu, R. Song, and M. Davidian. On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

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

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

Y. Liu, Y. Wang, M. R. Kosorok, Y. Zhao, and D. Zeng. Robust hybrid learning for estimating personalized dynamic treatment regimens. *arXiv preprint arXiv:1611.02314*, 2016.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

W. Lu, H. H. Zhang, and D. Zeng. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5):493–504, 2013.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

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

S. A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6: 1073–1097, 2005.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Z. Qi and Y. Liu. D-learning to estimate optimal individual treatment rules. Technical report, Department of Statistics and Operation Research, University of North Carolina, 2017.

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

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.

A. Rohde, A. B. Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Y. Tao and L. Wang. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*, 2016.

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.

K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

C. J. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

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

T. T. Wu and K. Lange. Multicategory vertex discriminant analysis for high-dimensional data. *The Annals of Applied Statistics*, pages 1698–1721, 2010.

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.

C. Zhang and Y. Liu. Multicategory angle-based large-margin classification. *Biometrika*, 101(3): 625, 2014.

Y. Zhang, E. B. Laber, A. Tsiatis, and M. Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904, 2015.

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.

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, 2015a.

Y.-Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151, 2015b.

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.