Estimating Individualized Decision Rules with Tail Controls

Zhengling Qi\textsuperscript{1}, Jong-Shi Pang\textsuperscript{2}, and Yufeng Liu\textsuperscript{3}

\textsuperscript{1}Department of Decision Sciences, George Washington University
E-mail: qizhengling321@gmail.com
\textsuperscript{2}The Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, LA
E-mail: jongship@usc.edu
\textsuperscript{3}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
E-mail: yfliu@email.unc.edu

May 28, 2019

Abstract

With the emergence of precision medicine, estimating optimal individualized decision rules (IDRs) has attracted tremendous attention in many scientific areas. Most existing literature has focused on finding optimal IDRs that can maximize the expected outcome for each individual. Motivated by complex individualized decision making procedures and popular conditional value at risk (CVaR) measures, we propose a new robust criterion to estimate optimal IDRs in order to control the average lower tail of the subjects’ outcomes. In addition to improving the individualized expected outcome, our proposed criterion takes risks into consideration, and thus the resulting IDRs can prevent adverse events. The optimal IDR under our criterion can be interpreted as the decision rule that maximizes the “worst-case” scenario of the individualized outcome when the underlying distribution is perturbed within a constrained set. An efficient non-convex optimization algorithm is proposed with convergence guarantees. We investigate theoretical properties for our estimated optimal IDRs under the proposed criterion such as consistency and finite sample error bounds. Simulation studies and a real data application are used to further demonstrate the robust performance of our method.

Keywords: Conditional Value at Risk; Individualized decision rules; Non-convex optimization; Robustness; Tail controls

1 Introduction

Decision making is a long standing research problem in many scientific areas, ranging from engineering, management science to statistics. In the era of big data, the traditional “one fits
all” decision rules are no longer ideal in many applications due to data heterogeneity. A decision rule that works for certain subjects may not necessarily work for others. Motivated by this, it is desirable to make individualized decision rules (IDRs) that map from individual characteristics into available decision assignments. Developing effective IDRs has a wide range of applications. For example, a credit card company hopes to send a special offer for each targeted customer tailoring to his/her personal needs. An epidemiologist needs to decide whether to deliver a vaccine plan to a specific region in order to prevent the spread of diseases. In medical applications, IDRs can be developed for better prevention and treatment methods that are tailored to each individual patient. This is known as precision, or personalized, medicine.

Most existing methods in the literature, from data analytic perspective, are focused on estimating the optimal IDR that can maximize the expected outcome or minimize the expected loss for each subject, such as Qian and Murphy (2011) and Manski (2004). For example, we may want to learn an IDR \( d \) that maps a subject’s covariate \( X \in \mathcal{X} \) into a binary decision space \( A = \{1, -1\} \), i.e., \( d : \mathcal{X} \to A \), in order to maximize the expectation of \( R(d) \) or a utility function \( u \) of \( R(d) \). Here \( R(d) \) is the random outcome under IDR \( d \), which will be formally defined in the next section. Then the problem can be mathematically formulated as the following optimization problem:

\[
\max_{d \in \mathcal{D}} \mathbb{E}[R(d)],
\]

where \( \mathcal{D} \) is a structural class of IDRs specified by practitioners, such as decision trees. If the utility function \( u \) is used, then one could replace the objective function in (1) with \( \mathbb{E}[u(R(d))] \). The equivalent form of Problem (1) is

\[
\max_{t \in \mathbb{R}, d \in \mathcal{D}} t
\]

subject to \( t \leq \mathbb{E}[R(d)] \),

by using a hypographical representation (Rockafellar and Wets (2009)).

According to (2), it may be reasonable to restrict the expected outcome larger than some threshold when stochastic ups and downs of outcome \( R(d) \) can be safely averaged out. Then by implementing the IDR obtained by (2), we can guarantee on average the outcome will be better than some threshold \( t \). However, in practice, practitioners usually want to have a safe margin to protect against undesired outcomes, especially when the lower tails of the outcome are more important. For instance, in reliability engineering (Rockafellar and Royset (2010)), people are often interested in controlling the failure probability of some designed systems or structures such as buildings and bridges, instead of expected failure time. Here the selection of designs can be interpreted as IDRs based on environmental conditions, existing materials, etc. In precision medicine, sometimes the gain in the expected outcome may be very little by comparing two treatments while the tail of the potential outcome distribution is of direct interest (Wang et al. (2018)). Suppose two drugs are used to improve CD4 T cell amount of AIDS patients. Since normal range of CD4 cells is 500-1500, then in practice the event of a certain subject \( \{R_i(d) = 600\} \) may be treated as good as the event of \( \{R_i(d) = 800\} \). In contrast, the event \( \{R_i(d) = 50\} \) may be considered much worse
than $\{R_i(d) = 400\}$. Hence only using (2) to search for best IDRs may be insufficient when the tails are important or the variability of the outcomes needs to be controlled. For further illustrations, we consider the following toy example.

**Toy Example:** Figure 1 plots the conditional density of a random outcome $R$ under two treatments 1 and $-1$, given the patient’s gender, i.e., male or female. Each curve corresponds to a different Gaussian density curve. If we only consider the optimal IDR that maximizes the expected outcome, treatment 1 is more suitable for female while treatment $-1$ is better for male. However, the gain is quite little since the mean difference is only 0.1, and thus it is hard to distinguish between these two treatments given the gender information. However, if we consider the lower tails of outcomes caused by each treatment, in order to protect each person from risky scenarios, then treatment 1 is more favorable than treatment $-1$ for male, and similarly, treatment $-1$ is more preferable than treatment 1 for female. This treatment rule may be more reasonable than the previous one because we do not want to assign patients unstable treatments with potentially high risk. We will revisit this example in our numerical studies in Section 5 for further illustrations.

![Figure 1](image-url)

Figure 1: Plots of a motivating example. The dash and solid lines in the left plot show the probability densities of $\mathcal{N}(-0.1, 0.5)$ and $\mathcal{N}(0, 1)$ respectively. The dash and solid lines in the right plot correspond to the probability densities of $\mathcal{N}(0, 1)$ and $\mathcal{N}(-0.1, 0.5)$ respectively. In this example, male is more preferable to treatment 1, while female is more preferable to treatment $-1$.

As we can see from Figure 1, due to the complicated decision making procedure, only targeting on the expected outcome of each subject may not be sufficient. Risk control is necessary to prevent adverse events, i.e., the lower tails of outcomes. In this paper, motivated by the conditional value at risk (CVaR) used extensively in finance and risk management ([Rockafellar and Uryasev (2000)]),
we propose a new criterion that considers average lower tails of outcome to evaluate IDRs. The resulting IDR under our proposed criterion can optimize the outcome of each individual and provide a safe margin against adverse events jointly.

1.1 Related Literature

There is an increasing body of literature in learning optimal IDR under the framework of $[1]$. The literature spans across various fields such as statistics, economics and machine learning. In the statistics literature, most existing methods can be roughly divided into two categories: model based methods and direct search methods. Q-learning (Watkins (1989), Murphy (2005), Schulte et al. (2014)) and A-learning (Murphy (2003), Robins (2004)) are two representative model based methods. Other variants include Fan et al. (2017), Gunter et al. (2011), etc. For direct search methods, by viewing IDR problems as a weighted classification problem, Zhao et al. (2012) proposed to use the weighted support vector machine method to estimate the optimal IDR. Following that, various types of machine learning methods were proposed, such as Liu et al. (2018), Zhou et al. (2017), Zhao et al. (2015), Cui et al. (2017), Tao and Wang (2017), Laber and Zhao (2015), Zhang et al. (2015), Chen et al. (2018), etc. In addition, Tian et al. (2014), Qi and Liu (2018) and Qi et al. (2019) proposed to use regression methods to directly estimate the optimal IDR. Recently, Wang et al. (2018) proposed to use the quantile of $R(d)$ as a criterion to search for best IDRs, which is closely related to this paper. Their method can help to obtain robust optimal IDRs by controlling the lower quantile. However, it can be instable when the potential outcome distribution is discrete. More importantly, as we mentioned in the example of CD4 T cells, lower tails of the outcome should be treated differently: if both $\{R_i(d) = 50\}$ and $\{R_i(d) = 400\}$ are below the 25% quantile, the first event should be considered much worse than the latter one. More discussions on the difference between our criterion and the quantile one can be found in Section 2.

In the econometrics literature, Manski (2004) provided a comprehensive regret analysis of estimating optimal IDRs with a link to statistical decision theory (Savage (1951)). Later on, exact finite sample regret analysis was established by Stoye (2009) and Tetenov (2012). Under smooth parametric and semi-parametric settings, Hirano and Porter (2009) investigated asymptotic optimality and large sample properties of optimal IDRs. Other related work includes Chamberlain (2011), Bhattacharya and Dupas (2012), Kasy (2016). Recently, Kitagawa and Tetenov (2018) and Athey and Wager (2017) established rate-optimal regret bounds for learning optimal IDRs. All of these existing developments are based on expected outcome or expected utility. We also note that Dehejia (2008) studied the risk aversion of treatment effect evaluation by using the mean-variance trade off criterion.

We would like to point out the increasing literature of learning optimal IDRs in machine learning community, which is often referred as batch learning from bandit feedback. In particular, Beygelzimer and Langford (2009) formulated the IDR problem as a weighted supervised learning problem. Dudík et al. (2011) developed doubly robust estimators for IDR evaluations and used machine learning algorithms to search optimal IDRs. The follow up empirical applications can be found in
Li et al. (2010, 2011, 2015). Swaminathan and Joachims (2015) proposed to use variance regularized empirical risk approximation to Problem (1) in order to improve robustness and generalization performance of learning optimal IDRs. Recently, Kallus (2018) made use of matching techniques in causal inference to efficiently evaluate IDR and leveraged bi-level optimization techniques to search for the best IDRs, which improves the empirical performances. These existing methods do not consider risk control for estimation of IDRs. Finally, we note that in the reinforcement learning literature, CVaR has been used as constraints in Markov decision processes such as Tamar et al. (2015) and Chow et al. (2017).

1.2 Main Contributions and Outline

The main contributions of this paper can be summarized as follows. We leverage the CVaR criterion and propose a robust criterion to directly estimate the optimal IDR that can maximize the expected outcome while simultaneously controlling the average lower tails of the outcome. We discuss several important properties of our criterion and its practical usage by providing safety protection for implementing optimal IDRs. An efficient non-convex optimization algorithm is proposed to compute the solutions with a convergence guarantee. We establish several important theoretical properties of our estimator under the proposed criterion similar to the regret analysis in Zhao et al. (2012); Zhou et al. (2017); Kitagawa and Tetenov (2018); Athey and Wager (2017).

The remainder of this paper is organized as follows. In Section 2, supplementing the previous value function framework, we introduce a new criterion to estimate the optimal IDR by using the concept of CVaR in risk management. We present several properties of our proposed criterion. In Section 3, we discuss our statistical estimation procedure to compute optimal IDRs under our proposed criterion. An efficient non-convex optimization algorithm is presented by using some recent developments in difference of convex algorithms (DCA). In Section 4, we establish several important theoretical properties of our method based on statistical learning theory. We demonstrate our method via extensive simulation studies and a data application in Sections 5. In Section 6, we discuss several extensions of our proposed criterion from the perspectives of algorithm and modeling. Some technical results are provided in the supplementary materials.

2 Robust Criteria to Estimate Optimal IDRs

2.1 Notation and Basic Settings

We discuss our IDR problem under the Neyman-Rubin causal framework (Rubin (1974)). Suppose there is a random sample of size $n$ from a given population. For each subject in this sample, we observe covariate information $X_i \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^p$, for $i = 1, \ldots, n$. Each subject is associated with a pair of potential outcomes, $(R_i(1), R_i(-1))$, which are scalars. We can only observe either $R_i(1)$ or $R_i(-1)$ based on which treatment has been assigned. Here $A_i \in \mathcal{A} = \{1, -1\}$ denotes the treatment that $i$-th subject has received. Therefore, for each subject, we observe $(X_i, A_i, R_i)$, where $R_i = R_i(A_i)$. By doing that, we assume the stable unit treatment value assumption (SUTVA) (Rubin (1974)).
We define \( \pi(A|X) \) to be the probability of a subject being assigned treatment \( A \) given the covariates and denote \( P^d \) as the probability measure when treatment follows the decision rule \( d \). In addition, throughout this paper, we use the following two standard assumptions under the potential outcome framework.

**Assumption 2.1.** (*Ignorable assignment mechanism*). \( R_i(1), R_i(-1) \perp \perp A_i | X_i \), where \( \perp \perp \) denotes independence.

**Assumption 2.2.** (*Strong overlap assumption*). \( c \leq \pi(A|X) \leq (1-c) \) for some \( c > 0 \).

For simplicity, we assume the random outcome \( R \) has a bounded support. Without loss of generality, we assume that the larger \( R \) indicates the better condition a subject is in. Throughout this paper, we consider a randomized experiment so the propensity score \( \pi(A|X) \) is known. For observational studies, the proposed method can be applied as well by estimating the propensity score using various methods such as the logistic regression. An IDR \( d \) is defined as a measurable function mapping from the covariate space \( X \) into the treatment space \( A \). We also let \( L^r(T, F_1, P^d) \) be the space of all measurable functions such that \( \int_{T \in T} |f(T)| r P^d < \infty \), where \( F_1 \) is the corresponding \( \sigma \)-field generated by \( T := X \times A \times \mathbb{R} \).

### 2.2 Expected Value Function Framework

Before introducing our new criterion and methods, we first present the existing value function framework used by most existing methods, such as Manski (2004) and Qian and Murphy (2011). The value function is defined as

\[
V(d) = \mathbb{E}[R(d)] = \mathbb{E}_X [\mathbb{E}[R(d)|X]] = \mathbb{E}_X [\mathbb{E}[R(d)|X, A = d(X)] = \mathbb{E}_X [\mathbb{E}[R|X, A = d(X)]] = \mathbb{E} \left[ \frac{R1(A = d(X))}{\pi(A|X)} \right],
\]

where the second line is based on Assumption 2.1, the third line is based on the SUTVA assumption in Section 2.1 and the last line relies on Assumption 2.2. Based on this value function, an optimal IDR \( d_0 \) is defined as

\[
d_0 \in \arg\max_d V(d).
\]

Note that

\[
V(d) = \mathbb{E}[\mathbb{E}[R|X, A = 1]|d(X) = 1) + \mathbb{E}[R|X, A = -1]|d(X) = -1)] = \mathbb{E}[(\mathbb{E}[R|X, A = 1) - \mathbb{E}[R|X, A = -1)]|d(X) = 1)] + \mathbb{E}[\mathbb{E}[R|X, A = -1]],
\]

and then as a result,

\[
d_0(X) \in \arg\max_{a \in A} \mathbb{E}[R|X, A = a],
\]

almost surely. It is observed that under the value function framework, the optimal IDR is to select the treatment with the largest expected outcome among all treatments for each patient.
Despite the progress of developing optimal IDR s in the intersection of statistics, econometrics and machine learning, only focusing on obtaining the largest expected outcome for each individual can be too restrictive and sometimes may not be even safe. For example, doctors may want to know whether a treatment does the best to improve the worst scenario, in particular for a high risk patient. Without such a risk consideration, this may lead to potentially severe events, such as exacerbation or hospitalization in practice. Similar concerns may happen in the credit card company, where the “best” policy should not only improve the average profit for the company, but also reduce the chance of incurring a heavy loss. This motivates us to control risk exposure associated with the corresponding decision rules, in addition to maximizing the expected outcome of each individual.

2.3 Conditional Value at Risk

It is natural to consider some robust metrics such as quantiles of $R$ given $X$ and $A$ to measure the effect of a treatment ([Wang et al. (2018)]). The corresponding optimal IDR $\tilde{d}$ under the quantile can be defined as

$$\tilde{d} \in \arg\max_{d} Q_\gamma(R(d)),$$

where $Q_\gamma(R(d)) = \inf\{\alpha : P[R(d) < \alpha] \geq 1 - \gamma\}$ and $\gamma \in (0, 1)$, i.e., $\gamma$-quantile of $R(d)$. Analogous to Problems (1) and (2), an equivalent form of Problem (6) can be stated as follows:

$$\max_{t \in \mathbb{R}, d} \quad t$$

subject to

$$t \leq Q_\gamma(R(d)).$$

Based on the above representation, roughly speaking, the constraint set $\{t \leq Q_\gamma(R(d))\}$ implies that $(1 - \gamma) \times 100\%$ of the population under the given IDR are controlled to be greater than a certain threshold $t$. Thus under the quantile criterion, one can obtain a robust IDR that can improve almost $(1 - \gamma) \times 100\%$ of the population in some extent.

There are several potential drawbacks of using quantile in IDR problems. First of all, using the quantile criterion treats all the outcomes lower than $Q_\gamma(R(d))$ as the same. However, as we pointed out in one of our early examples, the CD4 T cell $\{R_i(d) = 400\}$ below the normal level is considered to be much better than $\{R_i(d) = 50\}$, therefore they should be treated differently in practice. Secondly, $Q_\gamma(R(d))$ is generally not continuous in $\gamma$, which may cause instability. For instance, if the outcome distribution is discrete and a small change happens in $\gamma$, the resulting optimal $\tilde{d}$ may change significantly. Lastly, from the computational perspective, the quantile makes the optimization Problem (6) hard to solve while [Nouiehed et al. (2017)] recently has proved that quantile of random functional can be represented as a difference-of-convex function.

In order to address the drawbacks of using quantiles, also known as Value-at-Risk, we propose to use conditional value at risk (CVaR), also known as expected tail loss, which was proposed by [Artzner et al. (1999)] in risk management. Consider a continuous random variable $Y$. Then the original $\gamma$-CVaR of $Y$ is defined as

$$S(F_Y) := \frac{1}{\gamma} E[Y I(Y \leq Q_\gamma(F_Y))],$$
where $F_Y$ is the corresponding probability distribution of $Y$. Based on this definition, the $\gamma$-CVaR can be interpreted as a truncated mean lower than the $\gamma$-quantile of $Y$. For the general setting, instead of assuming a continuous distribution of $Y$, $\gamma$-CVaR is defined as an optimal value of a concave maximization problem by the celebrated work of [Rockafellar and Uryasev (2000)], which is defined as follows:

\[
S(F_Y) := \sup_{\alpha \in \mathbb{R}} \left\{ \alpha - \frac{1}{\gamma} \mathbb{E}[\alpha - Y]_+ \right\}, \tag{9}
\]

where $[t]_+ = \max(0, t)$. The leftmost of the optimal solution set to (9) is $Q_\gamma(F_Y)$ [Rockafellar and Uryasev (2000, Theorem 1)]. Then one can see that the definition in (9) is equivalent to (8) when the outcome distribution of $Y$ is continuous. As a remark, we note that $Y$ is often referred to as a loss in the finance literature. However, here we call $Y$ an outcome in order to be consistent with our problem setting.

We would like to point out that $\gamma$-CVaR has several nice properties discussed by [Artzner et al. (1999)] and it is in general preferable to quantile measure ([Sarykalin et al. (2008))]. In particular, based on the interpretation of (8), $\gamma$-CVaR considers average outcomes lower than $\gamma$-quantile, which treats lower tails of outcome differently. This exactly satisfies our purpose. The concave maximization formulation in (9) demonstrates the continuity of $\gamma$-CVaR with respect to $\gamma$, which provides more stability measure compared with the quantile criterion. Furthermore, $\gamma$-CVaR is considered to be more computationally efficient than the quantile criterion because of the concave maximization formulation. In addition, [Pflug (2000)] or [Rockafellar and Uryasev (2002)] showed that $S(F_Y) \leq Q_\gamma(F_Y)$, which is more conservative than $\gamma$-quantile. This implies that larger $\gamma$-CVaR of a random outcome indicates larger corresponding $\gamma$-quantile. Clearly the reverse inequality does not necessarily hold. These nice properties motivate us to use it in the IDR problems. For the related theoretical discussion about CVaR, we refer to [Rockafellar and Uryasev (2002)] and the references therein.

### 2.4 A New Robust Criterion for IDR Problems

In this subsection, we borrow the concept of CVaR in order to conduct risk control to obtain a robust optimal IDR that can improve the lower tails of outcomes. Specifically, we define the following decision-rule based CVaR criterion as

\[
M_0(d) := \sup_{\alpha \in \mathbb{R}} \left\{ \alpha - \frac{1}{\gamma} \mathbb{E}[\alpha - R(d)]_+ \right\}. \tag{10}
\]

Note that the difference between the criterion (10) and the original CVaR is to let the outcome depend on IDR $d$. Given the coherence property of CVaR in [Artzner et al. (1999)], we have the corresponding properties of $M_0(d)$.

**Proposition 2.1.** The following properties of $M_0(d)$ hold.

(a) If $R$ is shifted by a constant $c$, then $M_0(d)$ is also shifted by the same constant $c$;

(b) If $R$ is multiplied by a positive constant $c$, then the corresponding $M_0(d)$ is also multiplied by the same constant $c$;
(b) Given two IDR\( s \) \( d_1 \) and \( d_2 \), if \( R(d_1) \leq R(d_2) \), a.s., then
\[
M_0(d_1) \leq M_0(d_2);
\] (11)

(d) Given IDR \( d \), \( M_0(d) \leq \min\{V(d), Q_\gamma(R(d))\} \).

In addition, if outcome \( R = c \), a.s., then \( M_0(d) = c \).

Remark 2.1. The above Proposition \( 2.1 \) justifies the use of \( M_0(d) \). In particular, (a) and (b) demonstrate that \( M_0(d) \) is not affected by a constant shift or multiplication. (c) implies that if one IDR is strictly better than the other, \( M_0(d) \) has the same preference. The last property (d) indicates that \( M_0(d) \) is more conservative than the expected outcome and the quantile criterion when evaluating IDR \( d \).

If the distribution of \( R(d) \) is continuous, then we could rewrite \( 10 \) as
\[
M_0(d) = \frac{\mathbb{E}[\mathbb{I}(R \leq Q_\gamma(R(d)))]}{\gamma}.
\] (12)

Note that
\[
P(R(d) < \alpha) = \mathbb{E}\left[\frac{\mathbb{I}(A = d(X))}{\pi(A|X)}\mathbb{I}(R < \alpha)\right]
= \mathbb{E}\left[\sum_{a \in A} \mathbb{I}(d(X) = a)P(R < \alpha|X, A = a)\right]
= \mathbb{E}[P(R < \alpha|X, A = d(X))].
\] (13)

Then \( Q_\gamma(R(d)) \) can be further expressed as
\[
Q_\gamma(R(d)) = \inf \{\alpha \mid \mathbb{E}[P(R < \alpha|X, A = d(X))] \geq \gamma\},
\] (14)

which can be interpreted as the average \( \gamma \)-quantile of \( R \) under the decision rule \( d \). Correspondingly \( M_0(d) \) can be understood as the \( \gamma \)-average CVaR.

According to Proposition \( 2.1(d) \), \( M_0(d) \) can be regarded as a lower bound of \( V(d) \) and \( Q_\gamma(R(d)) \). Thus maximizing \( M_0(d) \) can potentially maximize both \( V(d) \) and \( Q_\gamma(R(d)) \). Then the optimal IDR under our proposed robust criterion \( M_0(d) \) is defined as
\[
d_0 \in \text{argmax}_d M_0(d).
\] (15)

The interpretation of the optimal IDR with respect to \( M_0(d) \) is to select a treatment/decision with the largest \( \gamma \)-average CVaR. Moreover, if we again use hypographical representation as in \( 2 \) and \( 7 \), then we could formulate it as the following constraint optimization problem:
\[
\max_{d, t \in \mathbb{R}} \quad t \\
\text{subject to} \quad \sup_{\alpha \in \mathbb{R}} \left\{ \alpha - \frac{1}{\gamma} \mathbb{E}[\max\{0, \alpha - R(d)\}] \right\} \geq t.
\] (16)
Based on the formulation (16), \(d_0\) can be interpreted as the best IDR with the average lower tail of the outcome being larger than a certain threshold. Besides, based on statement (e) in Proposition (2.1), the constraint set in (16) implies \(\min\{V(d), Q_\gamma(R(d))\} \geq t\). Therefore, by using \(M_0(d)\), the resulting optimal \(d_0\) can control both \((1-\gamma) \times 100\%\) of the population’s outcomes and the expected outcome greater than some threshold \(t\).

By a similar derivation as in (3), we have the following proposition.

**Proposition 2.2.** \(M_0(d) = \sup_{\alpha \in \mathbb{R}} \left\{ \alpha - \frac{1}{\gamma} E^d\left[\frac{(A - d(X))}{\pi(A|X)}(\alpha - R)\right] \right\}\).

The definition of (10) and Proposition (2.2) give us a way to compute the optimal IDR \(d_0\) and \(\alpha_0\) jointly via optimizing

\[
(d_0, \alpha_0) \in \arg\max_{\alpha \in \mathbb{R}, d} \left\{ \alpha - \frac{1}{\gamma} E^d[(\alpha - R)_+] \right\}.
\]  

(17)

### 2.5 Duality Representation

Note that \(M_0(d)\) involves a concave maximization with respect to \(\alpha\). Thus it would be useful to investigate its dual representation by making use of convex duality theory in Rockafellar (1974).

To begin with, we first define the following set:

\[
\mathcal{V}_0^d := \{V \in L^1(\mathcal{T}, \mathcal{F}_1, P^d) \mid E^d[V] = 1, \ 0 \leq V(\omega) \leq \frac{1}{\gamma}, \ \text{for almost sure all } \omega \in \mathcal{T} \}.
\]  

(18)

We have the following theorem that gives the dual representation of \(M_0(d)\).

**Theorem 2.1.** \(M_0(d) = \inf_{V \in \mathcal{V}_0^d} E^d[V R]\).

According to the duality representation of \(M_0(d)\) and its proof in the supplementary material, we can define a conditional probability measure \(P^V(B) = \int_B V dP^d\) for any measurable set \(B \in \mathcal{T}\), where \(V \in \mathcal{V}_1^d\). Then \(V = \frac{dP^V}{dP^d}\). Define

\[
\zeta(u) = \begin{cases} 
0 & \text{if } 0 \leq u \leq \frac{1}{\gamma} \\
\infty & \text{otherwise}
\end{cases}
\]

Then we can further rewrite \(E^d[RV]\) in \(M_0(d)\) for \(V \in \mathcal{V}_0^d\) as

\[
E^d[RV] = E^d[R \frac{dP^V}{dP^d}] + E \left[ \zeta \left( \frac{dP^V}{dP^d} \right) \right]
\]

\[
= E_X [E_{P^V}[R]] + E_X \left[ \int \zeta \left( \frac{dP^V}{dP^d} \right) dP^d \right]
\]

\[
= E_{P^V}[R] + I_\zeta \left( \frac{dP^V}{dP^d} \right),
\]

where \(I_\zeta(\cdot)\) can be interpreted as the \(f\)-divergence distance between \(P^V\) and \(P^d\). Then \(M_0(d) = \inf_{P^V \ll P^d} E_{P^V}[R] + I_\zeta \left( \frac{dP^V}{dP^d} \right)\), where \(u \ll v\) means that the probability measure \(u\) is absolutely...
continuous with respect to the probability measure \( v \). Thus the optimal IDR can also be written as
\[
d_0 \in \arg\max_d \left\{ \inf_{P^V \ll P^d} \mathbb{E}_{P^V}[R] + I\left(\frac{dP^V}{dP^d}\right) \right\},
\]
which can be interpreted as choosing an optimal decision rule in terms of the worst expected outcome within the \( f \)-divergence distance from the original distribution \( P^d \).

According to our problem setting, define the density under \( P^d \) is \( f_0(x)I(d(x) = a)f_1(r|x,a) \). Since \( P^V \ll P^d \), then the density under \( P^V \) should be \( v_0(x)I(d(x) = a)v_1(r|x,a) \) for some conditional probability density \( v_0(x) \) and \( v_1(r|x,a) \). Then we can have \( \frac{v_0(x)v_1(r|x,a=d(x))}{f_0(x)f_1(r|x,a=d(x))} \) by the chain rule. Therefore, we can further rewrite \( M_0(d) \) as
\[
M_0(d) = \inf_{P^V} \left\{ \mathbb{E}_{P^V}[R] \mid P^V \ll P^d, \ 0 \leq \frac{v_0(x)v_1(r|x,a = d(x))}{f_0(x)f_1(r|x,a = d(x))} \leq \frac{1}{\gamma}, \text{ almost surely} \right\}. \tag{19}
\]
This gives us a natural link to distributionally robust statistical models that can evaluate a decision rule under ambiguity. Maximizing \( M_0(d) \) over the decision rule \( d \) is equivalent to identifying an optimal IDR that is robust to the contamination of both outcome \( R \) and covariates \( X \) characterized by a probability constraint set.

3 Statistical Estimation and Optimization

In this section, we discuss the estimation and optimization procedures for Problem (2.2) given observed data. The optimization in (17) can be rewritten as
\[
\max_{\alpha \in \mathbb{R}, d \in D} \mathbb{E} \left[ \left( \alpha - \frac{(\alpha - R)_+}{\gamma} \right) I(A = d(X)) \right], \tag{20}
\]
where \( D \) is some specified classes of decision rules such as the linear ones.

Consider the binary treatment setting and let \( d(X) = \text{sign}(f(X)) \). Suppose we observe independently and identically distributed data \((X_i, A_i, R_i); i = 1, \cdots, n\). Then we can estimate the optimal IDR via empirical approximation:
\[
\max_{\alpha \in \mathbb{R}, d \in D} \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = \text{sign}(f(X_i)))}{\pi(A_i|X_i)} \left( \alpha - \frac{(\alpha - R_i)_+}{\gamma} \right). \tag{21}
\]
It is well known that optimization over indicator functions is NP hard. Alternatively, we can replace the 0-1 loss function by the following smooth truncated loss,
\[
S(u) = \begin{cases} 
0 & \text{if } u \leq -\delta \\
(1 + u/\delta)^2 & \text{if } 0 > u \geq -\delta \\
2 - (1 - u/\delta)^2 & \text{if } \delta > u \geq 0 \\
2 & \text{if } u > \delta,
\end{cases}
\]
and then use a functional margin representation to express \( I(A_i = \text{sign}(f(X_i))) \) as \( I(A_i f(X_i) > 0) \) for each \( i \). The corresponding function plot of \( S(u) \) is shown in Figure 2 with \( \delta = 1 \). From the plot, we can see that the smooth approximation \( \frac{S(u)}{2} \) is very close to the 0-1 loss. The parameter \( \delta \) can control the closeness of this approximation. In practice, we can simply choose \( \delta = 1 \).

![Figure 2: Plot of smooth surrogate loss function with \( \delta = 1 \)](image)

By using the surrogate function, we can estimate the optimal IDR under \( M_0(d) \) via computing

\[
\min_{\alpha \in \mathbb{R}, f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \frac{S(A_i f(X_i))}{\pi(A_i | X_i)} (\alpha - \frac{(\alpha - R_i)_+}{\gamma}) + \lambda \frac{n}{2} \| f \|_{\mathcal{H}}^2,
\]

(22)

where \( \| f \|_{\mathcal{H}}^2 \) is some convex penalty function on \( f \). For example, if we consider \( \mathcal{H} \) be a Reproducing Kernel Hilbert Space (RKHS), then \( \| f \|_{\mathcal{H}}^2 \) could be the square of RKHS norm of \( f \). The estimated IDR is given by \( \hat{d}_0(X) = \text{sign}(\hat{f}(X)) \). Note that Problem (22) involves a non-convex and potentially non-smooth optimization problem. Recent developments in difference-of-convex (DC) optimization (Pang et al. (2016)) motivate us to use DC programming to efficiently solve this problem. Note that \( S(u) \) can be expressed as a difference of convex differentiable functions: \( S_1(u) - S_2(u) \) where

\[
S_1(u) = \begin{cases} 
0 & \text{if } u \leq -\delta \\
(1 + u/\delta)^2 & \text{if } -\delta < u \leq 0, \\
2 + 2u/\delta - 1 & \text{if } 0 < u 
\end{cases}
\]

and

\[
S_2(u) = \begin{cases} 
0 & \text{if } u \leq 0 \\
(u/\delta)^2 & \text{if } 0 < u \leq \delta, \\
2u/\delta - 1 & \text{if } u > \delta 
\end{cases}
\]

Define

\[
G^{(1)}(f, \alpha) := \frac{1}{n} \sum_{i=1}^{n} \frac{S(A_i f(X_i))}{\pi(A_i | X_i)} (- (\alpha - \frac{(\alpha - R_i)_+)}{\gamma})) + \lambda \frac{n}{2} \| f \|_{\mathcal{H}}^2,
\]

(23)
Thus the optimal solution set 

\[
\tilde{G}_j^{(1)}(f) = \frac{1}{n} \sum_{i=1}^{n} S(A_i f(X_i)) \left(-R_j - \frac{(R_j - R_i)_+}{\gamma}\right).
\]  

(24)

The following proposition gives us a way to express (22) as a DC function.

**Proposition 3.1.** The following two optimization problems have the same optimal value, i.e.,

\[
\min_{\alpha \in \mathbb{R}, f \in \mathcal{H}} G^{(1)}(f, \alpha) = \min_{f \in \mathcal{H}} \left\{ \tilde{G}^{(1)}(f) \right\},
\]

(25)

where \( \tilde{G}^{(1)}(f) := \min_{1 \leq j \leq n} \{ G_j^{(1)}(f) \} + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}} \). More importantly, the optimal solution sets of \( f \) to both problems are the same.

**Proof.** Note that for any given \( f \), \( G^{(1)}(f, \alpha) \) is a convex piecewise affine function with respect to \( \alpha \). Thus the optimal solution set \( \alpha^* \) should contain one of the knots, i.e., \( R_1, \ldots, R_n \). Then it follows that

\[
\min_{\alpha \in \mathbb{R}} G^{(1)}(f, \alpha) = \min_{j \in \{1, \ldots, n\}} G_j^{(1)}(f) + \frac{\lambda_n}{2} \| f \|^2_{\mathcal{H}}.
\]

Thus

\[
\min_{f \in \mathcal{H}, \alpha \in \mathbb{R}} G^{(1)}(f, \alpha) = \min_{f \in \mathcal{H}} \left\{ \min_{1 \leq j \leq n} \{ G_j^{(1)}(f) \} + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}} \right\},
\]

and correspondingly

\[
\arg\min_{f \in \mathcal{H}} G_1(f, \alpha) = \arg\min_{f \in \mathcal{H}} \left\{ \min_{1 \leq j \leq n} \{ G_j(f) \} + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}} \right\}.
\]

\[\square\]

Based on Proposition 3.1, instead of solving (22), we can equivalently solve the optimization problem in the right hand side of (25). Let \( c_{ij} = \frac{-(R_j - (R_j - R_i)_+)}{\pi(A_i \cdot X_i)} \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, n \), and note that \( c_{ij} \) is not necessarily nonnegative. Recall that \( S(u) = S_1(u) - S_2(u) \). Then we can further rewrite \( G_j^{(1)}(f) \) as

\[
G_j^{(1)}(f) = \frac{1}{n} \sum_{i=1}^{n} (\max(c_{ij}, 0) S_1(A_i f(X_i)) + \max(-c_{ij}, 0) S_2(A_i f(X_i)))
\]

\[= \frac{1}{n} \sum_{i=1}^{n} (\max(c_{ij}, 0) S_2(A_i f(X_i)) + \max(-c_{ij}, 0) S_1(A_i f(X_i))) \]

\[= F_j(f) - H_j(f), \]

where both \( F_j(f) \) and \( H_j(f) \) are convex functions with respect to \( f \) for \( j = 1, \ldots, n \). Then we can further decompose

\[
\tilde{G}^{(1)}(f) = \min_{1 \leq j \leq n} \{ F_j(f) - H_j(f) \} + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}}
\]

\[= \sum_{i=1}^{n} F_j(f) - \max_{1 \leq j \leq n} \{ H_j(f) + \sum_{k \neq j}^{n} F_k(f) \} + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}} \]

\[= F(f) - \max_{1 \leq j \leq n} h_j(f) + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}}, \]

(27)
as a DC function, where $h_j(f) := H_j(f) + \sum_{k \neq j}^n F_k(f)$. Note that $\tilde{G}(1)(f)$ is a potentially non-smooth function if there exits multiple $k$’s such that $h_k(f) = \max_{1 \leq j \leq n} h_j(f)$. As pointed by Pang et al. (2016), traditional DC programming may potentially converge to nonsense points. Motivated by Pang et al. (2016), we define $M_\epsilon(f) := \{ j \mid h_j(f) \geq \max_{1 \leq k \leq n} h_k(f) - \epsilon \}$, i.e., “$\epsilon$-argmax” index set and use the following enhanced probabilistic DCA summarized in Table 1 below.

Algorithm 1 Algorithm for (22)

1: Given a fixed $\epsilon > 0$, let $f^{(v)}$ be the solution at the $v$ iteration.
2: Randomly select $j \in M_\epsilon(f^{(v)})$, and compute $f^{(v+1)} \in \arg\min_{f \in H} \{ F(f) - \frac{\partial h_j(f^{(v)})}{\partial f} (f - f^{(v)}) + \lambda \frac{\|f\|_2^2}{2} \}$. (28)
3: The algorithm stops when $|\tilde{G}(1)(f^{(v)}) - \tilde{G}(1)(f^{(v+1)})| < \kappa$, for some pre-specified positive constant $\kappa$.

The proof of convergence to sharp stationary points by the above algorithm can be found in Pang et al. (2016). For the computation of the subproblem (28), efficient algorithms such as quasi-newton methods can be used.

4 Analysis of Statistical Convergence

In this section, we discuss the statistical theory related to our estimation under $M_0(d)$. We first define $M_0(d, \alpha) = \mathbb{E} \left[ \mathbb{I}(A = d(X)) \frac{(\alpha - \frac{1}{\gamma} (\alpha - R)_+)}{\pi(A|X)} \right]$, (29)
and denote the optimal solutions of maximizing (29) as $d_0 = \text{sign}(f_0)$ and $\alpha_0$. Since we use the surrogate loss function $S(u)$, we further define $M_S(f, \alpha) = \mathbb{E} \left[ \frac{S(Af(X))}{\pi(A|X)} (\alpha - \frac{1}{\gamma} (\alpha - R)_+) \right]$ as the surrogate value function. Our theoretical results are based on regret analysis similar to Zhou et al. (2017).

4.1 Fisher Consistency

We first establish Fisher consistency of estimating optimal ITRs under $M_S(f, \alpha)$ to justify the use of the surrogate loss $S(u)$, compared with $M_0(d, \alpha)$. The proof is different from the classical Fisher consistency in classification, which only involves one functional class of interest. Here we need to consider the effect of the surrogate function on estimating $\alpha$.

Theorem 4.1. For any measurable function $f$ and $\alpha$, if $(f^*_S, \alpha^*_S)$ maximizes $M_S(f, \alpha)$, then $(\text{sign}(f^*_S), \alpha^*_S)$ maximizes $M_0(d, \alpha)$.

Based on Theorem 4.1 instead of $M_0(d, \alpha)$, we can target on $M_S(d, \alpha)$ equivalently.
4.2 Excess Value Bound

Based on Theorem 4.1, we can further justify the use of the surrogate function \( S(u) \) by establishing the following excess value bound for the 0-1 loss in \( M_0(d, \alpha) \).

**Theorem 4.2.** For any measurable function \( f, \alpha \) and any probability distribution over \( (X, A, R) \),

\[
M_0(d_0, \alpha_0) - M_0(\text{sign}(f), \alpha) \leq M_S(f^*_S, \alpha^*_S) - M_S(f, \alpha).
\]

Theorem 4.2 gives us a way of bounding the difference between the optimal IDR and the estimated IDR under \( M_0(d, \alpha) \) by using \( M_S(d, \alpha) \) instead.

4.3 Convergence Rate

In order to obtain the finite sample performance of our estimated optimal IDR under \( M_0(d, \alpha) \), it is enough to focus on the difference of \( M_S(d, \alpha) \) between the estimated optimal IDR and the optimal ITR based on Theorem 4.2. Define

\[
(\hat{f}, \hat{\alpha}) = \arg\min_{f \in \mathcal{H}, \alpha \in \mathbb{R}} O_n(f, \alpha) + \frac{\lambda_n}{2} ||f||_H^2,
\]

where \( O_n(f, \alpha) := \frac{1}{n} \sum_{i=1}^{n} S(A_i f(X_i))(\frac{\alpha - R_i}{\gamma} - \alpha) \). For simplicity, in the following we consider \( \mathcal{H} \) be RKHS with Gaussian radial basis functions. The results can be extended to other scenarios such as the class of linear functions with the \( l_1 \) penalty. Define \( O_S(f, \alpha) = -M_S(f, \alpha) \) and let \( (f_{\lambda_n}, \alpha_{\lambda_n}) = \arg\min_{f \in \mathcal{H}, \alpha \in \mathbb{R}} O_S(f, \alpha) + \frac{\lambda_n}{2} ||f||_H^2 \). Then \( A(\lambda_n) := O_S(f_{\lambda_n}, \alpha_{\lambda_n}) + \frac{\lambda_n}{2} ||f_{\lambda_n}||_H^2 - O_S(f^*_S, \alpha^*_S) \) is considered to be the approximation error. The following theorem gives us a finite sample upper bound of our estimated optimal IDR and the optimal IDR based on \( M_0(d, \alpha) \).

**Theorem 4.3.** For any distribution \( P \) over \( (X, A, R) \) such that \( |R| \leq C_0 \), if \( A(\lambda_n) \leq C_1 \lambda_n^{w_1} \), where \( w_1 \in (0, 1] \), then with probability at least \( 1 - \epsilon \),

\[
M_0(d_0, \alpha_0) - M_0(\text{sign}(\hat{f}), \hat{\alpha}) \leq C_2 n^{\frac{w_1}{2w_1 + 1}},
\]

for some constant \( C_2 \).

The above theorem shows that the difference between our estimated IDR and the optimal IDR under \( M_0(d, \alpha) \) converges to 0 in probability under some conditions. The upper bound assumption on the approximation error \( A(\lambda_n) \) is analogous to those in the statistical learning literature such as Steinwart and Scovel (2007) to derive the convergence rate. The convergence rate is the same as those in Zhao et al. (2012) and Zhou et al. (2017). This is not surprising because the proposed \( M_0(d) \) can be roughly regarded as the truncated mean.

5 Numerical Studies

5.1 Simulation Studies

For all simulation settings, we consider binary-armed randomized trials with equal probabilities of patients being assigned to each treatment group. We use \( l_1\text{-DC-CVaR}, l_2\text{-DC-CVaR} \) and \( \text{GK-DC-CVaR} \) to represent the methods of estimating optimal IDRs under \( M_0(d) \) with three different
penalties on $f$ in Problem (22) respectively. Here “$l_1$” and “$l_2$” refer to the $l_1$ and $l_2$ penalties. “GK” represents using Gaussian radial basis functions with bandwidth $\varsigma$ to learn the optimal IDR. In all our numerical studies, we set $\gamma = 0.5$.

All tuning parameters are selected based on the 10-fold-cross-validation procedure. We select the tuning parameter that maximizes the empirical average of $M_0(d)$ on the validation data set. We compare our methods with the following three methods:

1. $l_1$-PLS by Qian and Murphy (2011) with basis function $(1, X, A, XA)$;
2. RWL by Zhou et al. (2017) with linear kernel;
3. RWL by Zhou et al. (2017) with Gaussian kernel.

Since we only consider the randomized design study, we do not compare those methods designed for observational studies such as Zhang et al. (2012).

### 5.1.1 A Motivating Example Revisit

Recall the motivating example in Figure 1 that shows the importance of risk controls in estimating optimal IDRs. In this subsection, we conduct some numerical analysis to further demonstrate this finding. In particular, the covariate of gender is generated by uniform distribution over $\{1, -1\}$, where 1 and $-1$ denotes male and female respectively. The corresponding outcome $R$ is generated by the following model:

$$R = I(XA = 1)\epsilon_1 + I(XA = -1)\epsilon_2,$$

where $\epsilon_1 \sim \mathcal{N}(-0.1, 1)$ and $\epsilon_2 \sim \mathcal{N}(0, 0.5)$. We consider training data with the sample size $n = 200$ and independently generated test data of size 10000. Based on test data, in Figure 3, we plot box plots of three different outcome distributions if treatments follow estimated IDRs by $l_1$-PLS, linear RWL, and $l_2$-DC-CVaR respectively. Based on these box plots, we can observe that since there is not much difference between these two treatments based on the expected outcome, the empirical mean of value functions resulted from these three methods are indistinguishable. However, besides maximizing the expected outcome for each individual, our methods also control the tails of subjects. The resulting outcome distribution by our method is more stable, and thus has less variability than those of $l_1$-PLS and linear RWL.
Figure 3: Box plots of value functions computed by three methods and two reference distributions. The left box plot corresponds to the result of $l_1$-PLS under the value function framework. The second box plot corresponds to those of RWL. The middle two box plots correspond to the result of our proposed method. The last two box plots correspond to two reference normal distributions.

5.1.2 Distribution Shift Examples

In this section, we demonstrate the superior performance of our methods under distribution shift of both covariates $X$ and outcome $R$ based on the duality representations of $M_0(d)$ in (19). We consider the sample size $n = 200$ and the dimension $p = 20$. The outcome $R$ is generated by the model: $R = 1 + x_1 + x_2 + A(x_1 - x_2 + x_3) + \epsilon$. We consider the following two distribution shift scenarios:

(1) Each covariate follows a two component Gaussian mixture distribution of $\mathcal{N}(0, 1)$ and asymmetric log-normal distribution $\text{lognorm}(0, 2)$ with probabilities of the two mixture component to be 0.7 and 0.3 respectively and $\epsilon$ follows the standard Gaussian distribution;

(2) Covariates $X$ are generated by the uniform distribution between $-1$ and $1$ and $\epsilon$ follows a two-component mixture distribution of $\mathcal{N}(0, 1)$ and asymmetric log-normal distribution $\text{lognorm}(0, 2)$ with probabilities of the two mixture component to be 0.7 and 0.3 respectively.

The first scenario considers the covariate distribution shift and the second scenario considers the outcome distribution shift. For simplicity, we only report misclassification error rates given by $l_1$-PLS, linear RWL and $l_2$-DC-CVaR in Table 1. For Scenario (1), since $l_1$-PLS assumes a linear model, its performance is not affected by the distributional shift of covariates. In contrast, RWL, which is based on maximizing the value function, depends heavily on correct approximations to the value function empirically. Thus the performance of RWL is worse than $l_1$-PLS under this
scenario. For the estimated optimal IDR under our proposed $M_0(d)$, the performance is superior to RWL because $M_0(d)$ considers the perturbation of the covariate distributional shift. For Scenario (2), since the estimated optimal IDR under $M_0(d)$ is a minimax estimator under the outcome distributional shift, the performances are much better than the other two methods developed under the value function framework.

Table 1: Comparisons of misclassification error rates (standard errors) for simulated examples with $n = 200$ and $p = 20$.

| Scenario (1) | Scenario (2) |
|--------------|--------------|
| $l_1$-PLS    | 0.07 (0.01)  |
| $l_2$-RWL    | 0.14 (0.006) |
| $l_2$-DC-CVaR| 0.08 (0.01)  |

5.1.3 Additional Simulation Scenarios

In this subsection, we further study the performance of our proposed methods via eight simulation examples. In order to have more flexibility, we consider to use $M_1(d)$ defined below by combining our proposed $M_0(d)$ and $V(d)$ together:

$$M_1(d) := 0.5V(d) + 0.5M_0(d).$$

The corresponding optimal IDR is denoted by $d_1$. All our results in Sections 3 and 4 can be naturally extended. In particular, since the surrogate function we use in Section 3 is a DC function, a small modification can be incorporated in Algorithm 1 to estimate $d_1$ empirically. Moreover, by making use of the existing results in the literature of expected value function such as Zhou et al. (2017), we can perform the similar regret analysis for $M_1(d)$ as that of $M_0(d)$ in Section 4. We consider the sample size $n = 200$ and the dimension $p = 20$. The covariates $X$ are generated by the uniform distribution between $-1$ and $1$. The outcome $R$ is generated by the model: $R = 1 + x_1 + x_2 + A\delta(X) + \epsilon$. We consider the following eight different combinations of $\delta(X)$ and $\epsilon$:

1. $\delta(X) = x_1 - x_2 + x_3$, and $\epsilon$ follows Gaussian normal $\mathcal{N}(0, 4)$;
2. $\delta(X) = x_1 - x_2 + x_3$, and $\log(\epsilon)$ follows Gaussian normal $\mathcal{N}(0, 2|1 + x_1 + x_2|)$;
3. $\delta(X) = x_1 - x_2 + x_3$, and $\log(\epsilon)$ follows Gaussian normal $\mathcal{N}(0, 2)$;
4. $\delta(X) = x_1 - x_2 + x_3$, and $\epsilon$ follows a Weibull distribution with shape parameter 0.3 and scale parameter 0.5;
5. $\delta(X) = 3.8(0.8 - x_1^2 - x_2^2)$, and $\epsilon$ follows Gaussian normal $\mathcal{N}(0, 4)$;
6. $\delta(X) = 3.8(0.8 - x_1^2 - x_2^2)$, and $\log(\epsilon)$ follows Gaussian normal $\mathcal{N}(0, 2|1 + x_1 + x_2|)$;
7. $\delta(X) = 3.8(0.8 - x_1^2 - x_2^2)$, and $\log(\epsilon)$ follows Gaussian normal $\mathcal{N}(0, 2)$;

18
\( \delta(\mathbf{X}) = 3.8(0.8 - x_1^2 - x_2^2) \), and \( \epsilon \) follows a Weibull distribution with shape parameter 0.3 and scale parameter 0.5.

We consider different shapes of error distributions to test the robustness of our methods, compared with other methods. The first four scenarios are of linear decision boundaries while the remaining four consider nonlinear decision boundaries. In order to evaluate different methods, we generate test data and use \( \text{sign}(\delta(\mathbf{X})) \) as the true optimal decision rule, since treatment \( A \) only appears in the interaction term \( \delta(\mathbf{X}) \). We evaluate different methods based on the misclassification error rates in Table 2, mean of value functions in Table 3, mean of 50% and 25% quantiles of value functions in Tables 4 and 5 respectively. Overall, our methods show competitive performances among all methods. In particular, for Scenarios (1) and (5), which are standard simulation settings in the literature, our proposed methods perform well in finding optimal IDRs. For Scenarios (2) and (6), the error distributions depend on the covariate information. Although the averages of empirical value functions of our proposed methods are smaller than those of RWL, the 50% and 25% quantiles of empirical value functions by our methods are much better. One possible reason is that methods under the value function framework ignore subjects with potentially high risks while only focusing on maximizing the value function. Thus the resulting IDRs by these methods may assign wrong treatments to patients and make them become even worse by delivering the corresponding IDR. Similar observations can be drawn from other simulation scenarios.

Table 2: Comparisons of misclassification error rates (standard deviations) for simulated examples with \( n = 200 \) and \( p = 20 \). From left to right, each column represents Scenarios (1)-(8) respectively. Each row represents one specific method. The last six rows correspond to our proposed methods.

|       | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
|-------|------|------|------|------|------|------|------|------|
| l₁-PLS | 0.28 | 0.5  | 0.44 | 0.45 | 0.44 | 0.52 | 0.5  | 0.49 |
| RWL   | 0.34 | 0.5  | 0.44 | 0.44 | 0.41 | 0.51 | 0.47 | 0.46 |
| RWL-GK| 0.5  | 0.5  | 0.5  | 0.5  | 0.38 | 0.52 | 0.46 | 0.43 |
| l₂-DC-CVaR | 0.32 | 0.17 | 0.22 | 0.13 | 0.43 | 0.5  | 0.48 | 0.46 |
| l₁-DC-CVaR | 0.32 | 0.16 | 0.19 | 0.1  | 0.44 | 0.5  | 0.49 | 0.49 |
| GK-DC-CVaR | 0.49 | 0.5  | 0.5  | 0.5  | 0.38 | 0.51 | 0.41 | 0.38 |

Table 3: Comparisons of average value functions (standard deviations) for simulated examples with \( n = 200 \) and \( p = 20 \). From left to right, each column represents scenarios (1)-(8) respectively. Each row represents one specific method. The last six rows correspond to our proposed methods.

|       | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
|-------|------|------|------|------|------|------|------|------|
| l₁-PLS | 1.51 | 205839.37 | 8.55 | 5.67 | 1.25 | 32840.58 | 8.42 | 5.68 |
| RWL   | 1.39 | 47651.35 | 8.58 | 5.70 | 1.37 | 42489.93 | 8.46 | 5.83 |
| RWL-GK | 1.01 | 241196.97 | 8.41 | 5.55 | 1.47 | 30912.36 | 8.53 | 5.96 |
| l₂-DC-CVaR | 1.44 | 240599.77 | 8.96 | 6.27 | 1.26 | 32777.67 | 8.53 | 5.78 |
| l₁-DC-CVaR | 1.44 | 240226.39 | 8.94 | 6.32 | 1.27 | 32917.95 | 8.51 | 5.68 |
| GK-DC-CVaR | 1.02 | 13046.46 | 8.40 | 5.54 | 1.49 | 34718.08 | 8.75 | 6.11 |
Table 4: Comparisons of 50% quantiles (standard deviations) of value functions for simulated examples with $n = 200$ and $p = 20$. From left to right, each column represents scenarios (1)-(8) respectively. Each row represents one specific method. The last six rows correspond to our proposed methods.

|       | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $l_1$-PLS | 1.52(0.13) | 2.21(0.16) | 2.71(0.2) | 1.93(0.19) | 1.28(0.24) | 2.35(0.22) | 2.86(0.18) | 2.03(0.2) |
| RWL   | 1.4(0.17) | 2.22(0.19) | 2.71(0.16) | 1.96(0.16) | 1.39(0.25) | 2.36(0.27) | 2.96(0.36) | 2.2(0.44) |
| RWL-GK | 1.01(0.07) | 2.22(0.05) | 2.59(0.05) | 1.81(0.04) | 1.5(0.23) | 2.25(0.65) | 3.01(0.58) | 2.32(0.58) |
| $l_2$-DC-CVaR | 1.45(0.14) | 2.8(0.08) | 3.11(0.09) | 2.46(0.08) | 1.28(0.19) | 2.42(0.14) | 2.97(0.16) | 2.17(0.22) |
| $l_1$-DC-CVaR | 1.45(0.14) | 2.82(0.06) | 3.13(0.08) | 2.48(0.05) | 1.28(0.19) | 2.43(0.06) | 2.91(0.11) | 2.04(0.09) |
| GK-DC-CVaR | 1.01(0.08) | 2.22(0.04) | 2.59(0.05) | 1.81(0.04) | 1.53(0.16) | 2.33(0.67) | 3.24(0.42) | 2.56(0.29) |

Table 5: Comparisons of 25% quantiles (standard deviations) of value functions for simulated examples with $n = 200$ and $p = 20$. From left to right, each column represents scenarios (1)-(8) respectively. Each row represents one specific method. The last six rows correspond to our proposed methods.

|       | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $l_1$-PLS | -1.3(0.14) | 1.02(0.16) | 1.28(0.26) | 0.69(0.23) | -1.71(0.25) | 0.68(0.18) | 0.89(0.18) | 0.25(0.19) |
| RWL   | -1.43(0.19) | 1.04(0.19) | 1.29(0.21) | 0.72(0.2) | -1.59(0.27) | 0.71(0.22) | 1.05(0.41) | 0.51(0.45) |
| RWL-GK | -1.82(0.08) | 1.02(0.04) | 1.12(0.05) | 0.53(0.04) | -1.47(0.23) | 0.72(0.57) | 1.2(0.63) | 0.71(0.57) |
| $l_2$-DC-CVaR | -1.38(0.15) | 1.82(0.12) | 1.86(0.14) | 1.46(0.13) | -1.71(0.2) | 0.75(0.11) | 1.01(0.2) | 0.41(0.25) |
| $l_1$-DC-CVaR | -1.37(0.15) | 1.84(0.1) | 1.9(0.12) | 1.5(0.07) | -1.7(0.2) | 0.75(0.05) | 0.93(0.13) | 0.25(0.09) |
| GK-DC-CVaR | -1.82(0.08) | 1.03(0.03) | 1.12(0.05) | 0.53(0.04) | -1.45(0.17) | 0.79(0.6) | 1.46(0.46) | 0.95(0.28) |

5.2 Real Data Applications

In this section, we perform a real data analysis to further evaluate our proposed robust criterion for estimating optimal IDRs. We use the clinical trial dataset from “AIDS Clinical Trials Group (ACTG) 175” in Hammer et al. (1996) to study whether there exists some subpopulations that are suitable for different combinations of treatments for AIDS. In this study, a total number of 2139 patients with HIV infection is randomly assigned into four treatment groups: zidovudine (ZDV) monotherapy, ZDV combined with didanosine (ddI), ZDV combined with zalcitabine (ZAL), and ddI monotherapy with equal probability. In this data application, we focus on finding optimal IDRs between two treatments: ZDV with ddI and ZDV with ZAL as our interest. The total number of patients receiving these two treatments are 1046.

Similar to the previous studies by Lu et al. (2013) and Fan et al. (2017), we select 12 baseline covariates for our model: age (year), weight(kg), CD4 T cells amount at baseline, Karnofsky score (scale at 0-100), CD8 amount at baseline, gender (1 = male, 0 = female), homosexual activity (1 = yes, 0 = no), race (1 = non white, 0 = white), 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 the outcome as the early stage (around 25 weeks) CD4+ T (cells/mm³) cell amount. Using this outcome, we can estimate the optimal IDR under our proposed robust criterion. To evaluate the performance of our proposed methods under robust criterion, we randomly divide the dataset into five folds and use four of them to estimate optimal IDRs by different methods. The remaining one fold of data is used to evaluate the performances of different methods. We repeat this procedure 200 times. For each method, we report the mean, 25% and 10% quantiles of empirical value functions. From Table 6 we can see that our proposed methods perform best among all methods. Furthermore, our method is also the best in terms of means of the empirical value functions due to the heavy right tail distribution of $R$ as shown in Figure 4. This is consistent with our findings in the simulation settings when the outcome distribution has potential heavy right tails. In terms of 25% and 10% quantiles of empirical value functions, we can see the advantages of our method are more significant than the other two as we focus more on the tails of outcomes for each individual.

Table 6: Results of value function comparison for the AIDS data. First column represents the means of empirical value functions. Second and third columns represent means of 50% and 25% quantiles of empirical value functions, respectively.

|        | $V_n(d)$  | 25% quantiles | 10% quantiles |
|--------|-----------|----------------|---------------|
| $l_1$-PLS | 401.11(0.06) | 285.98(0.07) | 216.03(0.09) |
| rwl    | 400.77(0.06) | 285.44(0.07) | 215.66(0.08) |
| $l_2$-DC-CVaR | **401.71(0.06)** | **287.28(0.07)** | **217.01(0.09)** |

Figure 4: Histogram of the early stage CD4+ T cell amount. It exhibits a heavy right tail behavior.
6 Extensions

In this paper, we propose a robust criterion to estimate optimal IDRs by considering individualized risk using the concept of CVaR. The resulting optimal IDRs can not only improve the individualized expected outcome, but also prevent adverse consequences by controlling the lower tails of the outcome. In this section, we discuss several extensions of our proposed criterion on the perspective of both the modeling and the algorithm.

6.1 Control of Individualized Lower Tails

One interesting question is whether we can control the individualized $\gamma$-CVaR instead of average $\gamma$-CVaR of the outcome for each subject. In this subsection, we discuss another criterion as an extension of $M_0(d)$ in (10) from the average level risk control to the individualized level. Basically, we define the decision-based individualized CVaR function as follows:

\[
M_2(d) := \sup_{\alpha \in L^1(\mathcal{X},\Xi,P_X)} \left\{ \mathbb{E}[\alpha(X)] - \frac{1}{\gamma} \mathbb{E}[(\alpha(X) - R(d))_+] \right\},
\]

(32)

where $\Xi$ is the $\sigma$-field generated by $\mathcal{X}$ and $P_X$ is the corresponding probability measure. In order to understand $M_2(d)$, we need to characterize the optimal $\alpha^*$ in (32) given any IDR $d$. Define the individualized $\gamma$-quantile as $Q_\gamma(R|X = x, A = a) := \inf\{\alpha : P(R < \alpha(x,a)|X = x, A = a) \geq 1 - \gamma\}$ and individualized $\gamma$-CVaR as $\text{CVaR}_\gamma(R|X, a) := \sup_{s \in \mathbb{R}}\{s - \frac{1}{\gamma} \mathbb{E}[(s - R)_+|X, A = a]\}$ given covariates $X$ and $A = a$. The following theorem gives an explicit expression of the optimal $\alpha^*$ by using the theory of variational analysis (Rockafellar and Wets (2009)).

**Theorem 6.1.** Given any decision rule $d$, $\alpha^*$ is optimal to the optimization problem in $M_2(d)$ if and only if

\[
\alpha^*(X) = Q_\gamma(R|X, A = d(X)) \quad \text{almost surely.}
\]

Thus

\[
M_2(d) = \mathbb{E}[\text{CVaR}_\gamma(R|X, A = d(X))].
\]

(34)

According to (33), the explicit form of $\alpha^*(X)$ in $M_2(d)$ can be interpreted as the individual $\gamma$-quantile by the decision rule $d$. Then $M_2(d)$ in (34) takes each individualized CVaR into consideration. The following proposition explains the relationships of $M_2(d)$ to $M_0(d)$ and $V(d)$.

**Proposition 6.1.** The following two inequalities hold: $M_0(d) \leq M_2(d) \leq V(d)$.

**Proof.** The first inequality follows the fact that any constant is an element of $L^1(\mathcal{X},\Xi,P_X)$. The second inequality is similar to (d) in Proposition 2.1.

The first inequality in Proposition 6.1 indicates that $M_2(d)$ improves $M_1(d)$ by extending $\alpha$ to incorporate the covariates information $X$. The second inequality in Proposition 6.1 justifies the
conservativeness of $M_2(d)$ as a lower bound of $V(d)$, which can also provide the risk control. In addition, since $\text{CVaR}_\gamma(R|X, a) \leq Q_\gamma(R|X, a)$, one can have $M_2(d) \leq \mathbb{E}[Q_\gamma(R|X, d(X))]$. Then the optimal IDR under $M_2(d)$ is defined as
\[ d_2 \in \arg\max_d M_2(d). \] (35)

Based on Theorem 6.1, we have the explicit expression for $d_2$ by the following proposition.

**Proposition 6.2.** The optimal IDR under the criterion $M_2(d)$ is given by
\[ d_2(X) \in \arg\max_{a \in A} \{ \text{CVaR}_\gamma(R|X, a) \} \quad \text{almost surely}. \]

Thus, under $M_2(d)$, the optimal IDR $d_2$ is equivalent to choosing a treatment that has the largest individualized $\gamma$-CVaR among all treatments. By a similar derivation, we can also express $M_2(d)$ in a dual form. Define
\[ \mathcal{W}_2^d := \{ W \in L^1(\mathcal{T}, \mathcal{F}_1, P^d) \mid 0 \leq W(\omega_2) \leq \frac{1}{\gamma} \text{ for almost sure } \omega_2 \in \mathcal{T}, \mathbb{E}[W|X, A = d(X)] = 1 \}. \] (36)

We have the following theorem that gives the dual representation of $M_2(d)$.

**Theorem 6.2.** $M_2(d) = \inf_{W \in \mathcal{W}_2^d} \mathbb{E}^d[W|R]$.

**Duality representation of $M_2(d)$:** Similarly, for $W \in \mathcal{W}_2^d$, if we define $P^W_X(B) = \int_B W \, dP^d_X$ for any measurable set $B \in \mathcal{T}$, then $W = \frac{dP^W_X}{dP_X^d}$. Thus the optimal IDR can also be written as
\[ d_2 \in \arg\max_d \left\{ \inf_{P^W_X \ll P^d_X} \mathbb{E}_{P^W_X}[R] + \mathbb{E}_X \mathbb{E}_X \left\{ \frac{dP^W_X}{dP^d_X} \right\} \right\}, \]

where $\mathbb{E}_X \left\{ \frac{dP^W_X}{dP^d_X} \right\} = \mathbb{E}_X \left[ \zeta \left( \frac{dP^W_X}{dP^d_X} \right) \right]$. Moreover, the conditional probability density $W$ with respect to $P^d$ given $X$ can be written as $\frac{\nu_1(r|x, a=d(x))}{\int_1(r|x, a=d(x))}$ by the chain rule, according to our problem setting. Therefore, we can also express $M_2(d)$ as
\[ M_2(d) = \inf_{P^W_X} \left\{ \mathbb{E}_{P^W_X}[R] \mid P^W_X \ll P^d_X, \quad 0 \leq \frac{w(r|x, a=d(x))}{\int_1(r|x, a=d(x))} \leq \frac{1}{\gamma}, \text{ almost surely} \right\}. \] (37)

Maximizing $M_2(d)$ over the decision rule $d$ is equivalent to identifying an optimal IDR that is robust to the contamination of the outcome $R$.

**Comparisons between $M_0(d)$ and $M_2(d)$:** From a duality representation perspective, we can see $M_0(d)$ and $M_2(d)$ have substantial differences with regard to their robustness. The “minimax” sense of $M_0(d)$ in [19] considers the scenario where both distributions of the covariates $X$ and outcome $R$ are perturbed from true underlying distributions. For $M_2(d)$, Proposition 6.1 shows that $M_2(d) \geq M_1(d)$, which means considering individualized CVaR improves the outcome of a given decision rule $d$. At the same time, however, it also indicates $M_2(d)$ is not as conservative as $M_0(d)$. This can also be justified by the “minimax” representation of (37), which considers the
contamination of outcome \( R \). In the end, both \( M_0(d) \) and \( M_2(d) \) are more robust than the value framework, i.e., \( V(d) \). Therefore \( M_0(d) \) and \( M_2(d) \) may have better ability to improve generalization by quantifying the inconsistency between training and test data.

Similar to (17), theoretically we can compute \( \alpha_2(X) \) and \( d_2 \) jointly via

\[
(d_2, \alpha_2(X)) \in \arg\max_{d, \alpha \in L^2(X, \Xi, P_X)} \left\{ E[\alpha(X)] - \frac{1}{\gamma} E^d[\alpha(X) - R]_+ \right\}.
\]

(38)

Although given the data we could use empirical approximation to compute \( d_2 \) and \( \alpha_2(X) \) by assuming them belong to some functional classes, the estimation of model-dependent \( \alpha_2(X) \) is challenging. Simply restricting \( \alpha_2(X) \) to be some smooth functions may lead to potentially model mis-specifications. According to the explicit form of \( \alpha^*(X) \) in Theorem 6.1, the true \( \alpha_2(X) \) may be a non-smooth function of \( X \). For example, if the individualized \( \gamma \)-quantile \( Q_{\gamma}(R|X, A) \) is linear in both \( X \) and \( A \), then in most cases \( \alpha^*(X) \) will be a non-smooth function of \( X \). Thus it would be desirable to explore a broader structure of \( \alpha(X) \) or develop some robust estimation methods to overcome potential model misspecifications of \( \alpha_2(X) \).

6.2 Flexible Models

In the previous real data application, we include the expected value function \( V(d) \) with the proposed \( M_0(d) \) in (31) to allow more flexibility to find the best IDRs. This is roughly equivalent to maximizing \( V(d) \) among all the decision rules with \( M_0(d) \) being larger than some threshold. Such an interpretation motivates us an approach to find the optimal IDR when observing multiple outcomes after receiving treatments. Without loss of generality, suppose we can observe a risk outcome \( H \) in addition to \( R \). In general, we prefer a smaller risk outcome. By some modification on \( M_0(d) \), we can search the optimal IDRs by

\[
\max_{d \in \mathcal{D}} V(d) \quad \text{subject to} \quad \min_{\alpha \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^n (R_i + \frac{1}{\gamma} (R_i - R_j)_+) S(A_i f(X_i)) \right\} \leq t,
\]

(39)

for some pre-specified constant \( t \), so that we can control the risk outcome with the CVaR criterion. By a similar analysis as in Section 2, the resulting IDRs using (39) are the best IDRs among all the decision rules with the risk outcome of \( (1 - \gamma) \times 100\% \) of the population being less than some threshold \( t \). Given the data, we could use the same techniques in Section 3 to estimate the IDR. Specifically, by using a surrogate function \( S \), we can solve the following optimization problem:

\[
\max_{d \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^n \frac{R_i S(A_i f(X_i))}{\pi(A_i | X_i)} + \frac{\lambda}{2} \| f \|_H^2 \quad \text{subject to} \quad \min_{1 \leq j \leq n} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{(R_j + \frac{1}{\gamma} (R_i - R_j)_+) S(A_i f(X_i))}{\pi(A_i | X_i)} \right\} \leq t,
\]

(40)

which again can be formulated as optimizing a DC function with a DC constraint as we can see the minimum of a finite number of DC functions in the constraint is a DC function. However,
several issues need to be solved before proceeding. The existence of the optimal solution for the optimization problem (40) should be demonstrated. The use of a surrogate function to replace the indicator function in (39) needs to be justified. Another challenge is to establish regret bounds for Problem (39). Existing techniques in statistical learning theory cannot be directly used since there is a stochastic term in the constraint of (39). It will be an interesting direction to pursue.

6.3 Flexible Decision Rules

In practice, decision makers may want to specify the class of decision rules $D$ that can include problem-specified constraints such as fairness, which was considered in the literature such as Bhattacharya and Dupas (2012) and Athey and Wager (2017). However, it may be difficult to incorporate such constraints by using the surrogate function in our estimation. Next we show how we can get around with this difficulty. We can first rewrite Problem 21 as

$$
\max_{1 \leq j \leq n} \left\{ \max_{d \in D} \frac{1}{n} \sum_{i=1}^{n} I(A_i = d(X_i)) \left( R_j - \frac{(R_j - R_i)_{+}}{\gamma} \right) \right\}, \quad (41)
$$

by the piecewise linearity with respect to $\alpha$ in (21). Based on this representation, we can learn $d_0$ by solving $n$ different weighted classification problems and choose the best one with the largest objective value. Based on the pre-specified $D$, for each weighted classification problem, we could implement the corresponding machine learning algorithm such as decision trees. However, it raises some computational concerns by using this algorithm since we need to solve $n$ subproblems, especially when $n$ is large. Here we provide a heuristic algorithm by iteratively solving $\alpha$ and $d$ for Problem (41) as shown in Table 2 to obtain $\hat{d}$ and $\hat{\alpha}$.

**Algorithm 2** Algorithm for (41)

1: Initial $\alpha^{(0)}$ be any number in $\{R_1, \cdots, R_n\}$.
2: For the $(k+1)$ iteration, given the last $\alpha^{(k)}$, compute $d^{(k)}$ via

$$
\max_{d \in D} \frac{1}{n} \sum_{i=1}^{n} I(A_i = d(X_i)) \left( \alpha^{(k)} - \frac{(\alpha^{(k)} - R_i)_{+}}{\gamma} \right), \quad (42)
$$

using any machine learning algorithm.
3: Given $d^{(k)}$, obtain $\alpha^{(k+1)}$ by solving

$$
\max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} I(A_i = d^{(k)}(X_i)) \left( \alpha - \frac{(\alpha - R_i)_{+}}{\gamma} \right). \quad (43)
$$

4: The algorithm stops when $\alpha^{(k)} = \alpha^{(k+1)}$.

Based on Problem (41), we establish the corresponding regret bound with respect to $(\hat{d}, \hat{\alpha})$. Denote $(d^*, \alpha^*) \in \arg\max_{d \in D, \alpha \in \mathbb{R}} M_0(d, \alpha)$ and $M_n(d, \alpha)$ as an empirical approximation of $M_0(d, \alpha)$. We also define the Vapnik-Chervonenkis dimension of $D$ as $VC(D)$. Then we have the following theorem.

25
Theorem 6.3. Given $R$ is uniformly bounded by $C_0$, then for any $\epsilon > 0$, with probability at least $1 - \epsilon$,

$$M_0(d^*, \alpha^*) - M_0(\hat{d}, \hat{\alpha}) \leq c_1 \sqrt{\frac{VC(D)}{n}},$$  \hspace{1cm} (44)$$

with some constant $c_1$ that depends on $C_0, \gamma, c$ and $\epsilon$. In particular,

$$M_0(d^*, \alpha^*) - \mathbb{E}[M_0(\hat{d}, \hat{\alpha})] \leq c_2 \sqrt{\frac{VC(D)}{n}}$$  \hspace{1cm} (45)$$

for some constant $c_2$.

While the result is the similar to that in Kitagawa and Tetenov (2018), the proof is much more involved because we need to deal with $\alpha$ and $d$ together.

References

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. Mathematical finance, 9(3):203–228, 1999.

S. Athey and S. Wager. Efficient policy learning. arXiv preprint arXiv:1702.02896, 2017.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’09, pages 129–138, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557040. URL http://doi.acm.org/10.1145/1557019.1557040

D. Bhattacharya and P. Dupas. Inferring welfare maximizing treatment assignment under budget constraints. Journal of Econometrics, 167(1):168–196, 2012.

G. Chamberlain. Bayesian aspects of treatment choice. In The Oxford Handbook of Bayesian Econometrics. Oxford University Press, 2011.

J. Chen, H. Fu, X. He, M. R. Kosorok, and Y. Liu. Estimating individualized treatment rules for ordinal treatments. Biometrics, 74(3):924–933, 2018.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. J. Mach. Learn. Res., 18(1):6070–6120, Jan. 2017. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=3122009.3242024

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.

R. Dehejia. When is ate enough? risk aversion and inequality aversion in evaluating training programs. In Modelling and Evaluating Treatment Effects in Econometrics, pages 263–287. Emerald Group Publishing Limited, 2008.
M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

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.

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

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.

K. Hirano and J. R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

N. Kallus. Balanced policy evaluation and learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8895–8906. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8105-balanced-policy-evaluation-and-learning.pdf.

M. Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 98(1):111–131, 2016.

T. Kitagawa and A. Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

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

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772758. URL http://doi.acm.org/10.1145/1772690.1772758.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, pages 297–306, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935878. URL http://doi.acm.org/10.1145/1935826.1935878.

L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on
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*, 37(26):3776–3788, 2018.

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

C. F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

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(Jul):1073–1097, 2005.

M. Nouiehed, J.-S. Pang, and M. Razaviyayn. On the pervasiveness of difference-convexity in optimization and statistics. *arXiv:1704.03535*, 2017.

J.-S. Pang, M. Razaviyayn, and A. Alvarado. Computing b-stationary points of nonsmooth dc programs. *Mathematics of Operations Research*, 42(1):95–118, 2016.

G. C. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic Constrained Optimization*, pages 272–281. Springer, 2000.

Z. Qi and Y. Liu. D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, 12(2):3601–3638, 2018.

Z. Qi, D. Liu, H. Fu, and Y. Liu. Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association*, 2019.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *The 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.

R. Rockafellar. *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics, 1974. doi: 10.1137/1.9781611970524. URL `https://epubs.siam.org/doi/abs/10.1137/1.9781611970524`.
R. T. Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability engineering & system safety*, 95(5):499–510, 2010.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

S. Sarykalin, G. Serraino, and S. Uryasev. Value-at-risk vs. conditional value-at-risk in risk management and optimization. *Tutorials in Operations Research*, pages 270–294, 2008.

L. J. Savage. The theory of statistical decision. *Journal of the American Statistical association*, 46(253):55–67, 1951.

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*, 29(4):640, 2014.

I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, pages 575–607, 2007.

J. Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015. URL [http://jmlr.org/papers/v16/swaminathan15a.html](http://jmlr.org/papers/v16/swaminathan15a.html)

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the cvr via sampling. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, pages 2993–2999. AAAI Press, 2015. ISBN 0-262-51129-0. URL [http://dl.acm.org/citation.cfm?id=2888116.2888133](http://dl.acm.org/citation.cfm?id=2888116.2888133)

Y. Tao and L. Wang. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*, 73(1):145–155, 2017.

A. Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.
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.

L. Wang, Y. Zhou, R. Song, and B. Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

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

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.

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, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151, 2015.

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.