Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Robust outcome weighted learning for optimal individualized treatment rules

Sheng Fu, Qinying He, Sanguo Zhang & Yufeng Liu

To cite this article: Sheng Fu, Qinying He, Sanguo Zhang & Yufeng Liu (2019) Robust outcome weighted learning for optimal individualized treatment rules, Journal of Biopharmaceutical Statistics, 29:4, 606-624, DOI: 10.1080/10543406.2019.1633657

To link to this article: https://doi.org/10.1080/10543406.2019.1633657

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2019, VOL. 29, NO. 4, 606–624
https://doi.org/10.1080/10543406.2019.1633657

Robust outcome weighted learning for optimal individualized treatment rules

Sheng Fu[a], Qinying He[b], Sanguo Zhang[c,d], and Yufeng Liu [e]

[a]Department of Industrial and Systems Engineering, National University of Singapore, Singapore; [b]College of Economics and Management, South China Agricultural University, Guangzhou, China; [c]School of Mathematical Science, University of Chinese Academy of Sciences, Beijing, China; [d]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China; [e]Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

ABSTRACT

Personalized medicine has received increasing attentions among scientific communities in recent years. Because patients often have heterogenous responses to treatments, discovering individualized treatment rules (ITR) is an important component of precision medicine. To that end, one needs to develop a proper decision rule using patient-specific characteristics to maximize the expected clinical outcome, i.e. the optimal ITR. Recently, outcome weighted learning (OWL) has been proposed to estimate optimal ITR under a weighted classification framework. Since most of commonly used loss functions are unbounded, the resulting ITR may suffer similar effects of outliers as the corresponding classifiers. In this paper, we propose robust OWL (ROWL) to build more stable ITRs using a new family of bounded and non-convex loss functions. Moreover, we extend the proposed ROWL method to the multiple treatment setting under the angle-based classification structure. Our theoretical results show that ROWL is Fisher consistent, and can provide the estimation of rewards' ratios for the resulting ITRs. We develop an efficient difference of convex functions algorithm (DCA) to solve the corresponding nonconvex optimization problem. Through analysis of simulated examples and a real medical dataset, we demonstrate that the proposed ROWL method yields more competitive performance in terms of the empirical value function and the misclassification error than several existing methods.

ARTICLE HISTORY
Received 25 September 2017
Accepted 16 September 2018

KEYWORDS
Angle-based classifiers; multiple treatments; non-convex optimization; precision medicine; robustness; soft and hard classification

1. Introduction

In modern medical studies, especially study of chronic diseases, patients can show significant heterogeneity in response to different treatments. For example, a treatment may work well for some patients with certain characteristics, but may have no effect for others (Ellsworth et al. 2010; Mancinelli et al. 2000). The target of personalized medicine is to maximize the clinical outcome or reward of patients. To improve the effect of treatments significantly, one should find proper individualized treatment rules (ITR), based on the patients' genomic or prognostic information, rather than a "one size fits all" approach. The optimal ITR is a function with maximum expected benefit from the treatment, which maps from the patient characteristics' space into the treatment decision space.

CONTACT Yufeng Liu yfliu@email.unc.edu Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA

Supplementary material can be accessed here.

In the recent literature, many machine learning techniques have been introduced to build the optimal ITR, especially for the scenario with binary treatments. There are two main strategies to estimate an ITR. One is the regression-based methods, which model the conditional mean of outcome based on covariates, treatments and their interaction effects (Fan et al. 2017; Qian and Murphy 2011; Tian et al. 2014; Xiao et al. 2019; Zhang et al. 2012). The other is the classification-based methods, which convert the estimation of the optimal ITR into a weighted classification problem. Zhao et al. (2012) proposed an outcome weighted learning (OWL) strategy which connects the ITR problem with the weighted support vector machine. Zhou et al. (2017) extended the OWL and proposed residual weighted learning (RWL) under the weighted classification framework. Furthermore, Liu et al. (2016) generalized OWL techniques to estimate binary dynamic treatment regimens. Laber and Zhao (2015) devised a new estimation method based on decision trees to obtain the optimal ITR.

Despite the great success in ITR estimation with binary treatments, extending the idea for multiple treatment scenarios is still not fully explored. For estimating ITRs with multiple treatments, one can design similar methods under the weighted multicategory classification framework. One typical approach is to use the sequential binary methods, such as one-versus-rest (OVR) and one-versus-one (OVO), which reduce the problem into multiple binary ones. Such extensions can be suboptimal, since they do not use the data jointly (Liu 2007). Another approach is to handle all treatments together using simultaneous classification methods. Recently, Chen et al. (2018) employed a data duplicate strategy to solve the ITR estimation problem with multiple ordinal treatments. To estimate ITRs with multiple treatments, Zhang et al. (2019) proposed a weighted angle-based classifier with a flexible margin-based loss function, which includes the original binary OWL as a special case.

For classification-based ITR estimation methods, the performance heavily depends on how well the corresponding classifier works. In the large-margin classification literature, it is known that classifiers using unbounded loss functions may suffer from the existence of extreme outliers. In practice, misclassified points that are far from points of their own classes may have large loss values and heavily affect the performance. Wu and Liu (2007) pointed out that truncation of the unbounded loss helps to decrease the impact of outliers and yield more stable classifier. Zhang et al. (2018) utilized the truncated hinge loss function for robust multicategory classification. Instead of truncating the loss functions, Wu and Liu (2013) and Fu et al. (2018) proposed new adaptively weighted large-margin classification techniques to achieve robustness. The truncated loss function is bounded and consequently can be more robust to outliers. Our motivation is to develop robust weighted classification methods for robust single-stage ITR estimation.

In this article, we design a new family of robust large-margin loss functions, which are smooth and bounded, and apply them to estimate robust ITRs. We name the proposed method as robust outcome weighted learning (ROWL). ROWL is a unified large margin angle-based ITR learning framework to handle binary and multiple treatment problems, and the estimated ITRs can be more stable to outliers. Moreover, the ITR estimator from ROWL is Fisher consistent, and can provide estimated ratio of clinical rewards for each treatment pair, which offers more information on the estimated ITR. For implementation of ROWL, we develop an efficient difference of convex functions algorithm (DCA) to solve the corresponding nonconvex problem.

The rest of this article is organized as follows. In Section 2, we briefly review ITR and the OWL for ITR learning with binary and multiple treatments, and propose a new family of loss functions and introduce our ROWL method. In Section 3, we present several nice statistical properties for ROWL, including Fisher consistency and theoretical ratios of clinical rewards under some conditions. We design the efficient DCA to solve the nonconvex optimization problem of ROWL in Section 4. Simulated examples as well as an application to a real medical dataset are used to demonstrate the effectiveness of ROWL in Section 5. Some discussions are given in Section 6. All proofs are included in the Appendix.

2. Methodology

We first review some basic concepts and notations of ITR estimation with binary and multiple treatments, and introduce several classification-based ITR estimation approaches in Section 2.1. We propose a new family of robust loss functions, and build the corresponding angle-based framework for estimating robust ITR with multiple treatments in Section 2.2.

2.1. *Individualized treatment rule and outcome weighted learning*

We assume the training data $\mathcal{T} = \{(x_i, a_i, r_i); i = 1, \ldots, n\}$ are from an underlying distribution $P(\boldsymbol{X}, A, R)$, where $x_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes the prognostic variables of a patient, the treatment $a_i \in \mathcal{A} = \{1, \ldots, k\}$ is independent of predictor $\boldsymbol{X}$, and $r_i \in \mathbb{R}$ is the corresponding clinical outcome of a patient, namely, reward. Without loss of generality, we assume larger values of R are more desirable. Thus, an ITR is a decision rule which maps from the space of prognostic variables $\mathcal{X}$ into the space of treatments $\mathcal{A}$. Among all possible rules, an optimal ITR is the one which maximizes the expected reward among all possible rules. Our learning target is to estimate the optimal ITR.

We denote the distribution of $(\boldsymbol{X}, A, R)$ by $\mathbb{P}$ and expectation with respect to $\mathbb{P}$ by $\mathbb{E}$. For any rule $A = D(\boldsymbol{X})$, it means that the treatment is determined by rule D, and let $\mathbb{P}^D$ be the conditional distribution $\{\boldsymbol{X}, A, R|A = D(\boldsymbol{X})\}$. The expectation with respect to $\mathbb{P}^D$ is $\mathbb{E}^D$. We assume that $\Pr(A = a) > 0$ for any $a \in \{1, \ldots, k\}$. It can be verified that $\mathbb{P}^D$ is absolutely continuous with respect to $\mathbb{P}$, and the Radon-Nikodym derivative $\frac{d\mathbb{P}^D}{d\mathbb{P}} = \frac{\mathbb{1}(a=D(X))}{\pi_a}$, where $\mathbb{1}(\cdot)$ is the indicator function, and $\pi_a = \Pr(A = a)$. Then, the expected reward under the given ITR D is

$$\mathbb{E}^D(R) = \int R d\mathbb{P}^D = \int R \frac{d\mathbb{P}^D}{d\mathbb{P}} d\mathbb{P} = \int R \frac{\mathbb{1}(A = D(X))}{\pi_A} d\mathbb{P} = \mathbb{E}\left[\frac{\mathbb{1}(A = D(X))}{\pi_A} R\right].$$

Consequently, the optimal ITR D^* can be defined as $D^* \in \arg\sup_D \mathbb{E}\left[\frac{\mathbb{1}(A=D(X))}{\pi_A} R\right]$. If D^* is an optimal ITR for any $x \in \mathcal{X}$, it indicates that the expected reward corresponding to D^* is larger than any other treatments in $\{1, \ldots, k\} \backslash D^*(x)$. One can show that finding D^* is equivalent to finding the minimizer of the following problem,

$$E\left[\frac{\mathbb{1}(A \neq D(X))}{\pi_A} R\right] = \int R \frac{\mathbb{1}(A \neq D(X))}{\pi_A} d\mathbb{P}. \tag{2.1}$$

The term in (2.1) can be regarded as a weighted misclassification error rate, which converts ITR learning into a classification problem. Based on the observed dataset $\mathcal{T}$, one can approximate the weighted misclassification error via the empirical loss

$$\frac{1}{n} \sum_{i=1}^{n} \frac{r_i}{\pi_{a_i}} \mathbb{1}(a_i \neq D(x_i)). \tag{2.2}$$

The main goal is to find a minimizer $\hat{D}_n$ of (2.2) with respect to D. However, because the involved 0–1 loss is discontinuous and nonsmooth, solving problem (2.2) directly can be NP-hard and intractable. To that end, one can apply a surrogate loss function instead of the 0–1 loss. In particular, when $k = 2$ with encoded treatment labels $\mathcal{A} = \{-1, +1\}$, for a single decision function f, $D(x)$ can be expressed as $\text{sign}(f(x))$ to estimate ITR. Similar to the ordinary binary classifier with a general surrogate loss $\ell(\cdot)$, under the standard *loss + penalty* regularization framework, the outcome weighted learning (OWL) method for ITR estimation solves the following problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{r_i}{\pi_{a_i}} \ell(a_i f(x_i)) + \lambda J(f), \tag{2.3}$$

where $\mathcal{F}$ is the candidate space for f, the first part of the objective function is the empirical loss term on the training dataset, $J(f)$ is a penalty term on f to prevent overfitting, and $\lambda > 0$ is a tuning parameter to control the balance of loss and penalty. For example, when the hinge loss is employed, problem (2.3) becomes SVM-type OWL (Zhao et al. 2012).

Although the OWL framework (2.3) allows the use of different loss functions to obtain various ITR estimators, there still exists several drawbacks in practice. First, the original OWL in Zhao et al. (2012) requires all outcomes be nonnegative, and such a requirement can be too strict. Second, the ITR estimation by OWL can be easily affected by a simple shift of outcomes, which makes the corresponding ITR unstable. To overcome these drawbacks, Zhou et al. (2017) proposed a two-step procedure, residual weight learning (RWL). Because clinical outcomes may not be comparable among subjects with different clinical covariates, they first built a regression model between clinical outcomes and clinical covariates. After removing common covariates effects, the residuals can be more comparable. Second, they solve the residual weighted classification problem with a smoothed ramp loss function. Due to the use of this special loss, Zhou et al. (2017) also showed that RWL can achieve robustness to some extent.

So far, the aforementioned OWL and RWL methods only focus on ITR estimation with binary treatments. When there are multiple treatments ($k>2$), the corresponding ITR learning can be more complicated. One direct approach is to use sequential binary methods, which may yield suboptimal performance (Liu 2007; Zhang et al. 2019). Another approach is to build a k-category simultaneous classifier using k functions with a sum-to-zero constraint (Lee et al. 2004; Liu and Yuan 2011; Zhang and Liu 2013). Such a constraint can help to reduce the parameter space, and ensure good statistical properties such as Fisher consistency. However, solving the corresponding optimization problem needs more extra computational cost.

To further improve multicategory classifiers, Zhang and Liu (2014) recently proposed a multicategory angle-based classification framework using $k-1$ functions without the sum-to-zero constraint, which can enjoy more efficient computation. In this paper, we focus on the angle-based classification to handle ITR problems with multiple treatments, and propose a new family of loss functions to achieve robust ITR estimation in Section 2.2.

2.2. *Robust outcome weighted learning*

The angle-based classification structure is well designed for multicategory classification problems. Under the angle-based framework, we propose a new robust OWL (ROWL) method to handle ITR problems with multiple treatments. First, we briefly introduce the angle-based structure. Define a simplex $\mathcal{W}$ in $\mathbb{R}^{k-1}$ with k-vertex vectors $\{\mathcal{W}_1, \ldots, \mathcal{W}_k\}$ standing for k class labels, such that

$$\mathcal{W}_j = \begin{cases} \frac{1}{\sqrt{k-1}}\mathbf{1}_{k-1}, & j = 1 \\ -\frac{1+\sqrt{k}}{(k-1)^{3/2}}\mathbf{1}_{k-1} + \sqrt{\frac{k}{k-1}}e_{j-1}, & 2 \le j \le k \end{cases},$$

where $\mathbf{1}_{k-1} \in \mathbb{R}^{k-1}$ is a vector of 1's, and $e_j \in \mathbb{R}^{k-1}$ is a vector of 0's except the jth element being 1. Consider a $(k-1)$-component decision function $\boldsymbol{f} = (f_1, \ldots, f_{k-1})^T \in \mathbb{R}^{k-1}$, which maps x from the original space into $\mathbb{R}^{k-1}$. The vector $\boldsymbol{f}$ can introduce k angles $\{\angle(\boldsymbol{f}(x), \mathcal{W}_j); j = 1, \ldots, k\}$, with respect to k vertices. The prediction rule for newly observed data x is $\hat{Y}(x) = \arg\min_j \angle(\boldsymbol{f}(x), \mathcal{W}_j) = \arg\max_j \langle \boldsymbol{f}(x), \mathcal{W}_j \rangle$. Therefore, for an arbitrary binary surrogate loss function $\ell(\cdot)$, Zhang and Liu (2014) proposed the large margin angle-based classification with the following optimization problem,

$$\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i} \rangle) + \lambda J(\boldsymbol{f}). \tag{2.4}$$

Without the redundant sum-to-zero constraint which is imposed on usual simultaneous classifiers, the angle-based classifier can achieve a faster computational speed, and better classification performance.

Among various large-margin classifiers, there are two main groups of methods: hard and soft classifiers (Liu et al. (2011)). Determined by its loss function, a hard classifier such as the SVM only focuses on estimating the decision boundary, while a soft classifier such as logistic regression can estimate the class conditional probability and the decision boundary simultaneously. The performances of soft and hard classifiers depend on the particular classification problems and the primary learning goal. The large-margin unified machine (LUM) loss proposed by Liu et al. (2011) covers a broad range of margin-based classifiers, including both hard and soft ones.

To generalize binary OWL methods to multicategory treatment scenarios, a natural direction is to connect multicategory angle-based classification methods with the OWL strategy. For instance, Zhang et al. (2019) proposed a multicategory outcome weighted margin-based learning method. Using LUM loss functions, they investigated the effects of soft and hard classifiers for ITR estimation.

Despite developments in angle-based classifiers, the corresponding classification-based ITRs can possibly suffer from the effect of potential outliers, which may result in unstable performance. In the classification literature, the choice of loss functions has an important impact on the performance of classifiers. Thus, finding a proper robust loss function is necessary and meaningful, and it is the key to build robust classifiers. In this paper, we propose a new family of robust loss functions as follows,

$$V(u) = \begin{cases} 1 + \frac{1}{\rho}[1 - e^{\rho u}], & \text{if } u < 0 \\ e^{-u}, & \text{if } u \geq 0 \end{cases}, \tag{2.5}$$

where $\rho > 0$ is the scale parameter to control the height of $V(u)$. The loss $V(u)$ is smooth and upper bounded by $1 + \frac{1}{\rho}$. It is a hybrid loss with two separate parts. The positive part is the exponential loss which is used in Adaboost. The negative part is the truncated exponential loss, which targets on controlling the influences of outliers. Such a loss function with a proper ρ can yield a robust and soft classifier.

The left panel of Figure 1 displays the plots of $V(\cdot)$ with several settings of ρ. We can see that the robust loss $V(\cdot)$ forms a very rich family. Note that ρ plays an important role for $V(\cdot)$, and it also determines the decaying speed and the height of the left part. When ρ increases, the left part of $V(\cdot)$ becomes more flat. In particular, we investigate several interesting settings of ρ as follows, which can connect some well-known loss functions and classifiers.

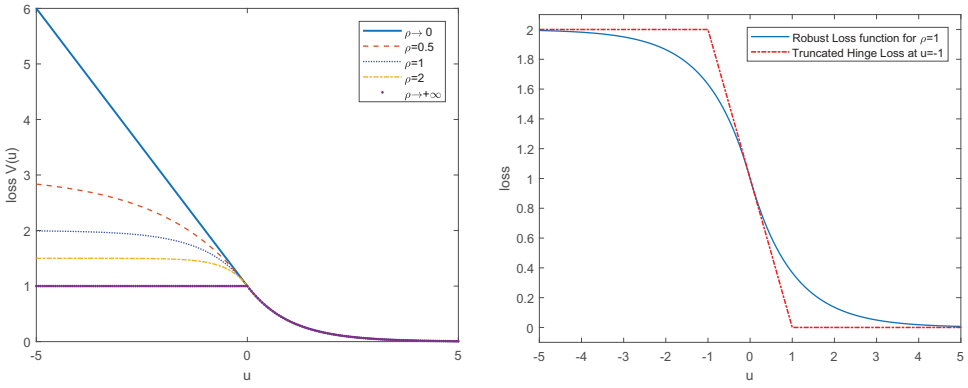

Figure 1. Plots of proposed loss *V(u)* with ρ = 0.5, 1, 2 and $\rho \to 0 + \infty$ (left), and plots of proposed loss and the truncated hinge loss (right).

If $\rho \to 0_+$, the limit of $V(u)$ is $1-u$ for $u<0$, and the positive part remains the same as before. It has an interesting connection with the hinge loss and exponential loss. As a consequence, the corresponding classifier can be viewed as a hybrid of the standard SVM and Adaboost, which is a special example of the LUM family (Liu et al. 2011). Such a smooth limiting loss results in a soft classifier.

If $\rho \to +\infty$, the limit of $V(u)$ is $\min(e^{-u}, 1)$, which can be regarded as the truncated exponential loss at the origin. Such a truncated loss was also studied by Wu and Liu (2007), and it can yield robust performance. The nonsmoothness of the limiting function creates the corresponding hard classifier.

If $\rho \in (0, +\infty)$, the robust loss $V(\cdot)$ is closely related to the truncated hinge loss proposed by Wu and Liu (2007). To illustrate this, we show the figures of $V(\cdot)$ with $\rho = 1$ and truncated hinge loss at -1 on the right panel of Figure 1. Note that $V(\cdot)$ can be viewed as an envelope of the truncated hinge loss, and it is upper bounded by the truncated hinge loss at $-\frac{1}{\rho}$. Therefore, similar to the truncated hinge loss, the proposed loss $V(\cdot)$ is robust to outliers as well. The robust smooth loss function $V(\cdot)$ leads to a soft classifier, while the nonsmooth truncated hinge loss only delivers a hard classifier.

Based on the above discussions, we can conclude that among different values of the scale parameter ρ, the proposed robust loss function $V(u)$ ranges from convex to nonconvex, and covers a spectrum of classifiers from soft to hard ones.

Besides robustness, another challenging problem is how to deal with observations with negative rewards. When there exists negative rewards, the resulting object function of the original OWL becomes nonconvex. In order to keep the objective function convex, Zhao et al. (2012) recommended to shift all rewards by a constant to ensure positiveness of all weights. Zhou et al. (2017) noted that such a constant shift for the rewards may yield suboptimal estimators, and proposed the nonconvex RWL method. To better handle the data with negative rewards, Chen et al. (2018) and Zhang et al. (2019) made a distinction between the positive and negative rewards, and proposed a new inverted loss function for negative rewards. Depending on the sign of observed clinical outcome value r, the modified loss function for ℓ is

$$\ell_r(u) = \begin{cases} \ell(u) & \text{if } r \geq 0, \\ \ell(-u) & \text{if } r < 0 \text{ (the inverted loss)}, \end{cases}$$

which can be simplified as $\ell_r(u) = \ell(\text{sign}(r) \cdot u)$. The inverted loss preserves the convexity of optimization, which efficiently alleviates the impact of nonnegative rewards.

Under the angle-based framework, we utilize the modified robust loss to estimate robust ITR with multiple treatments, and propose the ROWL method with the optimization problem as follows,

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}} V(\text{sign}(r_i) \cdot \langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i} \rangle) + \lambda J(\boldsymbol{f}). \tag{2.6}$$

The unified ROWL framework can deal with both binary and multicategory treatment problems, and estimate robust ITR with both positive and negative rewards. ROWL with $\rho > 0$ is a soft classifier-based ITR estimation technique. The objective function of ROWL with $\rho > 0$ is nonconvex, thus it is more complicated and challenging to implement than the ordinary convex OWL.

Similar to RWL in Zhou et al. (2017), we can also extend residual weighting techniques to ROWL. First, we estimate the main effect model between clinical outcomes and prognostic covariates, and assume the estimated regression model be $\tilde{r} = \hat{g}(x)$. Second, we obtain the fitted residuals $\{\hat{r}_i = r_i - \hat{g}(x_i), i = 1, \ldots, n\}$ as new "outcomes", and use them to replace the r_i in the ROWL framework (2.6). Thus, this residual-based ROWL generalizes the original binary RWL in Zhou et al. (2017) to handle multiple treatment problems.

3. Statistical properties

We explore some theoretical properties of the proposed ROWL, including Fisher consistency in Section 3.1, and estimation of the rewards' ratio for the resulting ITR in Section 3.2.

3.1. *Fisher consistency*

Fisher consistency is a desirable theoretical property of a classification loss function (Lin 2002; Liu 2007), and is also known as "classification calibration" (Bartlett et al. 2006). For angle-based classification ITR learning, Fisher consistency of multicategory cases is more involved than that of binary settings (Zhang et al. 2019). Here, we consider a general loss $\ell(\cdot)$, which includes $V(\cdot)$ as a special case. Based on outcome-weighted and angle-based ITR learning, one can define the conditional expected loss for a certain $x \in \mathcal{X}$ as follows,

$$S(x) = \mathbb{E}\left[\frac{|R|}{\pi_A}\ell(\text{sign}(R)\cdot\langle \boldsymbol{f}(\boldsymbol{X}), \mathcal{W}_A\rangle)|\boldsymbol{X} = x\right], \tag{3.1}$$

where the expectation is taken with respect to the marginal distribution of (R, A) for a given x. Set the theoretical minimizer of $S(x)$ as $\boldsymbol{f}^*(x) = \arg\min_f S(x)$. Note that $\boldsymbol{f}^*$ depends on the loss function $\ell(\cdot)$. For classification-based ITR learning, it is Fisher consistent when the predicted treatment based on $\boldsymbol{f}^*$ leads to the largest expected outcome, i.e. $\arg\max_a\langle \boldsymbol{f}^*(x), \mathcal{W}_a\rangle = \arg\max_j R_j(x)$, where $R_j(x) = \int (R|\boldsymbol{X} = x, A = j)d\mathbb{P}$ is the expected reward for the given treatment j at a fixed x.

Define the positive part of a conditional reward to be $R_j^+(x) = \int (R|\boldsymbol{X} = x, A = j)\mathbb{1}(R>0)d\mathbb{P}$, and the negative part to be $R_j^-(x) = \int (R|\boldsymbol{X} = x, A = j)\mathbb{1}(R<0)d\mathbb{P}$. One can check that $R_j(x) = R_j^+(x) + R_j^-(x)$. For treatment j on patients with the predictor vector x, correspondingly, $R_j^+(x)$ and $R_j^-(x)$ can be used to measure the positive and adverse effects. The next assumption requires that $R_j^+(x)$ and $R_j^-(x)$ of the best treatment for a given patient should not be small.

Assumption 1. *For a patient with the predictor vector x, denote the best treatment by j (i.e., $R_j(x) > R_i(x)$ for any $i \neq j$). The reward of the best treatment should be positive, i.e., $R_j(x) > 0$. Moreover, $R_j^+(x) \geq R_i^+(x)$ and $R_j^-(x) \geq R_i^-(x)$ for any $i \neq j$, and these two equalities cannot hold simultaneously.*

Assumption 1 is desirable and reasonable, and often necessary for practical problems. In particular, for any patient, we should expect that the best treatment does not have a large probability of adverse effects, and its adverse effects are relatively mild. We also hope that the best treatment has a high probability of positive effects, and its positive effects are relatively strong.

For ITR learning with binary treatments, Zhao et al. (2012) showed the SVM-type OWL enjoyed Fisher consistency for all non-negative rewards. For the multi-treatment ITR estimation with arbitrary rewards, Zhang et al. (2019) addressed a sufficient condition to achieve Fisher consistency, which required that the employed loss be convex and strictly decreasing. Our focus here is on the new robust nonconvex loss function $V(\cdot)$, which has not yet been considered by previous results.

For finding optimal ITR with multiple treatments, the following theorem shows another sufficient condition to achieve Fisher consistency.

Theorem 1. *For ITR learning using (3.1), and suppose Assumption 1 is valid, then the method is Fisher consistent if $\ell(\cdot)$ is differentiable with $\ell'(u) < 0$ for all u.*

Compared with the conditions of Theorem 5 in Zhang et al. (2019), Theorem 1 emphasizes on smoothness instead of convexity for loss function. Since the new robust loss $V(\cdot)$ with $\rho \in [0, +\infty)$ satisfy the conditions in Theorem 1, we can conclude that they are always Fisher consistent. However, the special case with $\rho \to +\infty$ and $V(u) = \min(e^{-u}, 1)$ may be an exception of Theorem 1, due to its nondifferentiability. The following theorem shows the Fisher consistency for this special loss.

Theorem 2. *For ITR learning using (3.1) with a specified loss* $\ell(u) = \min(e^{-u}, 1)$, *and suppose Assumption 1 is valid, then the corresponding method is Fisher consistent.*

Therefore, by Theorem 1 and 2, we can claim that for the loss function $V(\cdot)$ with $\rho \in [0, \infty]$, the corresponding ROWL method enjoys Fisher consistency, even there exists instances with negative rewards.

3.2. *Estimation for ratios of rewards*

In standard margin-based classification, besides label prediction, the conditional probability of belonging to each class can be very meaningful and informative as well. Thus, the class conditional probability estimation is an important problem in classification (Wang et al. 2008; Zhang and Liu 2013, 2014). In particular, for soft classification-based ITR learning, the estimated ratio of clinical rewards for each treatment pair is needed, rather than the class conditional probability. Such information may help doctors to further compare treatments for deciding prescriptions (Zhang et al. 2019). Here, we focus on the ratios of expected rewards for all treatment pairs.

Consider the angle-based ITR learning with a general smooth loss $\ell(\cdot)$ in (3.1). The following theorem demonstrates theoretical estimation for the ratio of expected rewards' pair under certain conditions.

Theorem 3 *For an arbitrary differentiable loss function* $\ell(\cdot)$ *with* $\ell'(u) < 0$, *if all k random rewards satisfy that* $R \geq 0$, *then for any* $i \neq j \in \{1, \ldots, k\}$, *we have*

$$c_{i,j} = \frac{R_i(x)}{R_j(x)} = \frac{\ell'(\langle \boldsymbol{f}^*(x), \mathcal{W}_j \rangle)}{\ell'(\langle \boldsymbol{f}^*(x), \mathcal{W}_i \rangle)}.$$

Note that Theorem 3 covers Theorem 6 in Zhang et al. (2019) as a special example, and does not require the convexity of loss function $\ell(\cdot)$. Furthermore, we can apply Theorem 3 for the robust loss $V(\cdot)$. For the ROWL using (2.6), once $\hat{\boldsymbol{f}}(x)$ is obtained for a new patient with clinical covariates x, we can estimate the ratio of rewards between ith and jth treatments by $\frac{V'(\langle \hat{\boldsymbol{f}}(x), \mathcal{W}_j \rangle)}{V'(\langle \hat{\boldsymbol{f}}(x), \mathcal{W}_i \rangle)}$.

4. Computational algorithms

When utilizing the unbounded loss $V(u)$ with $\rho = 0$ (denoted as $V_0(u)$), consequently, problem (2.6) is convex if $V_0(u)$ and $J(f)$ are convex. Then it can then be solved by classical optimization methods, such as those in Boyd and Vandenberghe (2004).

Here, we focus on developing optimization algorithms for the proposed ROWL (2.6) with nonconvex loss $V(u)(\rho > 0)$, which requires special nonconvex minimization techniques. Note that $V(u)$ can be decomposed as a difference of two convex functions, $V(u) = S(u) - T(u)$, where

$$S(u) = \begin{cases} 1 - u, & \text{if } u < 0 \\ e^{-u}, & \text{if } u \geq 0 \end{cases} \quad \text{and} \quad T(u) = \begin{cases} \frac{1}{\rho}[e^{\rho u} - 1] - u, & \text{if } u < 0 \\ 0, & \text{if } u \geq 0. \end{cases}.$$

Figure 2 illustrates the decomposition for $V(u)$ with $\rho = 1$. Utilizing the fact $V = S - T$, we design an efficient DCA to handle ROWL (An and Tao 1997; Liu et al. 2005; Wu and Liu 2007). DCA solves

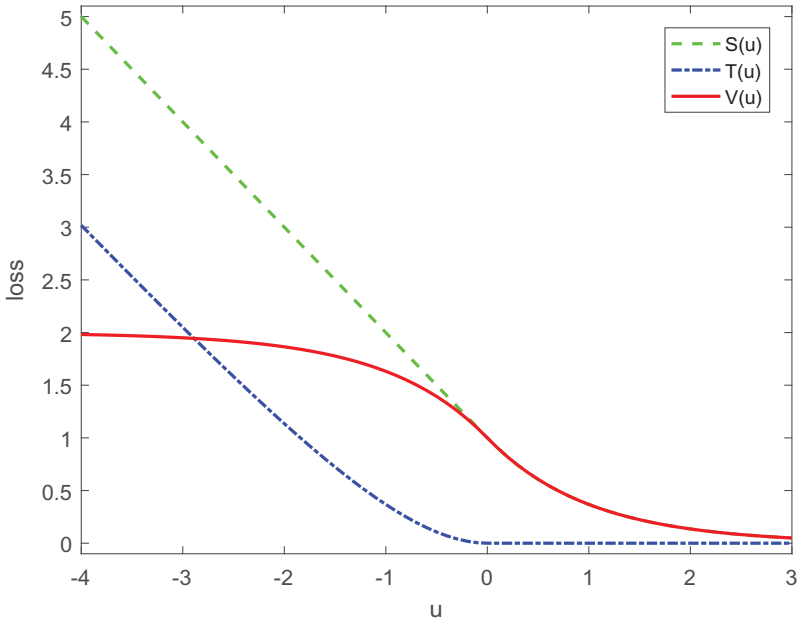

Figure 2. Plots of the functions *S(u)*; *T(u)*, and *V(u)* with $V = S - T$ and $\rho = 1$.

the nonconvex minimization problem via minimizing a sequence of convex subproblems, and its procedure can be summarized in Algorithm 1 as follows.

Algorithm 1. [The DCA procedure for minimizing $Q(\Theta) = Q_{vex}(\Theta) + Q_{cav}(\Theta)$]
1. Initialize Θ_0.
2. Repeat $\Theta_{t+1} = \arg\min_{\Theta} Q_{vex}(\Theta) + \left\langle \frac{\partial Q_{cav}(\Theta)}{\partial \Theta}\Big|_{\Theta=\Theta_t}, \Theta \right\rangle$ until convergence of Θ_t.

We show the technical details of DCA for the linear case in Section 4.1, and then extend them to the nonlinear setting through kernel mapping in Section 4.2.

4.1. *Linear learning*

Let $f_q(x) = \beta_q^T x_i + b_q (q = 1, \ldots, k-1)$, where $\beta_q \in \mathbb{R}^d$ and b_q are parameters of interest. We employ L_2 penalty $J(\boldsymbol{f}) = \frac{1}{2}\sum_{q=1}^{k-1} \beta_q^T \beta_q$ to prevent overfitting. For simplicity of the notations, denote the vector of parameters as $\boldsymbol{\Theta} = (b_1, \ldots, b_{k-1}, \beta_1^T, \ldots, \beta_{k-1}^T)^T$. Based on the decomposition $V = S - T$, the linear learning for the ROWL method solves the following optimization problem,

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \quad Q(\boldsymbol{\Theta}) = & \underbrace{\frac{\lambda}{2}\sum_{q=1}^{k-1} \beta_q^T \beta_q + \frac{1}{n}\sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}} S(\operatorname{sign}(r_i) \cdot \langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i} \rangle)}_{Q_{vex}(\Theta)} \\ & + \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}} T(\operatorname{sign}(r_i) \cdot \langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i} \rangle)}_{Q_{cav}(\boldsymbol{\Theta})}, \end{aligned} \tag{4.1}$$

where the defined $Q_{vex}(\boldsymbol{\Theta})$ and $Q_{cav}(\boldsymbol{\Theta})$ represent the convex and concave parts respectively. Then at the $(t+1)$th iteration, DCA solves the following optimization problem,

$$\begin{aligned}\min_{\boldsymbol{\Theta}} \quad & Q_{vex}(\boldsymbol{\Theta}) + \sum_{q=1}^{k-1}\left\langle \frac{\partial Q_{cav}(\boldsymbol{\Theta})}{\partial \beta_q}\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_t}, \beta_q \right\rangle + \sum_{q=1}^{k-1} b_q \frac{\partial Q_{cav}(\boldsymbol{\Theta})}{\partial b_q}\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_t} \\ = \quad & \frac{\lambda}{2}\sum_{q=1}^{k-1}\beta_q^T\beta_q + \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}} S(\mathrm{sign}(r_i)\cdot\langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i}\rangle) \\ & -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{|r_i|}{\pi_{a_i}} T'(\mathrm{sign}(r_i)\cdot\langle \boldsymbol{f}^{(t)}(x_i), \mathcal{W}_{a_i}\rangle)\mathrm{sign}(r_i)\cdot\langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i}\rangle\right\},\end{aligned} \tag{4.2}$$

where $f_q^{(t)}(x) = x^T\beta_q^{(t)} + b_q^{(t)}$ for $q = 1, \ldots, k-1$, and $T'(u)$ is the first order derivatives of $T(u)$.

The unconstrained subproblem (4.2) can be solved by the standard nonlinear convex minimization techniques. Here, we turn to a typical quasi–Newton algorithm, limited-memory Broyden-Fletcher-Goldfarb-Shanno (Nocedal 1980, L-BFGS) method. Since L-BFGS method can handle smooth objective functions efficiently, it works well in practice. For more details on L-BFGS, one can see Nocedal and Wright (2006).

4.2. *Kernel learning*

The kernel function $K(\cdot,\cdot)$ is a positive-definite function mapping from $\mathcal{X}\times\mathcal{X}$ to $\mathbb{R}$. When linear kernel $K(x,z) = x^Tz$ is applied, the corresponding classifier reduces to the linear case in Section 4.1. Various kernel functions can help to achieve complicated nonlinear decision boundaries. In the literature, the Gaussian kernel with a scale parameter $\sigma>0$, i.e. $K_\sigma(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$, is commonly used.

According to the representer theorem (Kimeldorf and Wahba 1971; Wahba 1990), we assume that $f_q(x_j) = \sum_{i=1}^n v_{i,q}K(x_i,x_j) + b_q = v_q^T\boldsymbol{K}_j + b_q (q = 1,\ldots,k-1)$, where $\boldsymbol{K}$ is an $n\times n$ kernel matrix for all training data with the (i,j)-th element being $\boldsymbol{K}(x_i,x_j)$, $\boldsymbol{K}_j$ is the jth column of $\boldsymbol{K}$, and $v_q\in\mathbb{R}^n$ is the vector of coefficients. Denote all interested parameters as $\boldsymbol{\Theta} = (b_1,\ldots,b_{k-1},v_1^T,\ldots,v_{k-1}^T)^T$. Then the optimization problem of kernel ROWL can be expressed as

$$\begin{aligned}\min_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}) = & \underbrace{\frac{\lambda}{2}\sum_{q=1}^{k-1} v_q^T\boldsymbol{K}v_q + \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}} S(\mathrm{sign}(r_i)\cdot\langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i}\rangle)}_{Q_{vex}(\boldsymbol{\Theta})} \\ & + \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}} T(\langle \boldsymbol{f}(x_i), \mathcal{W}_{a_i}\rangle)}_{Q_{cav}(\boldsymbol{\Theta})}.\end{aligned} \tag{4.3}$$

Following a similar procedure in Section 4.1, the convex subproblem of (10) at the $(t+1)$th iteration is given by

$$\begin{aligned}\min_{\boldsymbol{\Theta}} \quad & Q_{vex}(\boldsymbol{\Theta}) + \sum_{q=1}^{k-1}\left\langle \frac{\partial Q_{cav}(\boldsymbol{\Theta})}{\partial \mathbf{v}_q}\Big|_{\mathbf{Q}=\mathbf{Q}_t}, v_q \right\rangle + \sum_{q=1}^{k-1} b_q \frac{\partial Q_{cav}(\boldsymbol{\Theta})}{\partial b_q}\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_t} \\ = & \frac{\lambda}{2}\sum_{q=1}^{k-1} v_q^T K v_q + \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}} S(\mathrm{sign}(r_i)\cdot\langle f(x_i), \mathcal{W}_{a_i}\rangle) \\ & -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{|r_i|}{\pi_{a_i}} T'(\mathrm{sign}(r_i)\cdot\langle f^{(t)}(x_i), \mathcal{W}_{a_i}\rangle)\mathrm{sign}(r_i)\cdot\langle f(x_i), \mathcal{W}_{a_i}\rangle\right\},\end{aligned} \tag{11}$$

Table 1. Summary of the compared methods in numerical experiments.

methods	loss	binary	multicategory	angle-based	robustness
ℓ_1-PLS	least square	✓	✓	×	×
OWL_0	hinge loss	✓	OVO/OVR extensions	×	×
RWL	smoothed ramp	✓	OVO/OVR extensions	×	✓
LUM_MOML	LUM loss	✓	✓	✓	×
Proposed ROWL	robust loss	✓	✓	✓	$\begin{cases} \times, & \text{If } \rho \to 0 \\ \surd, & \text{If } \rho \neq 0 \end{cases}$

where $f_q^{(t)}(x) = \sum_{i=1}^n v_{i,q}^{(t)} K(x_i, x) + b_q^{(t)}$ for $q = 1, \ldots, k-1$. One can also employ L-BFGS algorithm to solve problem (11).

5. Numerical results

In this section, we conduct several simulation examples for both linear and nonlinear ITR boundaries to assess the performance of the proposed ROWL method. The examples with binary and multiple treatments, and with linear and nonlinear boundaries are carried out. We also apply ROWL method to an AIDS medical dataset to study its performance in practice. For comparisons, we implement several typical methods, the regression-based ITR learning (ℓ_1-PLS) in Qian and Murphy (2011), the standard SVM-type OWL (OWL_0) in Zhao et al. (2012) with extensions of one-versus-rest (OVR_OWL_0) and one-versus-one (OVO_OWL_0) for ITR problems with multiple treatments, the binary residual weighted learning (RWL) in Zhou et al. (2017) and its extensions for multicategory treatment settings (OVR_RWL and OVR_RWL), the LUM loss based ITR learning (LUM_MOML) proposed by Zhang et al. (2019), and the proposed ROWL method and its residual-based extension. We denote the original and residual-based ROWL methods by ROWL_0 and ROWL_1, respectively. For illustrations, we list the properties of these methods in terms of various perspectives in Table 1.

For simulated examples, we generate three independent datasets, the training, tuning and testing sets, and set the size of training and tuning sets be the same, and the size of testing set be 10 times as big as the training set. The training set is used to fit the model, the tuning set is used to find the best tuning parameters, and the test set is used to evaluate the fitted model performance. Let the scale parameter ρ of the robust loss $V(\cdot)$ vary in $\{0, 10^{-3}, 10^{-2}, \ldots, 10^4\}$, the regularization parameter λ vary in $\{10^{-3}, 10^{-2}, \ldots, 10^3\}$ for tuning, and we choose the best parameter through a grid search.

The evaluation value function is defined as $\mathbb{P}_n^*[\mathbb{1}(A = D(\boldsymbol{X}))R/\Pr(A)]/\mathbb{P}_n^*[1(A = D(\boldsymbol{X}))/\Pr(A)]$, where $\mathbb{P}_n^*$ denotes the empirical average of the testing dataset and $\Pr(A)$ is the probability of being assigned to the treatment A (Zhao et al. (2012)). The value function is regarded as a more comprehensive measure the difference between the estimated ITR and the true optimal ITR. We also record the misclassification errors to show the performances of all methods in terms of treatment assignments. We report the averages and standard deviations of these two measurements over 100 repetitions for all conducted settings.

5.1. *Linear boundary examples*

Example 1. We generate a simulated dataset in the following manner. First, we generate a 10-dimensional covariate vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$, consisting of independent $U[-1, 1]$ variables. Second, the treatment A is drawn from $\{-1, +1\}$ independently of $\boldsymbol{X}$ with equal probabilities. The outcome R is drawn from a normal distribution with mean $\mu = Q_0(\boldsymbol{X}) + \delta_0(\boldsymbol{X}) \cdot A$ and standard deviation 1, where Q_0 is the main effect between outcomes and clinical covariates, and $\delta_0(\boldsymbol{X}) \cdot A$ is the interaction effect between the treatment and clinical covariates. In fact, the treatment assignment rule is $\text{sign}(\delta_0(\cdot))$. To obtain linear decision boundaries, we

Table 2. Means and standard deviations (in parenthesis) of empirical value functions and misclassification errors evaluated on independent test set for the linear binary Example 1.

Methods		L_1-PLS		OWL_0		RWL		LUM_MOML		ROWL_0		ROWL_1	
n	*perc*	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc
100	0%	**2.987**	**0.017**	2.875	0.076	2.877	0.086	2.865	0.087	2.874	0.085	2.957	0.047
		(0.011)	(0.001)	(0.030)	(0.010)	(0.027)	(0.009)	(0.028)	(0.010)	(0.032)	(0.009)	(0.013)	(0.005)
	5%	**2.399**	**0.021**	2.221	0.095	2.273	0.093	2.247	0.092	2.232	0.102	2.366	0.056
		(0.010)	(0.002)	(0.043)	(0.013)	(0.030)	(0.010)	(0.036)	(0.011)	(0.044)	(0.012)	(0.013)	(0.005)
	10%	**3.157**	**0.033**	2.966	0.110	3.030	0.099	2.929	0.121	3.000	0.106	3.062	0.080
		(0.015)	(0.004)	(0.050)	(0.014)	(0.035)	(0.011)	(0.052)	(0.015)	(0.054)	(0.013)	(0.032)	(0.010)
300	0%	**2.916**	**0.008**	2.900	0.035	2.888	0.051	2.892	0.042	2.892	0.044	2.911	0.019
		(0.006)	(0.001)	(0.007)	(0.003)	(0.007)	(0.004)	(0.007)	(0.004)	(0.008)	(0.004)	(0.006)	(0.002)
	5%	**2.514**	**0.009**	2.484	0.040	2.475	0.058	2.479	0.047	2.473	0.055	2.504	0.027
		(0.006)	(0.001)	(0.015)	(0.005)	(0.009)	(0.005)	(0.015)	(0.005)	(0.009)	(0.006)	(0.007)	(0.003)
	10%	**2.741**	**0.013**	2.694	0.051	2.691	0.068	2.677	0.064	2.665	0.072	2.713	0.043
		(0.006)	(0.001)	(0.019)	(0.007)	(0.009)	(0.005)	(0.019)	(0.007)	(0.020)	(0.007)	(0.010)	(0.005)
1000	0%	**2.664**	**0.004**	2.659	0.024	2.647	0.046	2.658	0.026	2.656	0.031	2.661	0.014
		(0.003)	(0.000)	(0.003)	(0.001)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
	5%	**2.573**	**0.005**	2.566	0.027	2.552	0.050	2.565	0.028	2.561	0.034	2.570	0.016
		(0.004)	(0.000)	(0.004)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)
	10%	**3.115**	**0.007**	3.104	0.029	3.091	0.051	3.104	0.031	3.102	0.033	3.111	0.017
		(0.003)	(0.001)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.002)	(0.004)	(0.003)	(0.003)	(0.002)

consider the linear scenario with $Q_0(x) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$ and $\delta_0(x) = 1.8(0.3 - x_1 - x_2)$, and $\mu_i = Q_0(x_i) + \delta_0(x_i) \cdot a_i$ for instance (x_i, a_i).

To show the performance against possible outliers of all methods, we contaminate the outcomes in the training and tuning sets with a certain proportion, and denote the contamination percentage as *perc*. We contaminate the dataset by randomly selecting *perc* instances as outliers, and change the distributions of their rewards. For the outlier instance (x, a, r), we change the mean of outcome distribution as $\tilde{\mu} = Q_0(x) - \delta_0(x) \cdot a$, and draw a new $\tilde{r}$ from $N(\tilde{\mu}, 1)$, then the outcome contaminated point $(x, a, \tilde{r})$ is obtained. Denote the size of training set as n.

This simulated example is a typical ITR problem with binary treatments. Consider different scenarios with $n \in \{100, 300, 1000\}$ and $perc \in \{0\%(\text{no contamination}), 5\%, 10\%\}$. We report the results of sample means and standard deviations of the estimated value functions and the misclassification errors over 100 replications in Table 2. The standard deviations of these two measurements are presented in parenthesis, and the best performance for each scenario is in bold.

Example 2. We define three points $(c_1; c_2; c_3)$ of equal distances in $\mathbb{R}^{10}$ to represent the cluster centroids of the underlying true optimal treatments, where $c_1 = 5 \cdot (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \mathbf{0}_8)^{\mathsf{T}}$, $c_2 = 5 \cdot (\frac{\sqrt{3}-1}{2\sqrt{2}}, -\frac{\sqrt{3}+1}{2\sqrt{2}}, \mathbf{0}_8)^{\mathsf{T}}$ and $c_3 = 5 \cdot (-\frac{\sqrt{3}+1}{2\sqrt{2}}, \frac{\sqrt{3}-1}{2\sqrt{2}}, \mathbf{0}_8)^{\mathsf{T}}$. For each centroid c_j; $j = 1, 2, 3$, we generate its corresponding covariate vector x_i from a multivariate normal distribution $\boldsymbol{N}(c_j; \boldsymbol{I}_{10})$, where $\boldsymbol{I}_{10}$ represents a 10-dimensional identity matrix. The actually assigned treatment a_i follows a discrete uniform distribution $U\{1, 2, 3\}$. Based on the value of instance x_i and a_i, the outcome r_i is generated from a normal distribution $N(\mu(x_i, a_i, d_i); 1)$, where the mean function is $\mu(x_i, a_i, t_i) = 1 + \frac{1}{10}(\sum_{j=1}^{5} x_{i,j}^2 - \sum_{j=6}^{10} x_{i,j}^2) + 3 \cdot 1(a_i = d_i)$, and d_i represents the true underlying optimal treatment for x_i, which is determined by the cluster centroid. We also select *perc* instances as outliers, by changing the reward mean as $\tilde{\mu}_i = 1 + \frac{1}{10}(\sum_{j=1}^{5} x_{i,j}^2 - \sum_{j=6}^{10} x_{i,j}^2) - 3 \cdot 1(a_i = d_i)$ for the chosen contaminated instance (x_i, a_i, r_i), and sampling a new $\tilde{r}_i \sim N(\tilde{\mu}_i, 1)$ instead of r_i as an outlier. The training set is of size n.

Table 3. Means and standard deviations (in parenthesis) of empirical value functions and misclassification errors evaluated on an independent test for the linear Example 2 with $k = 3$.

Methods	$n = 300$						$n = 1000$					
	perc = 0%		*perc* = 5%		*perc* = 10%		*perc* = 0%		*perc* = 5%		*perc* = 10%	
	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc
L_1-PLS	6.049	0.149	6.179	0.107	5.822	0.230	6.133	0.123	6.289	0.072	6.137	0.121
	(0.081)	(0.027)	(0.071)	(0.023)	(0.090)	(0.030)	(0.074)	(0.025)	(0.060)	(0.020)	(0.076)	(0.025)
OVR_OWL$_0$	5.734	0.255	5.791	0.237	5.613	0.296	6.275	0.059	6.244	0.057	6.167	0.115
	(0.045)	(0.015)	(0.046)	(0.015)	(0.044)	(0.014)	(0.023)	(0.008)	(0.021)	(0.007)	(0.057)	(0.019)
OVO_OWL$_0$	6.009	0.162	5.866	0.262	5.716	0.262	6.290	0.036	6.281	0.057	6.262	0.081
	(0.045)	(0.015)	(0.053)	(0.018)	(0.054)	(0.018)	(0.021)	(0.007)	(0.026)	(0.008)	(0.031)	(0.010)
OVR_RWL	6.290	0.069	6.281	0.072	6.178	0.108	6.303	0.067	6.293	0.071	**6.409**	**0.030**
	(0.042)	(0.014)	(0.041)	(0.013)	(0.055)	(0.018)	(0.038)	(0.012)	(0.037)	(0.012)	(0.026)	(0.009)
OVO_RWL	6.185	0.105	6.189	0.102	6.168	0.111	**6.384**	**0.040**	**6.320**	**0.062**	6.289	0.071
	(0.052)	(0.017)	(0.048)	(0.015)	(0.056)	(0.019)	(0.033)	(0.011)	(0.039)	(0.013)	(0.042)	(0.014)
LUM_MOML	5.770	0.240	5.603	0.298	5.511	0.331	6.103	0.135	6.095	0.137	5.969	0.178
	(0.052)	(0.017)	(0.067)	(0.022)	(0.057)	(0.019)	(0.051)	(0.017)	(0.047)	(0.016)	(0.054)	(0.018)
ROWL$_0$	5.767	0.243	5.613	0.297	5.640	0.288	6.175	0.109	6.203	0.101	6.117	0.128
	(0.085)	(0.028)	(0.090)	(0.030)	(0.085)	(0.028)	(0.063)	(0.021)	(0.060)	(0.020)	(0.065)	(0.021)
ROWL$_1$	**6.322**	**0.058**	**6.320**	**0.061**	**6.201**	**0.102**	6.312	0.063	6.287	0.072	6.385	0.039
	(0.043)	(0.014)	(0.042)	(0.014)	(0.056)	(0.018)	(0.040)	(0.013)	(0.047)	(0.015)	(0.033)	(0.011)

To handle such an ITR example with 3 treatments, we implement two binary sequential schemes, i.e. one-versus-rest and one-versus-one, for the standard SVM-type OWL and binary RWL. We conduct several different scenarios with $n \in \{300, 1000\}$ and $perc \in \{0\%, 5\%, 10\%\}$. The other settings are the same as Example 1. We record the estimated value functions and misclassification errors, and summarize the sample means and standard deviations over 100 replications in Table 3.

From Table 2, we find that because of correct model specification, the ℓ_1-PLS method performs competitively compared with other methods. The ROWL methods give reasonable performance. When the sample size is large ($n = 300, 1000$), the residual-based ROWL performs very similarly to ℓ_1-PLS in terms of value functions. From Table 3, we can conclude that when n is smaller ($n = 300$), the residual-based method ROWL_1 outperforms other methods in terms of higher value functions and smaller misclassification errors among all settings. However, when the sample size increases, the RWL extensions perform better than our ROWL methods. From Tables 2 and 3, one can see that as the sample size increases, misclassification errors become smaller as expected in most settings.

5.2. *Nonlinear boundary examples*

Example 3. All the settings are the same as Example 1 except the specific setting of mean functions of outcomes. Here, we consider the nonlinear dependence with the explicit form $\mu(x, a) = Q_0(x) + \delta_0(x) \cdot a$, where $Q_0(x) = 1 + x_1^2 + x_2^2 + x_3^2 + 0.5x_4^2$ and $\delta_0(x) = 3.8(0.8 - x_1^2 - x_2^2)$ for $x \in \mathbb{R}^{10}$. The contaminated mean function is $\tilde{\mu}(x, a) = Q_0(x) - \delta_0(x) \cdot a$. This is a circle decision boundary example, where the patients inside the ring are assigned to one treatment, and another if outside the ring.

To achieve nonlinear learning, we adopt the Gaussian kernel $K(u_1, u_2) = \exp\left(\frac{-\|u_1 - u_2\|_2^2}{2\sigma^2}\right)$. For simplicity, we use the median of the between-class pairwise Euclidean distances of training inputs as estimated $\hat{\sigma}$ to avoid an extensive grid search of the tuning parameter σ (Brown et al. 2000; Liu and Yuan 2011; Wu and Liu 2007). The training dataset is of size n. We implement kernel learning for all compared methods in Example 1, and conduct several different scenarios with $n \in \{100, 300\}$ and $perc \in \{0\%, 5\%, 10\%\}$. The corresponding results are summarized in Table 4.

Table 4. Means and standard deviations (in parenthesis) of empirical value functions and misclassification errors evaluated on an independent test set for the binary nonlinear Example 3.

Methods	$n = 100$						$n = 300$					
	perc = 0%		*perc* = 5%		*perc* = 10%		*perc* = 0%		*perc* = 5%		*perc* = 10%	
	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc
L_1-PLS	2.673	0.357	3.012	0.356	2.834	0.358	2.572	0.358	2.815	0.352	2.616	0.352
	(0.028)	(0.007)	(0.030)	(0.006)	(0.037)	(0.008)	(0.018)	(0.003)	(0.027)	(0.005)	(0.022)	(0.004)
OWL_0	3.191	0.256	3.390	0.283	3.186	0.291	3.781	0.121	3.853	0.157	3.581	0.167
	(0.057)	(0.011)	(0.057)	(0.011)	(0.066)	(0.013)	(0.038)	(0.008)	(0.043)	(0.009)	(0.054)	(0.011)
RWL	2.640	0.353	2.991	0.353	2.779	0.358	2.660	0.343	2.866	0.343	2.700	0.336
	(0.024)	(0.004)	(0.025)	(0.004)	(0.026)	(0.006)	(0.033)	(0.006)	(0.032)	(0.006)	(0.038)	(0.008)
LUM_MOML	3.086	0.274	3.354	0.289	3.212	0.285	3.835	0.113	3.900	0.149	3.574	0.164
	(0.058)	(0.011)	(0.057)	(0.011)	(0.061)	(0.011)	(0.030)	(0.007)	(0.039)	(0.008)	(0.054)	(0.011)
ROWL_0	3.270	0.247	3.562	0.261	3.298	0.280	3.817	0.117	3.885	0.153	3.644	0.150
	(0.066)	(0.014)	(0.961)	(0.013)	(0.076)	(0.013)	(0.031)	(0.008)	(0.038)	(0.008)	(0.049)	(0.010)
ROWL_1	**3.327**	**0.230**	**3.656**	**0.237**	**3.434**	**0.246**	**3.947**	**0.084**	**4.058**	**0.109**	**3.761**	**0.125**
	(0.056)	(0.012)	(0.087)	(0.013)	(0.052)	(0.010)	(0.017)	(0.005)	(0.025)	(0.006)	(0.036)	(0.009)

Example 4. This is a four class example with nonlinear decision boundaries. The covariate vector is $\boldsymbol{X} = (X_1, \ldots, X_{10}) \in \mathbb{R}^{10}$ with each component following the uniform distribution $U[-1, 1]$. The optimal treatment d_i is determined by the sign of two underlying nonlinear functions $g_1(x) = x_1^2 + x_2^2 + \exp(0.5x_3)$ and $g_2(x) = x_4^2 - x_5^3 - x_6$. In particular, we set $d_i = d(x_i) = 1 + [\text{sign}(g_1(x_i) - m_1)]_+ + 2[\text{sign}(g_2(x_i) - m_2)]_+$, where $[u]_+ = \max(u, 0)$, and m_1 and m_2 are the medians of g_1 and g_2, respectively. The actually assigned treatment follows a discrete uniform distribution $U\{1, 2, 3, 4\}$. The actual outcome r_i is drawn from a normal distribution $N(\mu(x_i, a_i, d_i), 1)$, where $\mu(x_i, a_i, d_i) = (\sum_{j=1}^{5} x_{i,j} - \sum_{j=6}^{10} x_{i,j} - 1) + 5 \cdot 1(a_i = d_i)$. However, the outcome of contaminated sample is also from a normal distribution $N(\tilde{\mu}(\cdot), 1)$, where $\tilde{\mu}(x_i, a_i, d_i) = (\sum_{j=1}^{5} x_{i,j} - \sum_{j=6}^{10} x_{i,j} - 1) - 5 \cdot 1(a_i = d_i)$.

The proportion of the contaminated samples *perc* ranges in $\{0\%, 5\%, 10\%\}$, and the size of training set varies in $n \in \{300, 1000\}$. We employ the Gaussian kernel the same as in Example 3, and implement kernel learning for all compared methods in Example 2. The other settings are the same as Example 2, and we summarize the corresponding results in Table 5.

From Tables 4 and 5, similar to the linear examples, the results show that residual-based ROWL_1 enjoys better performance than other methods, and its robustness against outliers is very competitive. Increasing the sample size can make ROWL learning more accurate with larger estimated value functions and smaller misclassification errors. Especially for Example 4, when the percentage of contamination gets higher, the evaluated value functions for most methods become smaller, and the misclassification errors get larger.

5.3. *Real data application*

To show the performance of ROWL and its residual weighted extensions in practice, we apply them to a real medical data from AIDS Clinical Trials Group Protocol 175, which consists of 2139 subjects infected with the human immunodeficiency virus. This dataset was previously studied by Fan et al. (2017). There are four different treatments groups: zidovudine (ZDV) monotherapy, ZDV plus didanosine (ddI), ZDV plus zalcitabine (Zal) and ddI monotherapy, denoted as $A = 1, 2, 3, 4$, respectively. The sizes of these 4 treatment groups are $532, 522, 524, 561$, respectively. Therefore, it is a balanced ITR learning problem with 4 treatments.

The difference between early stage CD4 + T (cells/mm^3) cell amount and the baseline CD4 + T prior to trial can be set as the clinical outcome R. We choose 12 related covariates as prognostic

Table 5. Means and standard deviations (in parenthesis) of empirical value functions and misclassification errors evaluated on an independent test set for the nonlinear Example 4 with $k = 4$.

Methods	$n = 300$						$n = 1000$					
	perc = 0%		*perc* = 5%		*perc* = 10%		*perc* = 0%		*perc* = 5%		*perc* = 10%	
	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc	Value	Misc
L_1-PLS	2.391	0.319	2.187	0.361	1.950	0.409	2.944	0.211	2.800	0.240	2.665	0.267
	(0.019)	(0.004)	(0.028)	(0.005)	(0.029)	(0.006)	(0.009)	(0.002)	(0.012)	(0.002)	(0.015)	(0.003)
OVR_OWL$_0$	2.287	0.341	2.169	0.366	**1.980**	**0.401**	2.875	0.225	2.736	0.253	2.587	0.282
	(0.019)	(0.003)	(0.023)	(0.004)	(0.022)	(0.004)	(0.010)	(0.002)	(0.012)	(0.002)	(0.013)	(0.003)
OVO_OWL$_0$	2.117	0.374	2.045	0.392	1.881	0.424	2.650	0.270	2.562	0.288	2.503	0.300
	(0.024)	(0.004)	(0.027)	(0.005)	(0.023)	(0.004)	(0.014)	(0.003)	(0.016)	(0.003)	(0.020)	(0.004)
OVR_RWL	1.398	0.484	1.337	0.470	1.198	0.441	1.644	0.471	1.636	0.473	1.489	0.502
	(0.043)	(0.008)	(0.043)	(0.008)	(0.043)	(0.009)	(0.038)	(0.007)	(0.045)	(0.009)	(0.045)	(0.009)
OVO_RWL	2.135	0.368	2.014	0.396	1.821	0.434	2.814	0.236	2.753	0.249	2.617	0.276
	(0.035)	(0.007)	(0.033)	(0.006)	(0.035)	(0.007)	(0.018)	(0.004)	(0.019)	(0.004)	(0.028)	(0.005)
LUM_MOML	2.330	0.331	2.161	0.367	1.969	0.406	2.939	0.212	2.786	0.243	2.649	0.271
	(0.019)	(0.003)	(0.025)	(0.005)	(0.027)	(0.005)	(0.011)	(0.002)	(0.012)	(0.002)	(0.014)	(0.003)
ROWL$_0$	2.326	0.331	2.145	0.370	1.939	0.412	2.913	0.217	2.755	0.248	2.597	0.281
	(0.020)	(0.004)	(0.027)	(0.005)	(0.028)	(0.005)	(0.011)	(0.002)	(0.014)	(0.003)	(0.016)	(0.003)
ROWL$_1$	**2.394**	**0.318**	**2.221**	**0.356**	1.978	0.403	**2.980**	**0.204**	**2.812**	**0.237**	**2.667**	**0.267**
	(0.019)	(0.004)	(0.026)	(0.005)	(0.026)	(0.005)	(0.009)	(0.002)	(0.013)	(0.002)	(0.018)	(0.004)

variables X, including 5 continuous variables (age (years), weight (kg), Karnofsky score (scale of 0∼100), CD4 cell count at baseline and CD8 cell count (cells/mm^3) at baseline), and 7 binary variables (gender (0 = female, 1 = male), race (0 = white, 1 = non-white), homosexual activity (0 = no, 1 = yes), history of intravenous drug use (0 = no, 1 = yes), symptomatic status (0 = asymptomatic, 1 = symptomatic), antiretroviral history (0 = naive, 1 = experienced) and hemophilia (0 = no, 1 = yes)). For simplicity, we standardize the continuous variables with sample mean 0 and standard deviation 1. We observe that the variation of the outcome is very large. To obtain stable performance, we also normalize the rewards as well.

To show the robustness of the proposed methods, we contaminate the dataset with outliers in a different way from the simulated examples. We contaminate the treatment set with a certain percentage ($perc = 0, 5\%, 10\%, 15\%$ and 20%). We choose *perc* of all samples as outliers randomly, for the chosen instance (x_i, a_i, r_i), and assign the corresponding treatment to one of the remaining 3 treatments $\{1, 2, 3, 4\}\backslash a_i$ with equal probabilities as its new treatment $\tilde{a}_i$.

We randomly select 500 subjects as the training set, 500 as the tuning set, and the rest 1139 subjects as the testing set. For simplicity, we conduct linear learning for all the comparing methods. The means and corresponding standard deviations of the estimated value functions over 100 replications are reported in Table 6.

Table 6 shows that the proposed residual-based ROWL_1 performs slightly better than other methods, and it is more robust against outliers. This conclusion here is consistent with the simulation examples. However, the impact of contamination parameter *perc* is not clear for all

Table 6. Means and standard deviations (in parenthesis) of empirical value functions evaluated on an independent test set for the real AIDS dataset with $k = 4$.

Methods	$n = 500$									
	perc = 0%		*perc* = 5%		*perc* = 10%		*perc* = 15%		*perc* = 20%	
L_1-PLS	0.2214	(0.0069)	0.2134	(0.0063)	0.2235	(0.0060)	0.2225	(0.0064)	0.2180	(0.0066)
OVR_OWL$_0$	0.2250	(0.0086)	0.2313	(0.0057)	0.2342	(0.0067)	0.2296	(0.0066)	0.2236	(0.0080)
OVO_OWL$_0$	0.1904	(0.0056)	0.1845	(0.0053)	0.1528	(0.0067)	0.1377	(0.0067)	0.1833	(0.0080)
OVR_RWL	0.1821	(0.0111)	0.1952	(0.0103)	0.1895	(0.0104)	0.1921	(0.0102)	0.1764	(0.0105)
OVO_RWL	0.2624	(0.0058)	0.2498	(0.0063)	0.2452	(0.0065)	0.2420	(0.0076)	0.2476	(0.0074)
LUM_MOML	0.2687	(0.0059)	0.2543	(0.0054)	0.2498	(0.0052)	0.2488	(0.0066)	0.2479	(0.0068)
ROWL$_0$	0.2649	(0.0057)	0.2540	(0.0055)	0.2491	(0.0052)	0.2441	(0.0066)	0.2469	(0.0072)
ROWL$_1$	**0.2701**	(0.0055)	**0.2556**	(0.0049)	**0.2516**	(0.0052)	**0.2499**	(0.0061)	**0.2491**	(0.0070)

methods. One possible reason is that the way of contamination is different from the simulated examples. In particular, the real dataset has the contaminated treatment, while the simulated examples have the contaminated outcomes. In fact, these two kinds of contamination may happen.

Based on the intensive numerical experiments, we can conclude that the proposed residual weighted ROWL_1 is more stable to outliers, and we suggest the practitioners to use residual-based ROWL_1 to estimate ITRs with multiple treatments.

6. Conclusion

In this paper, we propose a family of robust loss functions to estimate the optimal ITRs with binary and multiple treatments. The robust loss function is upper bounded, and helps to obtain robust classifiers. The rich robust loss family covers both soft and hard classifiers through a scale parameter and connects several well-known classifiers. Following the angle-based framework, we can naturally extend binary treatment learning to multiple treatment settings. Based on the outcome weighted classification, we incorporate the angle-based structure and the new robust loss into ROWL to estimate robust ITR, which can directly handle ITR problems with both positive and negative outcomes. Under some mild conditions, ROWL enjoys Fisher consistency, and can provide estimation for the reward ratios of all pairs of treatments. Moreover, we develop an efficient DCA to solve the nonconvex minimization problem for ROWL. The results of simulated examples and real medical data indicate that the residual-based ROWL performs competitively in most cases. One possible future direction is to develop robust dynamic treatment regime techniques using the proposed robust loss functions.

Acknowledgments

The authors would like to thank the guest editor Jean Pan and reviewers, whose helpful comments and suggestions led to a much improved presentation. Liu's research was partially supported by NSF grants IIS1632951 and DMS1821231, and NIH grants R01GM126550 and P01CA142538.

Funding

This work was supported by the Division of Information and Intelligent Systems [1632951];Division of Mathematical Sciences [1821231];National Cancer Institute [142538];National Institute of General Medical Sciences [126550].

ORCID

Yufeng Liu http://orcid.org/0000-0002-1686-0545

References

An, L. T. H., and P. D. Tao. 1997. Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of Global Optimization* 11 (3):253–285.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101 (473):138–156.

Boyd, S., and L. Vandenberghe. 2004. *Convex optimization*. New York, NY: Cambridge University Press.

Brown, M. P., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* 97 (1):262–267.

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

Ellsworth, R. E., D. J. Decewicz, C. D. Shriver, and D. L. Ellsworth. 2010. Breast cancer in the personal genomics era. *Current Genomics* 11 (3):146–161.

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

Fu, S., S. Zhang, and Y. Liu. 2018. Adaptively weighted large-margin angle-based classifiers. *Journal of Multivariate Analysis* 166:282–299.

Kimeldorf, G., and G. Wahba. 1971. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* 33 (1):82–95.

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

Lee, Y., Y. Lin, and G. Wahba. 2004. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* 99 (465):67–81.

Lin, Y. 2002. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery* 6 (3):259–275.

Liu, Y. 2007. Fisher consistency of multicategory support vector machines. In *Artificial intelligence and statistics*, ed. M. Meila and X. Shen, 291–298. Proceedings of Machine Learning Research. http://proceedings.mlr.press/v2/

Liu, Y., X. Shen, and H. Doss. 2005. Multicategory ψ-learning and support vector machine: Computational tools. *Journal of Computational and Graphical Statistics* 14 (1):219–236.

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

Liu, Y., and M. Yuan. 2011. Reinforced multicategory support vector machines. *Journal of Computational and Graphical Statistics* 20 (4):901–919.

Liu, Y., H. H. Zhang, and Y. Wu. 2011. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association* 106 (493):166–177.

Mancinelli, L., M. Cronin, and W. Sadée. 2000. Pharmacogenomics: The promise of personalized medicine. *The AAPS Journal* 2 (1):29–41.

Nocedal, J. 1980. Updating quasi-newton matrices with limited storage. *Mathematics of Computation* 35 (151):773–782.

Nocedal, J., and S. J. Wright. 2006. *Sequential quadratic programming*. New York, NY: Springer.

Qian, M., and S. A. Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39 (2):1180–1210.

Tian, L., A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. 2014. 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.

Wahba, G. 1990. *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wang, J., X. Shen, and Y. Liu. 2008. Probability estimation for large-margin classifiers. *Biometrika* 95 (1):149–167.

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

Wu, Y., and Y. Liu. 2013. Adaptively weighted large margin classifiers. *Journal of Computational and Graphical Statistics* 22 (2):416–432.

Xiao, W., H. H. Zhang, and W. Lu. 2019. Robust regression for optimal individualized treatment rules. *Statistics in Medicine* 38 (11):2059–2073.

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

Zhang, C., J. Chen, H. Fu, X. He, Y. Zhao, and Y. Liu (2019). Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica Sinica*, to appear.

Zhang, C., and Y. Liu. 2013. Multicategory large-margin unified machines. *Journal of Machine Learning Research* 14 (1):1349–1386.

Zhang, C., and Y. Liu. 2014. Multicategory angle-based large-margin classification. *Biometrika* 101 (3):625–640.

Zhang, C., M. Pham, S. Fu, and Y. Liu. 2018. Robust multicategory support vector machines using difference convex algorithm. *Mathematical Programming* 169 (1):277–305.

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

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

Appendix: Technical proofs

Proof of Theorem 1. To prove the theorem, we need the following lemma, of which the proof can be found in Zhang and Liu (2014).

Lemma 1 (Zhang and Liu 2014, **Lemma 1)** *Suppose we have an arbitrary* $f \in \mathbb{R}^{k-1}$. *For any* $u, v \in \{1, \ldots, k\}$ *such that* $u \neq v$, *define* $T_{u,v} = \mathcal{W}_u - \mathcal{W}_v$. *For any scalar* $z \in \mathbb{R}$, $\langle (f + zT_{u,v}), \mathcal{W}_w \rangle = \langle f, \mathcal{W}_w \rangle$, *where* $w \in \{1, \ldots, k\}$ *and* $w \neq u, v$. *Furthermore, we have that* $\langle (f + zT_{u,v}), \mathcal{W}_u \rangle - \langle f, \mathcal{W}_u \rangle = -\langle (f + zT_{v,u}), \mathcal{W}_v \rangle + \langle f, \mathcal{W}_v \rangle$.

Recall the definition of the conditional expected loss $S(x)$ in (3.1), and the minimization problem with respect to the decision function f is equivalent to minimizing

$$\mathcal{L}(f|\boldsymbol{X} = x) = \sum_{j=1}^{k} [R_j^+ \ell(\langle f(x), \mathcal{W}_j \rangle) - R_j^- \ell(-\langle f(x), \mathcal{W}_j \rangle)],$$

where R_j^+ and R_j^- are defined in Section 3.1. Assume the theoretical minimizer of $\mathcal{L}(f|\boldsymbol{X} = x)$ is $f^*(x)$, i.e. $f^*(x) = \arg\min_f \mathcal{L}(f|X = x)$. For simple notations, we drop the dependence of $\mathcal{L}(f)$ and $\langle f, \mathcal{W}_j \rangle$ on x when there is no ambiguity. Then, the objective function becomes

$$\mathcal{L}(\boldsymbol{f}) = \sum_{j=1}^{k} [R_j^+ \ell(\langle \boldsymbol{f}, \mathcal{W}_j \rangle) - R_j^- \ell(-\langle \boldsymbol{f}, \mathcal{W}_j \rangle)]. \tag{A.1}$$

The property of ℓ implies that ℓ is strictly decreasing. Without loss of generality, let the treatment 1 be the best one. Then we need to prove that $\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle$ is the largest, as stated in the following fact.

Fact 1 *If* $R_1 > R_j, \forall j \neq 1$ *and Assumption 1 holds, then* $\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle > \langle \boldsymbol{f}^*, \mathcal{W}_j \rangle$.

Proof. We prove it by contradiction. Assume there exists a $j_0 \neq 1$, such that $\langle f^*, \mathcal{W}_1 \rangle < \langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle$. According to the monotone property of ℓ, we have that $\ell(\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle) > \ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle)$ and $\ell(-\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle) < \ell(-\langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle)$. From Assumption 1, we have $R_1^+ \geq R_j^+$ and $R_1^- \geq R_j^-$ for any $j \neq 1$, and the equalities cannot hold simultaneously. Then by Lemma 1, one can find a new $\boldsymbol{f}^{**}$, which satisfies

$$\langle \boldsymbol{f}^{**}, \mathcal{W}_1 \rangle = \langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_{j_0} \rangle = \langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_j \rangle = \langle \boldsymbol{f}^*, \mathcal{W}_j \rangle, \forall j \neq 1, j_0.$$

Then we have $\mathcal{L}(\boldsymbol{f}^{**}) - \mathcal{L}(\boldsymbol{f}^*) = (R_1^+ - R_{j_0}^+)[\ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle) - \ell(\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle)] + (R_1^- - R_{j_0}^-)[\ell(-\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle) - \ell(-\langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle)] < 0$, and this contradicts with the definition of $\boldsymbol{f}^*$.

Suppose there exists a $j_0 \neq 1$, such that $\langle \boldsymbol{f}^*, \mathcal{W}_1 \rangle = \langle \boldsymbol{f}^*, \mathcal{W}_{j_0} \rangle = s_0$. By Lemma 1, for any small $\varepsilon > 0$, one can construct a new $\tilde{f}$ with

$$\langle \tilde{f}, \mathcal{W}_1 \rangle = s_0 + \varepsilon, \langle \tilde{f}, \mathcal{W}_{j_0} \rangle = s_0 - \varepsilon, \langle \tilde{f}, \mathcal{W}_j \rangle = \langle \boldsymbol{f}^*, \mathcal{W}_j \rangle, \forall j \neq 1, j_0.$$

Then $\mathcal{L}(\tilde{f}) - \mathcal{L}(\boldsymbol{f}^*) = [(R_1^+ - R_{j_0}^+)\ell'(s_0) + (R_1^- - R_{j_0}^-)\ell'(-s_0)]\varepsilon + o(\varepsilon) < 0$, which is a contradiction. □

By the above fact, the proof of Theorem 1 is completed. ∎

Proof of Theorem 2. Consider the special loss $\ell(u) = \min(e^{-u}, 1)$, which is the non-smooth truncated exponential loss, the proof about fisher consistency becomes more complicated than Theorem 1. We divide the proof of Theorem 2 into several basic facts as follows.

Fact 2. *If* $R_1 > R_j, \forall j \neq 1$ *and Assumption 1 holds, then* $\boldsymbol{f}^* \neq \boldsymbol{0}$.

Proof. The conclusion says that $\boldsymbol{0}$ is not minimizer of $\mathcal{L}(f)$ in (A.1). We prove it by contradiction. Assume $\boldsymbol{f}^* = \boldsymbol{0}$, then we have $\mathcal{L}(\boldsymbol{0}) = \sum_{j=1}^{k} (R_j^+ - R_j^-)$. By Lemma 1, for any small $\varepsilon > 0$, one can find a $\boldsymbol{f}^{**}$ such that

$$\langle \boldsymbol{f}^{**}, \mathcal{W}_1\rangle = \varepsilon > 0, \langle \boldsymbol{f}^{**}, \mathcal{W}_2\rangle = -\varepsilon < 0, \langle \boldsymbol{f}^{**}, \mathcal{W}_j\rangle = 0, \forall j \neq 1, 2.$$

Hence, $\mathcal{L}(\boldsymbol{f}^{**}) = R_1^+\ell(\varepsilon) - R_1^- + R_2^+ - R_2^-\ell(\varepsilon) + \sum_{j=3}^k (R_j^+ - R_j^-)$. Due to $R_1^+ > R_1^- \geq R_2^-$ and $\ell(\varepsilon) < \ell(0)$, then we have $\mathcal{L}(\boldsymbol{f}^{**}) - \mathcal{L}(\mathbf{0}) = (R_1^+ - R_2^-)(\ell(\varepsilon) - 1) < 0$. That is a contradiction. □

Fact 2 indicates that $\boldsymbol{f}^* \neq \mathbf{0}$, then there inner product $\langle \boldsymbol{f}^*, \mathcal{W}_j\rangle$s may have different signs. In other words, there exists at least one j such that $\langle \boldsymbol{f}^*, \mathcal{W}_j\rangle > 0$. If the treatment 1 is the best one, we need to prove that $\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle$ is the largest, as shown in the following facts.

Fact 3. *If $R_1 > R_j, \forall j \neq 1$ and Assumption 1 holds, then $\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle > 0$.*

Proof. We prove it by contradiction. If there exists a $j_0 \neq 1$, such that $\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle \leq 0 < \langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle$, then we have $\ell(\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle) > \ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle)$ and $\ell(-\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle) \leq \ell(-\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle)$. By Lemma 1, one can find a new $\boldsymbol{f}^{**}$, which satisfies

$$\langle \boldsymbol{f}^{**}, \mathcal{W}_1\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_{j_0}\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_1\rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_j\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_j\rangle, \forall j \neq 1, j_0.$$

Then we have $\mathcal{L}(\boldsymbol{f}^{**}) - \mathcal{L}(\boldsymbol{f}^*) = (R_1^+ - R_{j_0}^+)[\ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle) - \ell(\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle)] + (R_1^- - R_{j_0}^-)[\ell(-\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle) - \ell(-\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle)] < 0$, which contradicts with the optimality of $\boldsymbol{f}^*$. □

Fact 4. *If $R_1 > R_j, \forall j \neq 1$ and Assumption 1 holds, then $\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle > \langle \boldsymbol{f}^*, \mathcal{W}_j\rangle$.*

Proof. We prove it by contradiction. If there exists a $j_0 \neq 1$, such that $0 < \langle \boldsymbol{f}^*, \mathcal{W}_1\rangle < \langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle$, then we have $\ell(\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle) > \ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle)$ and $\ell(-\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle) = \ell(-\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle) = 1$. By Lemma 1, one can find a new $\boldsymbol{f}^{**}$, which satisfies

$$\langle \boldsymbol{f}^{**}, \mathcal{W}_1\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_{j_0}\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_1\rangle, \langle \boldsymbol{f}^{**}, \mathcal{W}_j\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_j\rangle, \forall j \neq 1, j_0.$$

Then we have $\mathcal{L}(\boldsymbol{f}^{**}) - \mathcal{L}(\boldsymbol{f}^*) = (R_1^+ - R_{j_0}^+)[\ell(\langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle) - \ell(\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle)] < 0$, which contradicts with the optimality of $\boldsymbol{f}^*$.

Suppose there exists a $j_0 \neq 1$, such that $\langle \boldsymbol{f}^*, \mathcal{W}_1\rangle = \langle \boldsymbol{f}^*, \mathcal{W}_{j_0}\rangle = s_0 > 0$. For any small $\varepsilon > 0$, one can construct a new $\tilde{f}$ with

$$\langle \tilde{f}, \mathcal{W}_1\rangle = s_0 + \varepsilon > 0, \langle \tilde{f}, \mathcal{W}_{j_0}\rangle = s_0 - \varepsilon > 0, \langle \tilde{f}, \mathcal{W}_j\rangle = \langle \tilde{\boldsymbol{f}}^*, \mathcal{W}_j\rangle, \forall j \neq 1, j_0.$$

Then $\mathcal{L}(\tilde{f}) - \mathcal{L}(\boldsymbol{f}^*) = (R_1^+ - R_{j_0}^+)\ell'(s_0)\varepsilon + o(\varepsilon) < 0$, which is a contradiction. □

Combining the above three facts, we finish the proof of Theorem 2. □

Proof of Theorem 3. When there is no negative rewards, then $R_j \geq 0$ for all $j = 1, \ldots, k$. Consider a general smooth loss function $\ell(\cdot)$, then the optimization problem becomes

$$\min_{f \in \mathcal{F}} \mathcal{L}(f) = \min_{f \in \mathcal{F}} \sum_{j=1}^k R_j \ell(\langle f, \mathcal{W}_j\rangle). \tag{A.2}$$

Then the first-order condition for the optimization of (A.2) is

$$\frac{\partial \mathcal{L}}{\partial f} = \sum_{j=1}^k R_j \ell'(\langle f, \mathcal{W}_j\rangle) W_j = \mathbf{0}_{k-1}. \tag{A.3}$$

It is clear that the theoretical minimizer $\boldsymbol{f}^*$ is the solution of (A.3). Note that $\sum_{j=1}^k \mathcal{W}_j = \mathbf{0}_{k-1}$ and arbitrary $k-1$ of $\{\mathcal{W}_i; i = 1, \ldots, k\}$ are linearly independent. Hence one can conclude that $R_i\ell'(\langle \boldsymbol{f}^*, \mathcal{W}_i\rangle) = R_j\ell'(\boldsymbol{f}^*, \mathcal{W}_j\rangle)$ for $i \neq j$. Therefore, we can obtain the reward ratios accordingly. ∎