The Annals of Statistics
2024, Vol. 52, No. 2, 679–707
https://doi.org/10.1214/24-AOS2363

FINDING THE OPTIMAL DYNAMIC TREATMENT REGIMES USING SMOOTH FISHER CONSISTENT SURROGATE LOSS

BY NILANJANA LAHA[1,a], AARON SONABEND-W[2,b], RAJARSHI MUKHERJEE[2,c] AND TIANXI CAI[2,d]

[1]*Department of Statistics, Texas A&M University,* [a]*nlaha@tamu.edu*
[2]*Department of Biostatistics, Harvard University,* [b]*asonabend@gmail.com,* [c]*ram521@mail.harvard.edu,* [d]*tcai@hsph.harvard.edu*

Large health care data repositories such as electronic health records (EHR) open new opportunities to derive individualized treatment strategies for complicated diseases such as sepsis. In this paper, we consider the problem of estimating sequential treatment rules tailored to a patient's individual characteristics, often referred to as dynamic treatment regimes (DTRs). Our main objective is to find the optimal DTR that maximizes a discontinuous value function through direct maximization of Fisher consistent surrogate loss functions. In this regard, we demonstrate that a large class of concave surrogates fails to be Fisher consistent—a behavior that differs from the classical binary classification problems. We further characterize a nonconcave family of Fisher consistent smooth surrogate functions, which is amenable to gradient-descent type optimization algorithms. Compared to the existing direct search approach under the support vector machine framework (*J. Amer. Statist. Assoc.* **110** (2015) 583–598), our proposed DTR estimation via surrogate loss optimization (DTRESLO) method is more computationally scalable to large sample sizes and allows for broader functional classes for treatment policies. We establish theoretical properties for our proposed DTR estimator and obtain a sharp upper bound on the regret corresponding to our DTRESLO method. The finite sample performance of our proposed estimator is evaluated through extensive simulations. We also illustrate the working principles and benefits of our method for estimating an optimal DTR for treating sepsis using EHR data from sepsis patients admitted to intensive care units.

1. Introduction. Due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with biorepositories and other research registries, integrated large data sets have become available for *real world evidence* based precision medicine studies. These rich EHR data capture heterogeneity in response to treatment over time and across patients, thereby offering unique opportunities to optimize treatment strategies for individual patients over time. Sequential treatment decisions tailored to patients' individual characteristics at given decision time points are often referred to as dynamic treatment regimes (DTRs) in the statistical literature and reinforcement learning (RL) in the machine learning literature. An optimal DTR can be defined as the sequential treatment assignment rule that maximizes the expected counterfactual outcome, often referred to as the value function in the DTR literature.

To estimate the optimal DTR, the most traditional approaches rely on modeling the data-distribution or part of the data-distribution (Xu et al. (2016), Zajonc (2012)). The most popular among the latter class are the regression-based methods, including Q-learning, A-learning and marginal structural mean models (Murphy (2003), Orellana, Rotnitzky and Robins (2010), Robins (2004), Schulte et al. (2014), Watkins (1989)). The regression-based methods, especially Q-learning, offers the flexibility necessary for extension to a variety of

Received May 2022; revised August 2023.
Key words and phrases. Dynamic treatment regimes, classification, empirical risk minimization, nonconvex optimization.

settings including, but not limited to, semisupervised setting (Sonabend-W et al. (2023)), interactive model-building (Laber et al. (2014)), discrete outcomes or utilities (Moodie, Dean and Sun (2014)), etc. However, the underlying models in the regression-based approaches are often high-dimensional, and susceptible to mis-specification due to the sequential nature of the problem (Murphy, van der Laan and Robins (2001)). Although A-learning and marginal structural mean models are more robust to model misspecification, they still require the contrast of Q-functions to be correctly specified (cf. Schulte et al. (2014)). These limitations of the regression-based methods led the conception of the classification-based direct search methods, which in contrast, directly targets the counterfactual value function.

The classification-based approaches essentially rely on the representation of the counterfactual value function through importance sampling (Murphy, van der Laan and Robins (2001)), whose maximization can be framed as a classification problem with respect to the zero-one loss function (cf. Chen, Zeng and Kosorok (2016), Chen et al. (2017), Cui and Tchetgen Tchetgen (2021), Song et al. (2015), Zhao et al. (2012, 2015), Zhao (2016) and the references therein). The resulting objective function is not amenable to efficient optimization owing to the discontinuity of the zero-one loss. Therefore, following contemporary classification literature (cf. Bartlett, Jordan and McAuliffe (2006), Lin (2004)) the direct search methods aim to replace the zero-one loss with alternative smoother fisher consistent surrogate loss functions to facilitate efficient classification methods. The paradigm shift of estimating DTRs by finding classification rules is a powerful idea. Some authors indicate that existing direct search methods outperform regression-based counterparts when the number of stages is small (Kosorok and Laber (2019), Luedtke and van der Laan (2016)).

Although initially developed for the one-stage case, direct search method was introduced to the multistage DTR by the novel work of Zhao et al. (2015). Currently, it has two mainstream approaches. The first approach performs binary classification stagewise in a backward fashion (cf. BOWL method o Jiang et al. (2019), Zhao et al. (2015)). However, at stage t, this approach can only use those observations whose treatment assignment matches the optimal treatment stage $t + 1$ onward. As a result, the effective sample size of the initial stages dwindles rapidly, which can be problematic during practical implementation (Kallus (2019), Kosorok and Laber (2019)). The other approach builds on a simultaneous optimization method, which utilizes the whole data set for estimating each treatment assignment (simultaneous outcome weighted learning (SOWL), Zhao et al. (2015)). While it does not share the limitation of the BOWL-type approaches, this approach hinges on a sequential weighted classification problem, which is complicated by the dependent nature of the DTR setting. Zhao et al. (2015) solves this classification using a bivariate hinge-loss type surrogate. Although the idea behind simultaneous optimization is powerful, the implementation via nonsmooth hinge-loss surrogate leads to a number of issues, scalability being one of them; see Section C in the Supplementary Material (Laha et al. (2024)) for more details. It is natural to ask whether the hinge loss can be replaced by other surrogates. However, the answer is not immediate because unlike BOWL, the simultaneous classification does not yield to the binary classification theory on surrogate losses (Bartlett, Jordan and McAuliffe (2006)). Although multicategory and multilabel classifications have apparent resemblance with this classification problem, as we will see, they have fundamental differences. This gives rise to the need for a unified study of fisher consistent surrogate losses under the DTR setting. Our paper is the first step toward that end.

For the ease of presentation, we focus on $k = 2$ stage DTRs associated with two time points in this paper. However, the main methodology easily extends to general k-stage settings when $k > 2$. Similar to most current works in direct search methods, we consider only a binary treatment indicator, which is an important practical case (Laber and Davidian (2017)). Direct search with multilevel treatments would require substantially different techniques, and is out of the scope of the present paper.

1.1. *Main contributions*. In the sequel, we will refer to the classification problem resulting from the simultaneous optimization approach as "the DTR classification problem" for brevity. We will refer to our approach of achieving optimal *DTR estimation via surrogate loss optimization* as DTRESLO.

1.1.0.1. *Concave losses*. In Theorem 1, we establish that the above-bounded smooth concave surrogates fail to be Fisher consistent in the DTR context. The failure is not restricted to only smooth concave surrogates since our Theorem 2 also shows that nonsmooth hinge loss also fails to be Fisher consistent. Furthermore, we have not encountered any concave loss function that is Fisher consistent in the DTR context. Consequently, our findings naturally prompt the question of whether any concave loss function can indeed achieve Fisher consistency for this problem.

1.1.0.2. *A class of Fisher consistent surrogates for DTR estimation*. Given the limited promise of concave surrogate losses for this problem, we directed our attention toward the realm of nonconcave surrogates. We introduce a class of nonconcave Fisher consistent surrogate losses (see Theorem 3), which are amenable to efficient gradient-based algorithms, such as stochastic gradient descent. This facilitates the utilization of fast and scalable optimization methods. Since the resulting optimizing problem is nonconcave, convergence to the global maximum is not automatically guaranteed. However, the class of surrogate losses we consider do exhibit reliable empirical performance across all our simulation settings. Our approach offers flexibility for learning the DTRs so that practitioners can tailor the method to the data and problem at hand. In particular, the smoothness of our surrogate losses makes the optimization problem suitable to a broad range of standard machine learning algorithms including, but not limited to neural networks, wavelet series and basis expansion. Interpretable treatment rules are also achievable by coupling our DTRESLO method with interpretable classifiers, such as linear or tree-based classifiers. Finally, since we optimize the primal objective function, variable selection in our case is straightforward via addition of an l_1 penalty.

1.1.0.3. *Theoretical guarantee for a class of DTR estimators*. We provide sharp upper bound on the regret—the difference between the optimal value function and the value attained by the estimated treatment regime, with detailed analyses focused on searching for DTR within the neural network classifiers. We perform a sharp analysis of our approximation error (see Theorem 4) and estimation error under Tsybakov's small noise condition (Tsybakov (2004)). Corollary 1 shows that the regret of our DTRESLO method with neural network classifiers decays at a fast rate, provided the optimization error is small. Here by *fast*, we mean decay rate faster than $n^{-1/2}$ is achievable. It turns out that this rate also matches the minimax rate of risk decay (up to a polylogarithmic factor) of binary classification under assumptions similar to ours (Audibert and Tsybakov (2007)). Since two-stage DTR is unlikely to be simpler than one-stage DTR, we conjecture that that our rate is minimax-optimal (up to a polylogarithmic factor) in two-stage DTR under our assumptions. In the special case when treatment effects are bounded away from zero, we show that our regret decays at the rate of $O(1/n)$ up to a polylogarithmic order.

The rest of the article is organized as follows. In Section 2, we outline the problem and discuss the mathematical formulation. In Section 3, we discuss Fisher consistency in the DTR setting, show that a large class of concave surrogates fail to be Fisher consistent and establish the Fisher consistency of a family of nonconcave surrogates. In Section 4, we construct a method for estimating the optimal DTRs using the Fisher consistent surrogates, and discuss the potential sources of error that contribute to the regret. Section 5 and Section 6 are devoted toward obtaining theoretical upper bounds of the regret of our DTRESLO method. Section 5 focuses on approximation error, which is combined with the estimation error in Section 6 to

yield the final regret bound. Then, in Section 7, we illustrate our DTRESLO method's empirical performance with extensive simulations. We continue with a discussion in Section 8. The application of DTRESLO to a sepsis cohort and the proofs of our theoretical results have been deferred to the Supplementary Material (Laha et al. (2024)) due to space constraints.

1.2. *Notation.* We let $\overline{\mathbb{R}}$ denote the extended real line $\mathbb{R} \cup \{\pm\infty\}$ and write $\mathbb{R}_+$ for the positive half line $\{x \in \mathbb{R} : x > 0\}$. Denote by $\mathbb{N} = \{1, 2, \ldots\}$ the set of all natural numbers and for any integer t, we let $[t] = \{1, 2, \ldots, t\}$. We also let $\mathbb{Z}$ denote the set of all integers. For $m \in \mathbb{N}$, we let $\|\cdot\|_m$ denote the l_m norm, that is, for $v \in \mathbb{R}^m$, $\|v\|_m = (\sum_{i=1}^m |v_i|^m)^{1/m}$. If $v \in \mathbb{N}^m$, we denote by $|v|_1$ the quantity $\sum_{i=1}^m v_i$. We let $B_m(0, K)$ denote the l_2-ball in $\mathbb{R}^m$ centered at the origin with radius $K > 0$.

For any probability measure P and measurable function f, we denote by $\|f\|_{P,k}$ the norm $(\int |f(x)|^k \, dP(x))^{1/k}$. We will also denote this norm by $L_k(P)$. Also, Pf will denote the integral $\int f \, dP$. For a concave function $f : \mathbb{R}^k \mapsto \mathbb{R}$, the domain $\mathrm{dom}(f)$ will be defined as in (Hiriart-Urruty and Lemaréchal ((2001), p. 74)), that is, $\mathrm{dom}(f) = \{x \in \mathbb{R}^k : f(x) > -\infty\}$. For $f : \mathbb{R}^2 \mapsto \mathbb{R}$, we denote by f_{12} the partial derivative

$$f_{12}(x, y) = \frac{\partial^2 f(x, y)}{\partial x \partial y}.$$

For any differentiable function $f : \mathbb{R}^k \mapsto \mathbb{R}$, ∇f will denote the gradient of f, and the super-level set of f at level c will be defined by $\{x \in \mathbb{R}^k : f(x) \geq c\}$. For any $x \in \mathbb{R}$, we denote by $\sigma(x)$ the ReLU activation function $x_+ = \max(x, 0)$. For any set A, use the notation $1[x \in A]$ to denote the event $\{x \in A\}$. Also, we denote by $\mathrm{int}(A)$ the interior of the set A. The cardinality of A will be denoted by $|A|$. Throughout this paper, we use the convention $\pm\infty \times 0 = 0$. In this paper, we will use C and c to denote generic constants, which may vary from line to line.

Many results in this paper are asymptotic (in n) in nature, and thus require some standard asymptotic notations. If a_n and b_n are two sequences of real numbers, then $a_n \gg b_n$ (and $a_n \ll b_n$) implies that $a_n/b_n \to \infty$ (and $a_n/b_n \to 0$) as $n \to \infty$, respectively. Similarly, $a_n \gtrsim b_n$ (and $a_n \lesssim b_n$) implies that $\liminf_{n\to\infty} a_n/b_n = C$ for some $C \in (0, \infty]$ (and $\limsup_{n\to\infty} a_n/b_n = C$ for some $C \in [0, \infty)$). Alternatively, $a_n = o(b_n)$ will also imply $a_n \ll b_n$ and $a_n = O(b_n)$ will imply that $\limsup_{n\to\infty} a_n/b_n = C$ for some $C \in [0, \infty)$.

2. Mathematical formalism. We focus on the DTR estimation under a longitudinal setting where data are collected over time periods indexed by $t \in \{1, 2\}$. Let $O_t \in \mathcal{O}_t \subset \mathbb{R}^{p_t}$ denote the p_t dimensional vector of patient clinical variables collected at time t and $p = \max(p_1, p_2)$. At a given time t, a binary treatment decision $A_t \in \{\pm 1\}$ is made for the patient and a response to such treatment $Y_t \in \mathbb{R}$ is observed. Without loss of generality, we assume higher values of response Y_t are desirable. Let us denote the distribution underlying the observed random vector $\mathcal{D} = (O_1, A_1, Y_1, O_2, A_2, Y_2)$ by $\mathbb{P}$. Suppose we sample n i.i.d. observations from $\mathbb{P}$. The corresponding empirical distribution function will be denoted by $\mathbb{P}_n$. Since treatment decisions are often made based on all previous states including prior treatments and responses, we define the patient history by

$$H_1 = O_1, \quad \text{and} \quad H_2 = (O_1, Y_1, O_2, A_1),$$

where H_1 and H_2 take values in sets $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. We denote by $\pi_1(a_1|H_1)$ and $\pi_2(a_2|H_2)$ the propensity scores $\mathbb{P}(A_1 = a_1|H_1)$ and $\mathbb{P}(A_2 = a_2|H_2)$, respectively.

Our goal is to find the treatment regime $d = (d_1, d_2) : \mathcal{H}_1 \times \mathcal{H}_2 \mapsto \{\pm 1\} \times \{\pm 1\}$ that maximizes the expected sum of rewards $Y_1(d) + Y_2(d)$,

$$V(d_1, d_2) = \mathbb{E}_d\big[Y_1(d) + Y_2(d)\big],$$

where $Y_t(d)$ is the potential outcome associated with time $t \in \{1, 2\}$, and $\mathbb{E}_d$ is the expectation with respect to the data distribution under regime d. To this end, first we make some assumptions on the observed data distribution $\mathbb{P}$ so that $V(d_1, d_2)$ becomes identifiable under $\mathbb{P}$.

Assumptions for identifiability

I. *Positivity*: There exists a constant $C_\pi \in (0, 1)$ so that $\pi_t(A_t|H_t) > C_\pi$ for all H_t, $t = 1, 2$.

II. *Consistency*: The observed outcomes Y_t and covariates O_t agree with the potential outcomes and covariates under the treatments actually received; see Robins (1994), Schulte et al. (2014) for more details.

III. *Sequential ignorability*: For each $t = 1, 2$, the treatment assignment A_t is conditionally independent of the future potential outcomes Y_t and future potential clinical profile O_{t+1} given H_t. Here, we take O_3 to be the empty set.

Our version of sequential ignorability follows from Murphy, van der Laan and Robins (2001), Robins (1997). Assumptions I–III are standard in DTR literature (e.g. Murphy, van der Laan and Robins (2001), Schulte et al. (2014), Sonabend-W et al. (2023), Zhao et al. (2015)).

Under Assumptions I–III, $\mathbb{E}_d(Y_1 + Y_2)$ can be identified under $\mathbb{P}$ as (Zhao et al. (2015))

$$\mathbb{E}_d\big[Y_1(d) + Y_2(d)\big] = \mathbb{P}\bigg[\frac{(Y_1 + Y_2)1[A_1 = d_1(H_1)]1[A_2 = d_2(H_2)]}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}\bigg].$$

The treatment effect contrasts are defined as follows:

$$\mathcal{T}_1(H_1) = \mathbb{E}\big[Y_1 + U_2^*(H_2)|A_1 = 1, H_1\big] - \mathbb{E}\big[Y_1 + U_2^*(H_2)|A_1 = -1, H_1\big], \tag{1}$$

$$\mathcal{T}_2(H_2) = \mathbb{E}[Y_1 + Y_2|A_2 = 1, H_2] - \mathbb{E}[Y_1 + Y_2|A_2 = -1, H_2], \tag{2}$$

$$\text{where} \quad U_2^*(H_2) = \max_{a_2 \in \{\pm 1\}} \mathbb{E}[Y_2|H_2, A_2 = a_2]. \tag{3}$$

The above quantities are also called the optimal blip-to-zero function, or sometimes simply the blip function in the literature (Luedtke and van der Laan (2016), Robins (2004), Schulte et al. (2014)). We will also refer to them as the first-stage and the second-stage conditional treatment effects. For the blip functions or the conditional treatment effects to be well defined, we need the conditional expectations in (1) and (2) to be finite, which is not automatically guaranteed by Assumptions I–III. Therefore, we introduce another assumption to ensure that the treatment effects are well defined.

Assumption IV. For any $h_2 \in \mathcal{H}_2$ and $a_2 \in \{-1, 1\}$, the conditional expectation $\mathbb{E}[|Y_1| + |Y_2||H_2 = h_2, A_2 = a_2] < \infty$. For any $h_1 \in \mathcal{H}_1$ and $a_1 \in \{-1, 1\}$, the conditional expectation $\mathbb{E}[Y_1 + U_2^*(H_2)|H_1 = h_1, A_1 = a_1] < \infty$. Furthermore, $\mathbb{E}[|Y_1 + Y_2|] < \infty$.

In addition to ensuring the well-definedness of treatment effects, Assumption IV also serves as a technical requirement in our proofs and enhances the interpretability of our theoretical findings. While we expect that many of our theoretical results would hold even without this assumption, the proofs would become more intricate and cumbersome. It is important to note that Assumption IV is not overly stringent, since in most of our applications, Y_1 and Y_2 represent measurements and are automatically bounded.

We define the optimal DTR d^* to be the maximizer of $\mathbb{E}_d[Y_1(d) + Y_2(d)]$ over all possible regimes $d = (d_1, d_2)$ such that $d_1 : \mathcal{H}_1 \mapsto \{\pm 1\}$ and $d_2 : \mathcal{H}_2 \mapsto \{\pm 1\}$. Under Assumptions I–III, the optimal policy d^* can be identified as follows (Chakraborty and Moodie (2013), Zhao et al. (2015)):

$$\begin{aligned} d_2^*(H_2) &= \arg\max_{a_2 \in \{\pm 1\}} \mathbb{E}[Y_2|H_2, A_2 = a_2], \\ d_1^*(H_1) &= \arg\max_{a_1 \in \{\pm 1\}} \mathbb{E}\big[Y_1 + U_2^*(H_2)|H_1, A_1 = a_1\big], \end{aligned} \tag{4}$$

where U_2^* is as defined in (3). Since the optimal decision rules remain unchanged if a constant C is added to both Y_1 and Y_2, in what follows, unless otherwise mentioned, we assume that $Y_1, Y_2 > C$ for some $C > 0$. This trick was also used in Zhao et al. (2015).

REMARK 1 (Uniqueness of d_1^* and d_2^*). It is worth noting that d_1^* and d_2^* defined in (4) may not be unique because they are allowed to take any value in $\{\pm 1\}$ at the boundary. To elaborate on this further, suppose some H_2 satisfies $\mathbb{E}[Y_2|H_2, A_2 = 1] = \mathbb{E}[Y_2|H_2, A_2 = -1]$. Such values of H_2 constitute the decision boundary for the second stage. Then both versions $d_2(H_2) = 1$ and $d_2'(H_2) = -1$ qualify as optimal rule for at H_2. Similarly, for d_1^*, we can show that if H_1 belongs to the first-stage decision boundary

$$\{h_1 \in \mathcal{H}_1 : \mathbb{E}[Y_1 + U_2^*(H_2)|A_1 = 1, h_1] = \mathbb{E}[Y_1 + U_2^*(H_2)|A_1 = -1, h_1]\},$$

then $d_1^*(H_1)$ can take either value $+1$ or -1. Thus, d_1^* is not unique either. Consequently, to avoid confusion, we let $d_1^* = 1$ and $d_2^* = 1$ at both first- and second-stage decision boundaries. Note that under this convention, $d_1^*(H_1) = 1[T_1(H_1) \geq 0]$ and $d_2^*(H_2) = 1[\mathcal{T}_2(H_2) \geq 0]$. In what follows, we shall also refer to this optimal rule as "the optimal rule." □

There is an alternative way of formulating d^*. If (f_1^*, f_2^*) is a maximizer of

$$V(f_1, f_2) = \mathbb{P}\left[\frac{(Y_1 + Y_2)1[A_1 f_1(H_1) > 0]1[A_2 f_2(H_2) > 0]}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}\right] \tag{5}$$

over the class

$$\mathcal{F} = \{(f_1, f_2)|f_1 : \mathcal{H}_1 \mapsto \mathbb{R}, f_2 : \mathcal{H}_2 \mapsto \mathbb{R} \text{ are measurable}\}, \tag{6}$$

then $\text{sign}(f_1^*)$ and $\text{sign}(f_2^*)$ yield the optimal rules d_1^* and d_2^*, respectively (Zhao et al. (2015)). If f_1^* and f_2^* take the value zero, then d_1^* and d_2^* can be either $+1$ or -1. Finally, even if d_1^* and d_2^* are unique, f_1^* and f_2^* need not be unique.

At this stage, although it is intuitive to consider maximization of the sample analogue of $V(f_1, f_2)$ to estimate the optimal decision rule, the nonconcavity and discontinuity of the zero-one loss function render the maximization of $V(f_1, f_2)$ computationally hard. To deal with issues of similar flavor, the classification literature (cf. Bartlett, Jordan and McAuliffe (2006)) suggests using a suitable surrogate to the zero-one loss function. We appeal to this very intuition and consider

$$V_\psi(f_1, f_2) = \mathbb{P}\left[\frac{(Y_1 + Y_2)\psi(A_1 f_1(H_1), A_2 f_2(H_2))}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}\right], \tag{7}$$

where ψ is some bivariate function. For example, Zhao et al. (2015) takes $\psi(x, y) = \min(x - 1, y - 1, 0)$, the bivariate concave version of the popular hinge loss $\phi(x) = \max(1 - x, 0)$.

Suppose there exist functions $f_1 : \mathcal{H}_1 \mapsto [-\infty, \infty]$ and $f_2 : \mathcal{H}_2 \mapsto [-\infty, \infty]$ so that

$$V_\psi(\tilde{f}_1, \tilde{f}_2) = \sup_{(f_1, f_2) \in \mathcal{F}} V_\psi(f_1, f_2), \tag{8}$$

where $\mathcal{F}$ is as defined in (6). Note that $\tilde{f}_1$ and $\tilde{f}_2$ may not be unique. Each $(\tilde{f}_1, \tilde{f}_2)$ lead to the decision rules $\tilde{d}_1(H_1) = \text{sign}(\tilde{f}_1(H_1))$ and $\tilde{d}_2(H_2) = \text{sign}(\tilde{f}_2(H_2))$. If $\tilde{f}(H_t) = 0$, then $\tilde{d}_t(H_t)$ can be either $+1$ or -1. We let $\tilde{f}_1$ and $\tilde{f}_2$ to be extended-valued functions because the supremum on the right-hand side of (8) may not be attained in $\mathcal{F}$ for some surrogates. It may happen that the supremum of V_ψ over $\mathcal{F}$ is attained at some f_1 and f_2, which satisfies $f_1(H_1) = \infty$ or $-\infty$ (alternatively, $f_2(H_2) = \infty$ or $-\infty$). Although $\tilde{f}_t$ can be extended valued, it does not create much technical issues because (a) $\tilde{d}_t$ is always 1 or -1 for $t = 1, 2$, and $\tilde{d}_t$ is the object of interest here.

Finally, we define excess risk in line with the excess risk in context of classification. Letting $V^* = V(f_1^*, f_2^*)$ and $V_\psi^* = V_\psi(\tilde{f}_1, \tilde{f}_2)$, we define the respective regret and ψ-regret of using (f_1, f_2) by $V^* - V(f_1, f_2)$ and $V_\psi^* - V_\psi(f_1, f_2)$, respectively. Note that regret and the ψ-regret are always nonnegative.

Throughout our paper, we will compare our DTR classification with binary classification. Therefore, we will fix the notation for binary classification. In the setting of binary classification, we have observations X taking value in an Euclidean space $\mathcal{X}$. Each X is associated with a label A, which plays the same role as our treatment assignments. The optimal rule or the Bayes rule assigns label $+1$ if $\eta(X) = P(A = 1|X) > 1/2$ and label -1 otherwise (cf. Bartlett, Jordan and McAuliffe (2006)). If $\eta(X) = 1/2$, both labels are optimal. The Bayes rule minimizes the classification risk $\mathcal{R}(f) = \mathbb{P}(Af(X) < 0)$ over all measurable functions $f : \mathcal{X} \to \mathbb{R}$. Also, we denote $\mathcal{R}^* = \inf_f \mathcal{R}(f)$, where the infimum is taken over all measurable functions. Replacing the zero-one loss with the surrogate ϕ results in the ϕ-risk $\mathcal{R}_\phi(f) = \mathbb{E}[\psi(Af(X))]$. We let $\mathcal{R}_\phi^*$ denote the optimized ϕ-risk.

Some parallels with the DTR classification setting are immediate. For example, $V(f_1, f_2)$, V^*, $V_\psi(f_1, f_2)$ and V_ψ^* correspond to $R(f)$, R^*, $R_\phi(f)$ and R_ϕ^*, respectively. Next, defining the maps $\eta_1 : \mathcal{H}_1 \mapsto \mathbb{R}$ and $\eta_2 : \mathcal{H}_2 \mapsto \mathbb{R}$ by

$$\eta_1(H_1) = \frac{\mathbb{E}[Y_1 + U_2^*(H_2)|A_1 = 1, H_1]}{\mathbb{E}[Y_1 + U_2^*(H_2)|A_1 = 1, H_1] + \mathbb{E}[Y_1 + U_2^*(H_2)|A_1 = -1, H_1]}, \tag{9}$$

$$\eta_2(H_2) = \frac{\mathbb{E}[Y_1 + Y_2|A_2 = 1, H_2]}{\mathbb{E}[Y_1 + Y_2|A_2 = 1, H_2] + \mathbb{E}[Y_1 + Y_2|A_2 = -1, H_2]}, \tag{10}$$

we observe that η_1 and η_2 play the same role in DTR setting as the conditional probability η in context of binary classification. To elaborate, from the definitions of d_1^* and d_2^* in (4), it follows that $d_t^*(H_t) = +1$ if $\eta_t(H_t) > 1/2$, and -1 otherwise. Note also that the first-stage and second-stage decision boundaries can be represented by the sets $\{h_1 : \eta_1(h_1) = 1/2\}$ and $\{h_2 : \eta_2(h_2) = 1/2\}$.

Throughout this paper, we occasionally make statements such as $\eta_1(H_1) \geq 1/2$, $T(H_2, A_2) \geq 0$, $d_1^*(H_1) \neq \tilde{d}_1(H_1)$, etc. Since H_1, H_2, A_1, A_2, etc. are random variables, quantities like $\eta_1(H_1)$, $\eta_2(H_2)$, $T(H_2, A_2)$, $d_2^*(H_1)$, $d_1^*(H_1)$ are also random. To avoid any confusion, we wish to clarify that when such statements are made, it implies that the stated conditions hold for all realizations of H_1, H_2, A_1, A_2, etc.

3. Fisher consistency. A desirable ψ should ensure that $\tilde{d}$ is consistent with d^*. To concertize the idea, we need the concept of Fisher consistency.

DEFINITION 1. The surrogate ψ is called Fisher consistent if for all $\mathbb{P}$ satisfying Assumption I–IV, any $\{f_{1n}, f_{2n}\}_{n\geq 1} \subset \mathcal{F}$ that satisfies

$$V_\psi(f_{1n}, f_{2n}) \to V_\psi^*, \quad \text{also satisfies } V(f_{1n}, f_{2n}) \to V^*.$$

Our definition of Fisher consistency is in line with classification literature (Bartlett, Jordan and McAuliffe (2006)). Note that Definition 1 does not require $\tilde{f}_1$ and $\tilde{f}_2$ to exist or be measurable. However, if $\tilde{f}_1$ and $\tilde{f}_2$ do exist, and they are in $\mathcal{F}$, then Fisher consistency implies $V(\tilde{d}) = V^*$, indicating $\tilde{d}$ is a candidate for d^*. In context of binary classification, the surrogate ϕ is Fisher consistent if and only if $\mathcal{R}_\phi(f_n) \to \mathcal{R}_\phi^*$ implies $\mathcal{R}(f_n) \to \mathcal{R}^*$, where f_n's are measurable functions mapping $\mathcal{X}$ to $\mathbb{R}$.

REMARK 2 (Characterization of Fisher consistency). In many classification problems, for example, binary, multicategory or multilabel classification, Fisher consistency can be directly characterized by convex hulls of points in the image space of ψ, and the related notion is known as calibration (Bartlett, Jordan and McAuliffe (2006), Gao and Zhou (2011),

Tewari and Bartlett (2007), Zhang (2010)). For example, Theorem 1 of Bartlett, Jordan and McAuliffe (2006) shows that a surrogate ϕ is Fisher consistent for binary classification if and only if the following condition holds.

CONDITION 1. $\phi : \mathbb{R} \mapsto \mathbb{R}$ satisfies

$$\sup_{x:x(2\eta-1)\leq 0} \big(\eta\phi(x) + (1-\eta)\phi(-x)\big) < \sup_{x\in\mathbb{R}} \big(\eta\phi(x) + (1-\eta)\phi(-x)\big)$$

for all $\eta \in [0, 1]$ such that $\eta \neq 1/2$.

However, due to the sequential nature of the DTR set-up, it is not easy to represent Fisher consistency in terms of analytical properties of ψ. This complicates the analysis of Fisher consistency in the DTR set-up. □

Traditionally, the first preference of surrogate losses have been the concave (convex in context of minimization) surrogates because they ensure unique optimum (Chen et al. (2017)). In the binary setting, a univariate concave surrogate ϕ is Fisher consistent if and only if it is differentiable at 0 with positive derivative (see Bartlett, Jordan and McAuliffe (2006), Theorem 6)). Many commonly used univariate concave losses satisfy these conditions. We display some of these in Figure G.2 in Section G in the Supplementary Material (Laha et al. (2024)). An important geometric property of these functions is that they mimic the graph of the zero-one loss function. After proper shifting and scaling, their image lies below that of the zero-one loss function (see Figure G.2 in Section G of the Supplementary Material (Laha et al. (2024))). Of course, concavity is not necessary for classification calibration, and this geometric property is shared by nonconcave, classification calibrated losses as well (see Lemma 9 of Bartlett, Jordan and McAuliffe (2006)).

There are also classes of concave surrogates, which are Fisher consistent for multicategory classification with respect to the zero-one loss (Duchi, Khosravi and Ruan (2018), Neykov, Liu and Cai (2016), Tewari and Bartlett (2007)) or for multilabel classification with respect to Hamming loss (Gao and Zhou (2011)). In that light, it is not unnatural to expect concave surrogates will succeed in the DTR classification setting as well. Unfortunately, as we will see in the next section, this simple-minded extension of binary classification may not hold.

DTR classification bears resemblance with multilabel classification (Dembczyński et al. (2012)) but additional complication arises since H_2 contains H_1. Also, the Fisher consistency literature on multilabel classification (Gao and Zhou (2011)) is based on Hamming loss and partial ranking loss, which are substantially different from the zero-one loss. Our problem also exhibits similarity with multiclass classification (Duchi, Khosravi and Ruan (2018)). However, a big difference arises because of the sequential structure. Had d_1 been a map from $\mathcal{H}_2$ to $\{\pm 1\}$ similar to d_2, existing theory on multiclass classification (Duchi, Khosravi and Ruan (2018)) could be readily used to provide conditions for a general function ψ to be Fisher consistent. However, during the treatment assignment d_1, one has no knowledge of A_1, Y_1 and O_2. Tewari and Bartlett (2007) and Zhang (2010) develop tools for general classification set-ups, but these tools are too generalized for explosion of the specific sequential structure of DTR classification. In fact, it is the binary classification, which seems to have the most parallels with DTR classification.

3.1. *Concave surrogates.* In this section, we will establish that a large class of concave surrogates fail to be Fisher consistent for DTR estimation. We first consider the case of smooth concave losses because they amend to gradient based optimization methods with good scalability properties. We will start our discussion with an example. The smooth concave function $\phi(x) = -\exp(-x)$ is Fisher consistent in the binary classification setting. Let

us consider its bivariate extension $\psi(x, y) = -\exp(-x - y)$. It turns out that $\tilde{d}_1(H_1)$ takes the form

$$\arg\max_{a_1 \in \{\pm 1\}} \mathbb{E}[h\big(Y_1 + \mathbb{E}[Y_2|H_2, A_2 = 1], Y_1 + \mathbb{E}[Y_2|H_2, A_2 = -1\big)|H_1, A_1 = a_1], \tag{11}$$

where $h(x, y) = \sqrt{xy}$. However, $d_1^*(H_1)$ takes the same form but with $h(x, y) = \max(x, y)$. In general, therefore, $\tilde{d}_1(H_1)$ and $d_1^*(H_1)$ do not agree. To see this, consider the toy example when $Y_1 = 1$ and

$$\begin{aligned} Y_2 = 4 \cdot 1[A_1, A_2 = 1] + 3 \cdot 1[A_1 = 1, A_2 = -1] + 5 \\ \cdot 1[A_1 = -1, A_2 = 1] + 1[A_1, A_2 = -1]. \end{aligned} \tag{12}$$

In this case, $d_1^*(H_1) = -1$ but $\tilde{d}_1(H_1) = 1$ for all H_1, and clearly, ψ is not Fisher consistent. If we consider other examples of smooth concave ψ, for example, logistic or quadratic loss, we obtain different h, but for these examples as well, h is quite different from the nonsmooth $h(x, y) = \max(x, y)$.

The above heuristics indicate that the criteria of DTR Fisher consistency may be incompatible with smooth concave losses. Theorem 1 below concretize the above heuristics for an important class of concave smooth losses. Theorem 1 assumes that ψ is closed and strictly concave. We say a function is closed if it is upper semicontinuous everywhere, or equivalently, if its superlevel sets are closed (Hiriart-Urruty and Lemaréchal ((2001), p. 78)). The function h is strictly concave if for any $\lambda \in (0, 1)$, and $x, y \in \text{dom}(h)$,

$$h\big(\lambda x + (1 - \lambda)y\big) > \lambda h(x) + (1 - \lambda)h(y).$$

THEOREM 1. *Suppose ψ is closed, strictly concave and bounded above. In addition, ψ has continuous second-order partial derivatives and ψ_{12} has continuous partial derivatives on int(dom(ψ)). Then ψ cannot be Fisher consistent for two-stage DTR.*

Following are some examples of ψ, also shown in Figure G.3 in Section G of the Supplementary Material (Laha et al. (2024)), which satisfy the assumptions of Theorem 1. (a) *Exponential*: $\psi(x, y) = -\exp(-x - y)$, (b) *Logistic*: $\psi(x, y) = -\log(1 + e^{-x} + e^{-y})$ (c) *Quadratic*: $\psi(x, y) = z^T Qz + b^T z + c$ where $z = (x, y)^T$, Q is negative definite, $b \in \mathbb{R}^2$ and $c \in \mathbb{R}$.

The proof of Theorem 1 is given in Section I of the Supplementary Material (Laha et al. (2024)). Our counterexample for Theorem 1 is based on a pathological case where O_2 and Y_1 are deterministic functions of H_1. We chose this case because it grants technical simplification. The realistic cases are no more likely to yield under concave surrogates than this simple pathological case. The calculations underlying the proof of Theorem 1 become severely technically challenging when the second-stage covariates are potentially random given H_1.

A main difficulty in proving Theorem 1 is that even under our pathological case, $\tilde{f}_1(H_1)$ and $\tilde{f}_2(H_2)$ do not have closed-form expressions. They are implicitly defined as maximizers of complex functionals of ψ. Therefore, if we consider a very large class of ψ's, characterization of $\tilde{f}_1(H_1)$ and $\tilde{f}_2(H_2)$ becomes difficult. The assumptions on ψ ensure that the class of ψ's under consideration is manageable, mitigating some technical difficulties in the characterization of $\tilde{f}_1(H_1)$ and $\tilde{f}_2(H_2)$. The latter is essential for learning the behavior of the signs of $\tilde{f}_1(H_1)$ and $\tilde{f}_2(H_2)$. Thus the assumptions on ψ are required for technical reasons in the proof. That is to say that our conditions on ψ are probably not necessary, and the assertions of Theorem 1 may hold even without these assumptions. In fact, we are not aware of any concave surrogates that are Fisher consistent in this context. We defer further discussion on the assumptions in Theorem 1 to Section A.1 of the Supplementary Material (Laha et al. (2024)).

The smoothness assumption in Theorem 1 is a technical assumption. Specifically, the existence of a gradient of ψ makes the proof simpler. However, we believe that the result may continue to hold without this condition, albeit with a more technically involved proof. In particular, the negative result in Theorem 1 is unlikely to be an artifact of the smoothness of ψ in Theorem 1, and may hold for broader classes of concave functions. In support of this claim, in Section 3.2, we demonstrate that a concave variant of the bivariate hinge loss $\min(x-1, y-1, 0)$, a commonly used nonsmooth concave loss, is not Fisher consistent. In fact, to our knowledge, there exists no concave surrogate, whether smooth or not, that is Fisher consistent for the DTR classification problem. These observations lead us to suspect that no concave loss is Fisher consistent for the DTR problem.

While we do not have an intuitive explanation for the apparent failure of concave functions, we attempt at making one heuristic reasoning. Even in the one-stage case of binary classification, it was observed that Fisher consistency requires the surrogates to mimic the shape of the zero-one loss to some extent. It appears to us that for Fisher consistency in two-stage DTR, the function ψ has to mimic the shape of the bivariate zero-one loss function (see Figure G.3a in Section G of the Supplementary Material, Laha et al. (2024)) more closely than that was necessary in binary classification (see Figure G.2 in Section G of the Supplementary Material, Laha et al. (2024)). In other words, the nonconcavity of the zero-one loss function at the origin pushes the concave losses to failure, thereby necessitating search for ψ among nonconcave losses, which we will study in Section 3.3.

Smooth concave or convex surrogates fail to be Fisher consistent in many other complex machine-learning problems. For example, Gao and Zhou (2011) shows known convex surrogates are not Fisher consistent for multilabel classification with ranking loss. Ranking is another notable example, where convex losses fail for a number of losses including the pairwise disjoint loss (Calauzenes, Usunier and Gallinari (2012)). In fact, in the latter case, the existence of a Fisher consistent concave surrogate would imply that the feedback arc-set problem is polynomial-time solvable (Duchi, Mackey and Jordan (2010)), which is conjectured to be NP complete (Karp (1972)). The DTR classification problem shares one common feature with the above-stated machine-learning problems where these surrogate losses fail. It does not *organically* reduce to a sequence of weighted binary classification problems, which appears to be a common element of all classification problems that are solvable via convex surrogates, for example, multicategory loss with zero-one loss function (Tewari and Bartlett (2007)), multilabel classification with partial ranking and hamming loss (Gao and Zhou (2011)), ranking with Hamming loss (Calauzenes, Usunier and Gallinari (2012)), ordinal regression with absolute error loss (Pedregosa, Bach and Gramfort (2017)), etc. Here, we emphasize the word "organic" because DTR classification does reduce to sequences of binary classification if it is framed as a sequential classification via exclusion of data points at each stage; cf. BOWL (Zhao et al. (2015)).

3.2. *Hinge loss*. In this section, we demonstrate the Fisher inconsistency of the nonsmooth loss function $\psi(x, y) = \min(x, y, 1)$, which is a bivariate version of the univariate hinge loss $\min(x, 1)$. The Fisher inconsistency of the hinge loss provides support to the conjecture that the Fisher inconsistency of concave surrogates extends beyond the class of smooth losses. The specific form of the hinge loss we examine has also been explored by Zhao et al. (2015) as well. See Figure G.3b in Section G of the Supplementary Material (Laha et al. (2024)) for a pictorial representation of this loss. If desired, readers may choose to bypass this section and proceed directly to Section 3.3, which focuses on the study of Fisher consistent losses.

Zhao et al. (2015) suggested a location transformation of the outcomes Y_1 and Y_2 so that they become positive, which is in alignment with our discussion in Section 2. Since we mainly

focus on their implementation of the hinge loss, we will take Y_1 and Y_2 to be positive for the time being. For our hinge loss, it turns out that we can especially characterize the solution $\tilde{d}$. The following inequality will be crucial for understanding the form of $\tilde{d}$ in this case:

$$\begin{aligned}&\big|\mathbb{E}\big[T\big(H_2, d_2^*(H_2)\big)|H_1 = h_1, A_1 = 1\big] - \mathbb{E}\big[T\big(H_2, d_2^*(H_2)\big)|H_1 = h_1, A_1 = -1\big]\big| \\ &\quad > \mathbb{E}\big[T\big(H_2, -d_2^*(H_2)\big)|H_1 = h_1, A_1 = 1\big] \\ &\qquad + \mathbb{E}\big[T\big(H_2, -d_2^*(H_2)\big)|H_1 = h_1, A_1 = -1\big],\end{aligned} \tag{13}$$

where we remind the readers that $T(H_2, a_2) = Y_1 + \mathbb{E}[Y_2|H_2, A_2 = a_2]$. Note that the left-hand side of (13) is the absolute value of the first-stage blip function or conditional treatment effect defined in (1). Thus (13) can be interpreted as a lower bound condition, indicating the minimum strength required for the first-stage conditional treatment effect. Further implications of (13) will be discussed after introducing Theorem 2, which demonstrates the necessity of (13) for the uniqueness of $\tilde{d}_1(H_1)$.

THEOREM 2 ($\tilde{d}_1$ and $\tilde{d}_2$ for hinge loss). *Suppose $\psi(x, y) = \min(x, y, 1)$. Further, suppose Assumptions I–IV hold and Y_1 and Y_2 are bounded below by some positive constant.*

1. *First stage*: *If* (13) *holds for some $h_1 \in \mathcal{H}_1$, then $\tilde{d}_1(h_1) = d_1^*(h_1)$. If* (13) *does not hold, then $\tilde{d}_1(H_1) = \{1, -1\}$.*
2. *Second stage*: *If $h_2 \equiv (h_1, a_1, y_1, o_2) \in \mathcal{H}_2$ is such that a_1 and h_1 satisfy $a_1 = \tilde{d}_1(h_1)$, then $\tilde{d}_2(h_2) = d_2^*(h_2)$. For all other h_2, $\tilde{d}_2(h_2) = \{-1, 1\}$.*

Theorem 2 follows from Theorem A.1 in Section A.2.1 of the Supplementary Material (Laha et al. (2024)), which is proved using straightforward algebra and elementary convex analysis results. The first observation from Theorem 2 is that the condition for $d_2^*(h_2) = \tilde{d}_2(h_2)$ is actually not restrictive. If the first-stage treatment allocation follows $\tilde{d}_1$, then $A_1 = \tilde{d}_1(H_1)$, and hence $\tilde{d}_2(H_2)$ matches with $d_2^*(H_2)$. However, the first stage appears to be more challenging for the hinge loss because when (13) fails to hold, this loss is unable to discriminate between the two treatment strategies in the first stage. If $d_1^*(H_1)$ is unique, then d_1^* and $\tilde{d}_1$ will disagree in such situations. As a trivial example of such a scenario, consider the illustration in (12). In this case, the absolute value of the first-stage conditional treatment effect is five but the threshold in the right-hand side of (13) is eight for all H_1. Thus $d_1^*(H_1)$ is unique, and it is always -1 but (13) does not hold in this example, thereby confirming Fisher inconsistency. We provide more examples of the failure of (13) in Section A.2.2 of the Supplementary Material (Laha et al. (2024)). Given that (13) represents a minimal strength condition for the first-stage conditional treatment effect, the above discussion indicates that the hinge loss requires a sufficiently strong first-stage conditional treatment effect to accurately identify the first-stage optimal treatment.

Similar to many other concave losses, the univariate version of hinge loss is Fisher consistent for the single-stage problem (Zhao et al. (2012)). However, Fisher inconsistency of Hinge loss has been observed in some classification problems involving more than two classes (see Liu (2007) for a detailed account). Hinge loss is also not Fisher consistent for maximum score estimation problem in linear binary response model (Feng, Ning and Zhao (2022)). In our case, the inconsistency stems from the first-stage treatment assignment, which aligns with the previous examples of concave losses in Section 3.1. This happens because the final-stage (in our case the second-stage) treatment assignment in DTR resembles a single-stage weighted classification problem, where concave surrogates work. The inherent difficulty of DTR manifests in the treatment assignments of the early stages. This is unsurprising because the early-stage treatment assignments need to take into account the potential outcomes of all future stages.

Some additional remarks are pertinent concerning the location transformation employed to ensure the positivity of outcomes because the location transformation makes it more challenging to satisfy (13). To see this, consider a hypothetical situation where (13) holds at $H_1 = h_1$ for some data distribution. If we perform a location shift by transforming Y_1 to $Y_1 + C$ and Y_2 to $Y_2 + C$, the left-hand side of (13) increases by C, while the right-hand side grows by $3C$. Therefore, if C is large enough, (13) will no longer hold for the location transformed data. Given the positivity of Y_1 and Y_2 does not ensure Fisher consistency anyway, one may question the form of $\tilde{d}$ when Y_1 and Y_2 are allowed to take nonpositive values. We delve deeper into this topic in Section A.2.1 of the Supplementary Material (Laha et al. (2024)).

REMARK 3. As previously mentioned, the SOWL method proposed by Zhao et al. (2015) is based on the bivariate hinge loss described in Theorem 2. With the hinge loss, the agreement between $\tilde{d}$ and d^* relies on the fulfillment of (13) when $d_1^*(H_1)$ is unique. From a high level, this condition requires the first-stage conditional treatment effect to be larger than some threshold. There are both distributions satisfying (13) for all $h_1 \in \mathbb{H}_1$, resulting in $\tilde{d} = d^*$; and distributions that violate (13) for some $h_1 \in H_1$, resulting in $\tilde{d} \neq d^*$. For specific examples and further elaboration, refer to Section A.2.2 of the Supplementary Material (Laha et al. (2024)).

3.3. *Construction of Fisher consistent surrogates.* In this section, we construct Fisher consistent loss functions for two-stage DTR classification. Noting the connection between binary classification and DTR classification, we consider bivariate loss functions of form $\psi(x, y) = \phi_1(x)\phi_2(y)$ where ϕ_1 and ϕ_2 themselves are univariate loss functions. The most intuitive choice of ϕ_i's would be the Fisher consistent losses for one-stage DTR. However, $\phi_i(x) = -\exp(-x)$ is Fisher consistent in one stage (Bartlett, Jordan and McAuliffe (2006), Chen et al. (2017)) although the product $\phi_1(x)\phi_2(y)$ is inconsistent for the two-stage setting (see Section 3). The above indicates that ϕ_i's Fisher consistency is insufficient for ψ to mimic the bivariate zero-one loss function effectively.

In fact, our calculations hint that ϕ_2 needs to share a particular property of the zero-one loss function, that is, for some constant $C > 0$,

$$\sup_{x \in \mathbb{R}} \big(\eta\phi_2(x) + (1 - \eta)\phi_2(-x)\big) = C \max(\eta, 1 - \eta). \tag{14}$$

The above property is satisfied by the sigmoid function, which is nonconcave, and Fisher consistent for binary classification (Bartlett, Jordan and McAuliffe (2006)). Interestingly, (14) alone does not guarantee the fisher consistency of $\psi = \phi_1\phi_2$. For instance, the loss $\phi(x) = \min(x + 1, 1)$ satisfies (14) with $C = 2$ (cf. Bartlett, Jordan and McAuliffe (2006)) but $\psi(x, y) = \phi(x)\phi(y)$ is not Fisher consistent for DTR when the number of stages is more than two. Therefore, (14) is not a sufficient for Fisher consistency. Now we introduce a sufficient condition for Fisher consistency.

CONDITION 2. ϕ is a strictly increasing function such that:

1. $\phi(x) > 0$ for all $x \in \mathbb{R}$.
2. For all $x \in \mathbb{R}$, $\phi(x)$ satisfies $\phi(x) + \phi(-x) = C_\phi$ where $C_\phi > 0$ is a constant.
3. $\lim_{x\to\infty} \phi(x) = C_\phi$ and $\lim_{x\to-\infty} \phi(x) = 0$.

We will show in the upcoming Theorem 3 that Condition 2 is sufficient for Fisher consistency in the sense that if ϕ satisfies Condition 2, then $\psi(x, y) = \phi(x)\phi(y)$ is Fisher consistent. A ϕ satisfying Condition 2 is Fisher consistent for binary classification, and it also satisfies (14) (see Lemma J.2 in Section J.1 of the Supplementary Material, Laha et al. (2024)).

Notably, this ϕ possesses another important property. When $C_\phi = 1$ and ϕ is continuous, ϕ becomes the distribution function of an unbounded symmetric random variable. In contrast, the previously mentioned univariate hinge loss $\phi(x) = \min(x + 1, 1)$ lacks this property. Specifically, when smooth, ϕ can be perceived as a smooth version of the 0–1 loss, smoothed via a symmetric distributional kernel. Consequently, it can be inferred that surrogates satisfying Condition 2 closely approximate the 0–1 loss. That being said, we do not yet know if Condition 2 is necessary for Fisher consistency in the DTR problem.

We provide some examples of functions satisfying Condition 2 below.

EXAMPLE 1. The following odd functions are nondecreasing with range $[-1, 1]$: (1) $f_a(x) = \frac{x}{1+|x|}$, (2) $f_b(x) = \frac{2}{\pi}\arctan(\frac{\pi x}{2})$, (3) $f_c(x) = \frac{x}{\sqrt{1+x^2}}$, (4) $f_d(x) = \tanh(x)$, where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Then $\phi_J(x) = 1 + f_J(x)$ satisfies Condition 2 with $C_\phi = 2$, for $J = a, b, c, d$. See Figures G.4a and G.4b in Section G of the Supplementary Material (Laha et al. (2024)) for the pictorial representation of these functions and the corresponding ψ's.

Our approach involving nonconcave surrogates leads to nonconvex optimization problems, prompting the question of how it differs from directly optimizing the original value function. While both approaches lead to nonconvex optimization, our method results in a smooth optimization problem. In contrast, direct maximization of the value function would lead to a discontinuous optimization problem with jump discontinuities. Moreover, the objective function resulting from the latter optimization problem is flat at the regions of continuity.

Our heuristic analysis on optimization error in Section D of the Supplementary Material (Laha et al. (2024)) suggests that the surface of DTRESLO optimization problem may possess favorable properties for certain policy classes. In such cases, we find that the optimization error incurred for DTRESLO with gradient descent-type algorithms may be small under specific conditions. In contrast, gradient descent-type methods would likely fail for the discontinuous problem resulting from direct value function maximization, and no known condition or method guarantees small optimization error for these methods in such problems (Xu, Wang and Fang (2014)). This makes direct optimization of the value function considerably more challenging than our method. Objectives with 0–1 loss appear naturally in various machine learning problems. As far as we know, In current statistical machine learning literature, direct optimization of such objectives is avoided, and instead, the original 0–1 loss is replaced with a more well-behaved surrogate loss, whether convex or not, whenever such a surrogate is available (Calauzenes, Usunier and Gallinari (2012), Feng, Ning and Zhao (2022), Gao and Zhou (2011), Horowitz (1992), Mukherjee, Banerjee and Ritov (2021), Pedregosa, Bach and Gramfort (2017), Xu, Wang and Fang (2014)).

The class specified by Condition 2 has been mentioned in various machine learning problems, often presented in forms appropriate for a minimization problem. In certain instances, it is referred to as the smoothed 0–1 loss. In some of these machine learning problems, this class has been proposed in situations where convex surrogates have demonstrated inconsistency. For example, Gao and Zhou (2011) has shown that this class of surrogates is Fisher consistent for multilabel classification with ranking loss, where convex surrogates are inconsistent. Feng, Ning and Zhao (2022) established Fisher consistency related guarantees for such surrogates in a diverse range of problems including Covariate-adjusted Youden index estimation, one-bit compress sensing and maximum score estimation in binary response model; see also Mukherjee, Banerjee and Ritov (2021), Xu, Wang and Fang (2014). Especially for maximum score estimation, Feng, Ning and Zhao (2022) showed that common convex surrogates such as exponential and hinge loss are inconsistent (Feng, Ning and Zhao (2022)). Finally, the surrogate loss used for multivariate ψ-learning in the context of multicategory classification is a nonsmooth member of our class (Liu and Shen (2006)). The authors of that work claim that this nonconcave surrogate outperforms SVM, which relies on hinge loss.

3.3.1. *Fisher consistency of ψ.* Instead of directly proving Fisher consistency, we will bound the true regret $V^* - V(f_1, f_2)$ in terms of the ψ-regret $V_\psi^* - V_\psi(f_1, f_2)$. The benefit of such a bound is that the rate of convergence of the true regret will be readily given by that of the ψ-regret, which we actually minimize.

As mentioned earlier, the true regret and the ψ-regret parallel the excess risk $\mathcal{R}(f) - \mathcal{R}^*$ and the ϕ-excess risk $\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*$ in binary classification. The relationship between the latter has been well studied. For Fisher consistent ϕ, Bartlett, Jordan and McAuliffe (2006) show that

$$h_\phi\big(\mathcal{R}(f) - \mathcal{R}^*\big) \le \mathcal{R}_\phi(f) - \mathcal{R}_\phi^*,$$

where h_ϕ is a convex function satisfying $h_\phi(0) = 0$. In view of the fact that the univariate sigmoid loss leads to a linear h_ϕ (Bartlett, Jordan and McAuliffe ((2006), Example 4)), it is reasonable to expect that a similar inequality holds when $\psi(x, y) = \phi(x)\phi(y)$ with ϕ as in Condition 2, as confirmed in following theorem.

THEOREM 3. *Suppose $Y_1, Y_2 > 0$ and Assumptions I–IV hold. Let $\psi(x, y) = \phi(x)\phi(y)$ with ϕ satisfying Condition 2 with some $C_\phi > 0$. Then*

$$V^* - V(f_1, f_2) \le \frac{(V_\psi^* - V_\psi(f_1, f_2))}{(C_\phi/2)^2}. \tag{15}$$

Theorem 3 immediately implies Fisher consistency because if $V_\psi(f_{1n}, f_{2n})$ converges to V_ψ^* for some $(f_{1n}, f_{2n}) \in \mathcal{F}$, then $V(f_{1n}, f_{2n}) \to V^*$ as well. Theorem 3 is proved in Section J.2 of the Supplementary Material (Laha et al. (2024)).

As mentioned earlier, a necessary requirement for Fisher consistency is an agreement between $\tilde{d}$ and d^*. Proving the latter is also a key step in the proof of Theorem 3. Let us provide some intuition as to why the $\tilde{d}$ corresponding to our ψ may agree with d^*.

We mentioned earlier that any ϕ satisfying Condition 2 is Fisher consistent for binary classification. It can be shown that Fisher consistency for binary classification translates to Fisher consistency for the single-stage case under Assumptions I–IV (Chen et al. (2017)). Using this insight, we can show that the second-stage treatment allocation $\tilde{d}_2(H_2)$ matches with $d_2^*(H_2)$ for our ψ. Regarding the first stage, after some algebraic manipulation, we can demonstrate that $\tilde{d}_1(H_1)$ takes the form in (11) analogous to the exponential loss, but with $h(x, y) = \max(x, y)$. This particular form of h is primarily driven by (14) and the positivity of ϕ. Since d_1^* satisfies (11) with $h(x, y) = \max(x, y)$, the above leads to $\tilde{d}_1 = d_1^*$.

We want to remind the readers that the assumption $Y_1, Y_2 > 0$ is not restrictive. As mentioned earlier, in cases where the observed outcomes are not positive, a location transformation can be applied to ensure positivity without altering the optimal treatment policy d^* and, consequently, $\tilde{d}$. We also want to emphasize that Theorem 3, as well as all our upcoming theorems, do not distinguish between continuous and discrete outcomes. Therefore, our method and theory apply to discrete and binary outcomes, which are of interest in many applications.

REMARK 4 (Scaling of the ϕ's). The scaling factor $C_\phi/2$ appears in the regret of (15) because ϕ differ from the zero-one function in scale by a factor of $C_\phi/2$. To understand the impact of the scaling factor in the regret bound, suppose $\phi_2 = a\phi_1$ for some $a > 0$, and $\psi_t(x, y) = \phi_t(x)\phi_t(y)$ for $t = 1, 2$. Then

$$\frac{(V_{\psi_1}^* - V_{\psi_1}(f_1, f_2))}{C_{\phi_1}^2/4} = \frac{(V_{\psi_2}^* - V_{\psi_2}(f_1, f_2))}{C_{\phi_2}^2/4}.$$

Thus, the regret bound in (15) does not depend on the scale of ϕ. Nevertheless, during our implementation, we take the scaling factor $C_\phi/2$ to be one so that the surrogate loss is at the same scale as the original zero-one loss.

We would like to emphasize a crucial point. While our algorithm is capable of handling large sample sizes, it is important to note, as we will discuss in Section D of the Supplementary Material (Laha et al. (2024)), that guarantees regarding its convergence to the global maximum are scarce. This limitation is a common challenge encountered in nonconcave optimization problems. However, it is important to recognize that we employ nonconcave losses due to the apparent absence of Fisher consistent concave losses. In other words, nonconcave optimisation may be the only viable choice if one aims to solve the DTR problem through simultaneous optimization. This highlights the inherent difficulty of the DTR problem.

To circumvent nonconcave optimization while retaining theoretical guarantees, one has two options: employing a stagewise, Fisher consistent optimization method like BOWL or opting for a regression-based approach such as Q-learning. However, it is important to note that BOWL achieves Fisher consistency at the expense of reduced sample size in the first stage Zhao et al. (2015). Our simulations in Section 7 indicate that BOWL does not outperform our proposed method. Additionally, our simulations demonstrate that BOWL exhibits significantly longer runtimes compared to our method for large sample sizes.

4. Main methodology. In this section, we describe how we use the Fisher consistent surrogate derived in Section 3.3 to estimate the optimal treatment regimes. For the remainder of this paper except Section A.2 of the Supplementary Material (Laha et al. (2024)), unless otherwise mentioned, ϕ will denote a univariate surrogate satisfying Condition 2, and ψ will denote the bivariate surrogate $\psi(x, y) = \phi(x)\phi(y)$ where ϕ satisfies Condition 2. Define the empirical ψ-value function

$$\widehat{V}_\psi(f_1, f_2) = \mathbb{P}_n\left[\frac{(Y_1 + Y_2)\psi(A_1 f_1(H_1), A_2 f_2(H_2))}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}\right]. \tag{16}$$

Because $\mathbb{P}$ is unknown, we maximize $\widehat{V}_\psi(f_1, f_2)$ instead of $V_\psi(f_1, f_2)$. Ideally, one should maximize $\widehat{V}_\psi(f_1, f_2)$ over $\mathcal{F}$ but brute force search over $\mathcal{F}$ is impossible unless $\mathcal{H}_1$ and $\mathcal{H}_2$ are discrete spaces with finite cardinality. Therefore, in practice, one may optimize $\widehat{V}_\psi(f_1, f_2)$ over a nested class $\mathcal{U}_1 \subset \cdots \subset \mathcal{U}_n \subset \mathcal{F}$, where $\mathcal{U}_n$ is some rich class of classifiers, preferably a universal class (see Zhang, Liu and Tao (2022)). We will discuss them in more detail later in Section 5. Whatever is the choice of $\mathcal{U}_n$, maximization of $\widehat{V}_\psi(f_1, f_2)$ over $(f_1, f_2) \in \mathcal{U}_n$ generally leads to a nonconvex optimization problem.

The surrogate loss based DTR optimization allows flexibility in the choice of $\mathcal{U}_n$ and the modification of the empirical loss $\widehat{V}_\psi(f_1, f_2)$ to accommodate high-dimensional covariates, nonlinear effects and variable selection. One can maximize $\widehat{V}_\psi(f_1, f_2) + \mathcal{P}(f_1, f_2)$ instead of $\widehat{V}_\psi(f_1, f_2)$ to enable variable selection and attain stable estimation, where $\mathcal{P}(f_1, f_2)$ is a penalty term. One can include complex basis functions in $\mathcal{U}_n$ to incorporate nonlinear effects. For example, tree and list based methods (Laber and Zhao (2015), Sun and Wang (2021), Zhang et al. (2018)) as well as neural networks (see Section 6.1.2 for details) can be potentially adapted to construct $\mathcal{U}_n$. Moreover, our method can be extended to K stages by taking $\psi(x_1, \ldots, x_k) = \prod_{i=1}^k \phi(x_i)$.

4.1. *Decomposition of errors.* In this section, we will discuss the decomposition of ψ-regret of DTRESLO into three sources of errors. To that end, let us denote our classifiers by $(\widehat{f}_{n,1}, \widehat{f}_{n,2})$. We will provide upper bounds for the ψ-regret $V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})$, which readily produces an upper bound for the true regret $V(\widehat{f}_{n,1}, \widehat{f}_{n,2})$ by Theorem 3. Before going into further detail, we point out that

$$V(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = \mathbb{P}\left[(Y_1 + Y_2)\frac{1[A_1\phi(\widehat{f}_{n,1}(H_1)) > 0]1[A_2\phi(\widehat{f}_{n,2}(H_2)) > 0]}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}\right]$$

is a random quantity because here we assume that H_t, A_t, Y_t $(t = 1, 2)$ are drawn from $\mathbb{P}$ independent of $(\widehat{f}_{n,1}, \widehat{f}_{n,2})$. The same holds regarding the ψ-regret $V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})$. We decompose the ψ-regret according to three sources of errors: (i) approximation error due to the approximation of $\mathcal{F}$ by $\mathcal{U}_n$; (ii) estimation error due to the use of finite sample and (iii) optimization error due to the possibility of not achieving global maximization for $\widehat{V}_\psi(f_1, f_2)$ since ψ is nonconcave. We define the optimization error as

$$\texttt{Opt}_n = \sup_{(f_1,f_2)\in\mathcal{U}_n} \widehat{V}_\psi(f_1, f_2) - \widehat{V}_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}).$$

We first provide some heuristics and intuitions for the error decomposition. For the time being, let us assume that $\arg\max_{(f_1,f_2)\in\mathcal{U}_n} V^*_\psi(f_1, f_2)$ is attained at some $(\tilde{f}_{n,1}, \tilde{f}_{n,2}) \in \mathcal{U}_n$. The existence of $(\tilde{f}_{n,1}, \tilde{f}_{n,2})$ is not guaranteed in general, and even if they exist, $(\tilde{f}_{n,1}, \tilde{f}_{n,2})$ will be hard to characterize for an arbitrary $\mathcal{U}_n$. We thus do not assume the existence of $(\tilde{f}_{n,1}, \tilde{f}_{n,2})$ in our proof. We define the map $\xi_{f_1,f_2,g_1,g_2} : \mathcal{H}_1 \times \mathcal{O}_2 \times \mathbb{R}^2 \times \{\pm 1\}^2 \mapsto \mathbb{R}$ by

$$\begin{aligned}&\xi_{f_1,f_2,g_1,g_2}(\mathcal{D})\\ &\quad= \frac{(Y_1+Y_2)\{\psi(A_1 g_1(H_1), A_2 g_2(H_2)) - \psi(A_1 f_1(H_1), A_2 f_2(H_2))\}}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)}.\end{aligned} \tag{17}$$

Elementary algebra shows that the ψ regret can be decomposed as follows:

$$\begin{aligned}&V^*_\psi - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})\\ &\quad= \underbrace{V^*_\psi - V_\psi(\tilde{f}_{n,1}, \tilde{f}_{n,2})}_{\text{Approximation error}} + (V_\psi - \widehat{V}_\psi)(\tilde{f}_{n,1}, \tilde{f}_{n,2}) - (V_\psi - \widehat{V}_\psi)(\widehat{f}_{n,1}, \widehat{f}_{n,2})\\ &\qquad+ \widehat{V}_\psi(\tilde{f}_{n,1}, \tilde{f}_{n,2}) - \sup_{(f_1,f_2)\in\mathcal{U}_n} \widehat{V}_\psi(f_1, f_2) + \underbrace{\sup_{(f_1,f_2)\in\mathcal{U}_n} \widehat{V}_\psi(f_1, f_2) - \widehat{V}_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})}_{\text{Optimization error: } \texttt{Opt}_n}\\ &\quad\le \text{Approximation error} + \underbrace{\big|(\mathbb{P}_n - \mathbb{P})[\xi_{f_1,f_2,\tilde{f}_{n,1},\tilde{f}_{n,2}}]\big|}_{\text{Estimation error}} + \texttt{Opt}_n.\end{aligned} \tag{18}$$

Clearly, $\texttt{Opt}_n$ depends on the optimization method used to maximize $\widehat{V}_\psi(f_1, f_2)$ over $\mathcal{U}_n$. We study the optimization error for linear DTR classes. For the sake of brevity, we have moved the discussion on the optimization error to Section D of the Supplementary Material (Laha et al. (2024)). The primary emphasis of this paper revolves around the estimation error and the approximation error. In our sharp analysis of the ψ-regret, the estimation error bound depends on the approximation error in an intricate manner; see Section B.1 of the Supplementary Material (Laha et al. (2024)) for more details. To keep our presentation short and focused, we present the main results on the approximation error in Section 5, and present the final regret bound in Section 6. Additional discussion on the regret decay is moved to Section B of the Supplementary Material (Laha et al. (2024)). Owing to the potential nonconcavity, the sharp analysis of the ψ-regret is significantly more subtle than existing results and approaches in the literature. We elaborate on this more in a detailed discussion presented in Section B.1 of the Supplementary Material (Laha et al. (2024)).

In what follows, similar to Zhao et al. (2015), we assume that the propensity scores, that is, π_1 and π_2 are known. This will hold in particular under a clinical trial like SMART (Kosorok and Laber (2019)), but not for observational data. When π_1 and π_2 are unknown, they can be estimated using a logistic regression model. This additional estimation step will not change the approximation error but the estimation error will likely change.

5. Approximation error. To establish the convergence rate of the approximation error, we require two assumptions. First, we require the standard assumption that the outcomes are bounded.

ASSUMPTION A. Outcomes Y_1, Y_2 satisfy $\max(Y_1, Y_2) \le C_y$.

The second assumption is the DTR version of Tsybakov's small noise assumption (Audibert and Tsybakov (2007), Tsybakov (2004)). Recall the blip functions/conditional treatment effects $\mathcal{T}_1$ and $\mathcal{T}_2$ defined in (1) and (2), respectively. Because we assume Y_1 and Y_2 are bounded away from zero, $\eta_t(H_t) = 1/2$ if and only if $\mathcal{T}_t(H_t) = 0$ for $t = 1, 2$. Therefore, the treatment boundary $\{h_t : \eta_t(h_t) = 1/2\}$ can also be formulated as $\{h_t : \mathcal{T}_t(h_t) = 0\}$.

In classification literature, it is well noted that obtaining a fast rate of convergence (faster than $n^{-1/2}$) requires control on the distribution of the random variable $\eta(X) - 1/2$ near the decision boundary $\{x : \eta(x) = 1/2\}$ to some degree, which gives rise to the so-called margin conditions (Audibert and Tsybakov (2007), Tsybakov (2004)). Similarly, in the DTR context, even with regression-based methods, regulation near the conditional treatment effect boundary $\{H_t : \mathcal{T}_t(H_t) = 0\}$ are generally required; see Appendix B.2 for a detailed discussion. Thus it is expected that we too would require control on the rate of decay of $\eta_1 - 1/2$ and $\eta_2 - 1/2$ near the treatment boundary $\{h_1 : \eta_1(h_1) = 1/2\}$ and $\{h_2 : \eta_2(h_2) = 1/2\}$, respectively, to obtain sharp bound on the ψ-regret. Among many variants of margin condition, we consider the Tsybakov small noise condition (Assumption MA of Audibert and Tsybakov (2007); see also Proposition 1 of Tsybakov (2004)), which has seen wide use in the literature (Audibert and Tsybakov (2007), Blanchard, Bousquet and Massart (2008), Steinwart and Scovel (2007)). The DTR formulation of Tsybakov's small noise condition takes the following form.

ASSUMPTION B (Tsybakov's small noise assumption). There exist a constant $C > 0$, a small number $t_0 \in (0, 1)$ and positive reals α_1, α_2 such that for all $t < t_0$,

$$P\big(0 < |\eta_1(H_1) - 1/2| \le t\big) \le Ct^{\alpha_1}, \qquad P\big(0 < |\eta_2(H_2) - 1/2| \le t\big) \le Ct^{\alpha_2}.$$

The parameters α_1 and α_2 are the *Tsybakov noise exponents*. We already noted that the Y_i's are bounded below. Since the outcomes are also bounded above by Assumption A, Assumption B is equivalent to saying

$$P\big(0 < \mathcal{T}_1(H_1) < t\big) + P\big(0 < \mathcal{T}_2(H_2) < t\big) \le Ct^{\alpha} \quad \text{for all } t \le t_0. \tag{19}$$

This alternative version is more common in precision medicine literature (Luedtke and van der Laan (2016), Qian and Murphy (2011)). See Section B.2 of the Supplementary Material (Laha et al. (2024)) for more details on the small noise assumption or similar assumptions in precision medicine literature. Finally, observe that if a stage has Tsybakov noise exponent α, then it also has noise exponent α' for all $\alpha' < \alpha$. Thus, to keep our calculations short, we assume that both stages have noise exponent α where $\alpha = \min(\alpha_1, \alpha_2)$. See Section B.2 of the Supplementary Material (Laha et al. (2024)) for further discussions on Assumption B.

Under Assumption B, it turns out that, our surrogates satisfying Condition 2 do not exhibit identical approximation error. The difference in the rate stems from the difference in their respective derivatives. Thus it will be convenient to split the above-mentioned surrogates into two types.

DEFINITION 2. We say a surrogate ϕ satisfying Condition 2 is of type A if there exists a constant $B_\phi > 0$ and $\kappa \ge 2$ such that $|\phi'(x)| < B_\phi(1 + |x|)^{-\kappa}$ for all $x \ne 0$. We say a surrogate ϕ satisfying Condition 2 is of type B if there exists a constant $B_\phi > 0$ and $\kappa > 0$ such that $|\phi'(x)| < B_\phi \exp(-\kappa|x|)$ for all $x \ne 0$.

TABLE 1
The type of the ϕ's in Example 1

$\phi(x)$	Type	B_ϕ	κ
(a) $x/(1+\lvert x\rvert)+1$	A	1	2
(b) $\frac{2}{\pi}\arctan(\pi x/2)+1$	A	2	2
(c) $x/\sqrt{1+x^2}+1$	A	$2^{3/2}$	3
(d) $1+\tanh(x)$	B	4	2

First, Definition 2 assumes ϕ to be smooth everywhere except perhaps at the origin. This restriction rules out nonsmooth ϕ's, but they are uninteresting from our implementation perspective anyways. All the ϕ's we have considered in Example 1 are differentiable at $\mathbb{R}/\{0\}$ (see Table 1; more details can be found in Section R of the Supplementary Material, Laha et al. (2024)). Second, type A merely means ϕ' decays polynomially in $|x|^{-\kappa}$, where type B ϕ's enjoy exponential decay of the derivative.

5.1. *Approximation error rate.* Theorem 4 summarizes the approximation error rate.

THEOREM 4. *Suppose $\mathbb{P}$ satisfies Assumptions I–IV, Assumption A and Assumption B with small noise coefficient α. Let $0 < a_n \to \infty$ be any sequence of positive reals. Further suppose there exist a small number $\delta_n \in (0,1)$ and maps $\tilde{h}_{n,1} : \mathcal{H}_1 \mapsto \mathbb{R}$ and $\tilde{h}_{n,2} : \mathcal{H}_2 \times \{0,1\} \mapsto \mathbb{R}$ so that*

$$\big\|\tilde{h}_{n,1} - (\eta_1 - 1/2)\big\|_\infty + \big\|\tilde{h}_{n,2} - (\eta_2 - 1/2)\big\|_\infty \le \delta_n,$$

where η_1 and η_2 are defined in (9) and (10), respectively. Then for any ϕ of type A, the following holds for any $\alpha' \in (0,\alpha)$ satisfying $\alpha - \alpha' < 1$:

$$V_\psi^* - V_\psi(a_n\tilde{h}_{n,1}, a_n\tilde{h}_{n,2}) \lesssim \begin{cases} a_n^{1-\kappa} + \min\big(\delta_n^{2+\alpha}a_n, \delta_n^{1+\alpha}\big) & \text{if } \kappa < 2+\alpha, \\ \dfrac{a_n^{-\frac{1+\alpha}{1+(\alpha-\alpha')/(\kappa-1)}}}{\alpha-\alpha'} + \min\big(\delta_n^{2+\alpha}a_n, \delta_n^{1+\alpha}\big) + \dfrac{\delta_n^{\alpha'+2-\kappa}}{(\alpha-\alpha')a_n^{\kappa-1}} & \text{if } \kappa \ge 2+\alpha. \end{cases}$$

Suppose ϕ is of type B. Then

$$V_\psi^* - V_\psi(a_n\tilde{h}_{n,1}, a_n\tilde{h}_{n,2}) \lesssim \frac{(\log a_n)^{1+\alpha}}{a_n^{1+\alpha}} + \min\big(a_n\delta_n^{2+\alpha}, \delta_n^{1+\alpha}\big) + a_n\delta_n\exp(-\kappa a_n\delta_n/2).$$

Theorem 4 entails that if $\{\tilde{h}_{n,1}, \tilde{h}_{n,2}\}$ approximates $\{\eta_1 - 1/2, \eta_2 - 1/2\}$ well in the sup-norm, then their scaled versions $\tilde{f}_{n,1} = a_n\tilde{h}_{n,1}$ and $\tilde{f}_{n,2} = a_n\tilde{h}_{n,2}$ incur small regret. It may appear a bit unusual in that we require $\tilde{f}_{n,i}$'s to be close to the functions $\eta_i - 1/2$'s, where V_ψ is actually maximized at $(\tilde{f}_1, \tilde{f}_2)$ (see Lemma J.1 in Section J of the Supplementary Material (Laha et al. (2024))). To that end, note that the extended real valued functions $\tilde{f}_i$'s cannot be approximated by any real valued f_i's because $\|f_i - \tilde{f}_i\|_\infty$ is infinity for all such f_i's. However, the proof of Theorem 4 ensures that $a_n(\eta_i - 1/2)$'s are good proxy for the $\tilde{f}_i$'s because $V_\psi(\tilde{f}_1, \tilde{f}_2) - V_\psi(a_n(\eta_1 - 1/2), a_n(\eta_2 - 1/2))$ is small. The bounds in Theorem 4 holds for any small δ_n, whose optimal rate will be found during the estimation error calculation. Theorem 4 bounds the approximation error because if $\mathcal{U}_n = \mathcal{U}_{1n} \times \mathcal{U}_{2n}$ is such that

$$\inf_{f_t \in \mathcal{U}_{tn}} \big\|f_t - (\eta_t - 1/2)\big\|_\infty < \delta_n, \quad t = 1, 2, \tag{20}$$

then Theorem 4 upper bound $V_\psi^* - \sup_{(f_1,f_2)\in\mathcal{U}_n} V_\psi(f_1, f_2)$.

A special scenario occurs when $\alpha = \infty$ in Assumption B. In this case, η_1 and η_2 are bounded away from zero on their respective domains. Under this condition, DTRESLO can obtain a regret decay rate up to $O_p(1/n)$, modulo a logarithmic term, in this situation when the optimization error is negligible. See Section B.3 of the Supplementary Material (Laha et al. (2024)) for a detailed discussion on this case.

6. Estimation error. In this section, we focus on the estimation error in (18), and provide sharp regret-bound for a selected set of classifiers by combining all sources of error. We assume that $(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \in \mathcal{U}_n = \mathcal{U}_{1n} \times \mathcal{U}_{2n}$, where $\mathcal{U}_{1n}, \mathcal{U}_{2n}$ are classes of functions. Our analysis in this section is fully nonparametric because our $\mathcal{U}_n$ is agnostic of the underlying data-generating mechanism. We first present a theorem on the regret decay rate of DTRESLO with general $\mathcal{U}_n$'s under Assumption B. Then we will proceed to study the specific example of neural networks. We also analyzed the regret decay rate of DTRESLO when $\mathcal{U}_n$ is the wavelet class, but this example has been deferred to Section B.4 of the Supplementary Material (Laha et al. (2024)) due to space constraints.

6.0.1. *Estimation error when $\mathcal{U}_n$ is a general function-class.* For general function-classes, we need some assumptions to control the complexity of $\mathcal{U}_n$. Such assumptions are widely used for bounding the expectation of the estimation error (Bartlett, Jordan and McAuliffe (2006), Bartlett and Mendelson (2002), Koltchinskii (2011)). To define complexity in the context of function-classes, we need to introduce the concept of the bracketing entropy. Given two functions f_l and f_u, the bracket $[f_l, f_u]$ is the set of all function f satisfying $f_l \le f \le f_u$. Suppose $\|\cdot\|$ is a norm on the function space and $\epsilon > 0$. Then $[f_l, f_u]$ is called an ϵ-bracket if $\|f_u - f_l\| < \epsilon$. For a function-class $\mathcal{G}$, we define the bracketing entropy $N_{[\,]}(\epsilon, \mathcal{G}, \|\cdot\|)$ to be the minimum number of ϵ-brackets needed to cover $\mathcal{G}$. This is a measure of the complexity of $\mathcal{G}$. We will see that the estimation error directly depends on the bracketing entropy of $\mathcal{U}_n$.

We will derive the estimation error of DTRESLO under the small noise assumption (Assumption B) when

$$N_{[\,]}(\epsilon, \mathcal{U}_{tn}, \|\cdot\|_\infty) \lesssim \left(\frac{A_n}{\epsilon}\right)^{\rho_n}, \quad t = 1, 2, \tag{21}$$

where $A_n, \rho_n > 0$. This leads to the regret bound of Theorem 5 that depends on A_n and ρ_n. The $\mathcal{U}_n$'s that satisfy (21) are called VC-type classes (p. 41 Koltchinskii (2011)). In all our examples, $\mathcal{U}_n$ will satisfy (21) for appropriate A_n and ρ_n.

THEOREM 5. *Suppose $\mathcal{U}_n$ is such that there exists $A_n > 0$ and $\rho_n \in \mathbb{R}$ so that (21) holds with $\liminf_n \rho_n > 0$, $\rho_n \log A_n = o(n)$, and $\liminf_n \rho_n \log A_n > 0$. Further suppose there exist $(\tilde{f}_{n,1}, \tilde{f}_{n,2}) \in \mathcal{U}_n$ so that*

$$\|\tilde{f}_{n,1}/a_n - (\eta_1 - 1/2)\|_\infty + \|\tilde{f}_{n,2}/a_n - (\eta_2 - 1/2)\|_\infty \le \left(\frac{\rho_n \log A_n}{n}\right)^{1/(2+\alpha)} \tag{22}$$

for some $a_n = n^a$ where $a > 1$. We also assume that $\mathbb{P}$ satisfies Assumptions I–IV, Assumption A and Assumption B with coefficient $\alpha > 0$. Then there exist $C > 0$ and $N_0 \ge 1$ such that for all $n \ge N_0$ and all $x > 0$, the following holds with probability at least $1 - \exp(-x)$:

$$V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \le C \max\left\{(1+x)^2 (\log n)^2 \left(\frac{\rho_n \log A_n}{n}\right)^{\frac{1+\alpha}{2+\alpha}}, \mathtt{Opt}_n\right\}.$$

Theorem 5 is proved in Section L of the Supplementary Material (Laha et al. (2024)). In the specific examples of neural networks and wavelets (see Section B.4 of the Supplementary Material (Laha et al. (2024)) for the latter), the regret bound in Theorem 5 leads to sharp rates provided η_1 and η_2 satisfy some smoothness conditions.

6.1. *Examples of regret bounds when $\mathcal{U}_n$ is the neural network class.* We will first state some assumptions on η_1 and η_2. Next, we will elaborate on the special cases when $\mathcal{U}_n$ corresponds to neural network classes.

6.1.1. *Smoothness assumption.* First, we explain why smoothness conditions on η_1 and η_2 are required. When we search for the DTRs among a class $\mathcal{U}_n$, for example, a class os neural networks or wavelets, instead of $\mathcal{F}$, we are basically restricting the search space. Although a class $\mathcal{U}_n$ with lower complexity helps lowering the estimation error, we require structural assumptions on η_1 and η_2 to ensure that η_1 and η_2 are well approximable by such $\mathcal{U}_n$'s—giving rise to the so-called complexity assumptions. Thus the complexity assumption enables the attainment of a small estimation error without necessarily blowing up the approximation error. See Audibert and Tsybakov (2007), Koltchinskii (2011) and Tsybakov (2004), among others, for a more detailed account of the necessity of complexity assumptions. If η_1 and η_2 are smooth, then they are well approximated by neural networks and basis expansion-type classes such as wavelets (Schmidt-Hieber (2020)). Therefore, we will assume our η_1 and η_2 to be smooth. To fix the idea, we define the Hölder classes with smoothness index $\theta > 0$ below.

Let $p \in \mathbb{N}$. A function $f : \mathcal{X} \subset \mathbb{R}^p \mapsto \mathbb{R}$ is said to have Hölder smoothness index $\theta > 0$ if for all $u = (u_1, \ldots, u_p) \in \mathbb{N}^p$ satisfying $|u|_1 < \theta$, $\partial^u f = \partial^{u_1} \partial^{u_2} \ldots \partial^{u_p} f$ exists and there exists a constant $C > 0$ so that

$$\frac{|\partial^u f(x) - \partial^u f(y)|}{|x - y|^{\theta - \lfloor \theta \rfloor}} < C \quad \text{for all } x, y \in \mathcal{X}.$$

For some $\mathcal{Y} > 0$, we denote by $\mathcal{C}_d^\theta(\mathcal{X}, \mathcal{Y})$ the Hölder class of functions given by

$$\left\{ f : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R} \Big| \sum_{u:|u|_1 < \theta} \|\partial^u f\|_\infty + \sum_{u:|u|_1 = \lfloor \theta \rfloor} \sup_{\substack{x,y \in \mathcal{X} \\ x \neq y}} \frac{|\partial^u f(x) - \partial^u f(y)|}{|x - y|^{\theta - \lfloor \theta \rfloor}} \leq \mathcal{Y} \right\}. \tag{23}$$

Since H_t may include categorical variables such as smoking status, we separate the continuous and categorical parts of H_t as $H_t = (H_{ts}, H_{tc}) \in \mathcal{H}_t = \mathcal{H}_{ts} \otimes \mathcal{H}_{tc}$, where $H_{ts} \in \mathcal{H}_{ts} \subset \mathbb{R}^{p_{ts}}$ and $H_{tc} \in \mathcal{H}_{tc} \subset \mathbb{R}^{p_{tc}}$ correspond to the continuous and categorical part of H_t, for $t = 1, 2$.

ASSUMPTION C (Smoothness assumption). $\mathcal{H}_1$ and $\mathcal{H}_2$ are compact and $\mathcal{H}_{1c}$ and $\mathcal{H}_{2c}$ are finite sets. Also, there exist $\theta > 0$ and $\mathcal{Y} > 0$ so that the following hold:

1. Let $\mathcal{X} = \mathcal{H}_{1s}$. For each $h \in \mathcal{H}_{1c}$, the map $\eta_1(\cdot, h) : \mathcal{X} \mapsto \mathbb{R}$ is in $\mathcal{C}_{p_{1s}}^\theta(\mathcal{X}, \mathcal{Y})$.
2. Let $\mathcal{X} = \mathcal{H}_{2s}$. For each $(h, a) \in \mathcal{H}_{2c} \times \{\pm 1\}$, the function $\eta_2(\cdot, h, a) : \mathcal{X} \mapsto \mathbb{R}$ is in $\mathcal{C}_{p_{2s}}^\theta(\mathcal{X}, \mathcal{Y})$.

We formulated the smoothness assumption in terms of η_1 and η_2 so that our results are consistent with contemporary classification literature. However, our proofs show that one could formulate the assumptions in terms of the smoothness of the blip functions in (1) and (2) as well. The compact support assumption for H_t, which is typically satisfied in real applications, is also commonly required in the DTR literature (Sonabend-W et al. (2023), Zhang et al. (2018), Zhao et al. (2012, 2015)). Under the compactness assumption, $\mathcal{H}_{1c}$ and $\mathcal{H}_{2c}$ are finite sets. Smoothness conditions as Assumption C have appeared in DTR literature in the context of nonparametric estimation (Sun and Wang (2021)). Compared to the parametric assumptions often imposed on the blip functions in Q-learning or A-learning (Schulte et al. (2014)), our smoothness assumptions are much weaker. Our smoothness assumption includes nondifferentiable functions as well.

6.1.2. *Neural network class.* We consider the neural network space in line with Schmidt-Hieber (2020)'s construction. Let $\mathcal{F}(L, W, s, \mathcal{Y})$ be the class of ReLU networks uniformly bounded by $\mathcal{Y} > 0$, with depth $L \in \mathbb{N}$, width vector W, sparsity $s \in \mathbb{N}$ and weights bounded by one. The output layer of the networks in $\mathcal{F}(L, W, s, \mathcal{Y})$ uses a linear gate. In this example, we consider that for $t = 1, 2$, the class $\mathcal{U}_{tn}$ corresponds to $\mathcal{F}(L_n, W_n, s_n, \mathcal{Y}_n)$ where L_n, W_n, s_n and $\mathcal{Y}_n$ may depend on n. To avoid cumbersome notation, we drop n from L_n, W_n, s_n and $\mathcal{Y}_n$, and simply denote them by L, p, s and $\mathcal{Y}$, respectively. One can control the complexity of this class via prespecifying upper bounds on the depth, width and sparsity of the network. Corollary 1 establishes the regret bound of DTRESLO with neural network classifier under Assumption B.

COROLLARY 1. *Suppose $\mathbb{P}$ satisfies Assumptions I–IV, Assumption A, Assumption B with parameter $\alpha > 0$ and Assumption C with parameter $\theta > 0$. Let $\mathcal{U}_{n,1}$ and $\mathcal{U}_{n,2}$ be of the form $\mathcal{F}(L, W, s, \infty)$ with appropriate W_1, where $\mathcal{F}(L, W, s, \infty)$ is as defined in Section 6.1.2. Suppose $L = c_1 \log n$, $s = c_2 n^{p/((2+\alpha)\theta+p)}$, and the maximal width $\max W \le c_3 s/L$ where $c_1, c_2, c_3 > 0$. Then there exist $N_0 > 0$ and $C > 0$ depending on $\mathbb{P}$ and ψ such that if $c_1, c_2, c_3 > C$, then for $n \ge N_0$ and any $x > 0$, the following holds with probability at least $1 - \exp(-x)$:*

$$V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \le C \max\{(1+x)^2 (\log n)^{\frac{6+4\alpha}{2+\alpha}} n^{-\frac{1+\alpha}{2+\alpha+p/\theta}}, \mathtt{Opt}_n\}.$$

The proof of Corollary 1 can be found in Section M.1 of the Supplementary Material (Laha et al. (2024)). The proof of Corollary 1 assumes that p is fixed, that is, it does not grow with n. The generic constant C in Corollary 1 may depend on p as well. Under Assumptions similar to A, B and C, the rate $n^{-\frac{1+\alpha}{2+\alpha+p/\theta}}$ is minimax in context of binary classification (Audibert and Tsybakov (2007)). Since two-stage weighted classification problem is not easier than binary classification, this rate is expected to be the minimax rate under our set-up as well. To the best of our knowledge, no other nonparametric DTR method has better guarantees for the regret under set-up similar to ours. See Section C of the Supplementary Material (Laha et al. (2024)) for a detailed comparison of the regret bound of DTRESLO with other nonparametric DTR methods such as BOWL/SOWL (Zhao et al. (2015)), nonparametric Q-learning, the list-based method of Zhang et al. (2018) and the stochastic tree-based reinforcement learning (ST-RL) method of Sun and Wang (2021).

7. Empirical analysis. We compare the performance of our DTRESLO method with the regression-based method Q-learning and the direct search methods BOWL and SOWL (Zhao et al. (2015)). For our DTRESLO method, we take $\phi(x) = 1 + 2/\pi \cdot \arctan(\pi x/2)$ because simulation shows that it has slightly better performance than the other smooth surrogates considered in Example 1. The code for implementing DTRESLO is provided at the second author's github page at Sonabend-W (2022). We consider several choices for the class of classifiers $\mathcal{U}_{1n}$ and $\mathcal{U}_{2n}$. When we consider the linear treatment policies, $\mathcal{U}_{1n}$ and $\mathcal{U}_{2n}$ are the class of all linear functions on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. We consider cubic spline, wavelets and neural network (NN) as the nonlinear treatment policies, with $\mathcal{U}_{1n}$ and $\mathcal{U}_{2n}$ being the respective function-classes in these cases. For the comparators, that is, Q-learning, BOWL and SOWL, we consider both linear and nonlinear policies as well. Following Zhao et al. (2015), we incorporate nonlinear policies for BOWL and SOWL using a reproducing kernel Hilbert space (RKHS) with RBF kernel; see Zhao et al. (2015) for more details. The nonlinear treatment policy for Q-learning is achieved by letting the Q-functions be in neural network classes. See Section E of the Supplementary Material (Laha et al. (2024)) for more details on the implementation of these methods.

We considered five broad simulation settings as detailed in Section E of the Supplementary Material (Laha et al. (2024)):

1. All covariates are discrete. Hence, an exhaustive search over $\mathcal{F}$ is possible using saturated models.

2. This is a setting with nonlinear decision boundaries in both stages. However, Y_2 does not depend on A_1.

3. This setting is inspired by Setting 2 of Zhao et al. (2015), where the outcome models, that is, $\mathbb{E}[Y_t|H_t]$'s are linear function of H_t for $t = 1, 2$. We will call this setting the linear setting.

4. This has highly nonlinear and even nonsmooth decision boundaries.

5. This setting has a higher number of covariates. In this case, the first-stage outcome model is linear, but the second-stage outcome model is nonlinear.

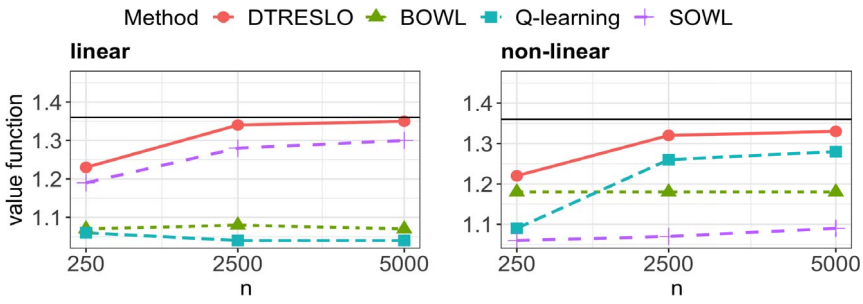
Setting 1 is a simple toy setting. The motivation behind including this setting is to verify the consistency of DTRESLO. We will use the linear setting 3 to check if linear treatment policies perform well when the outcome models are linear. On the other hand, we include settings 2 and 4 to examine if the methods with nonlinear treatment policies have an edge over those with linear treatment policies when the decision boundaries are nonlinear. Finally, setting 5 is included to compare the performance of different methods when the dimension of $\mathcal{O}_1 \cup \mathcal{O}_2$ is comparatively larger.

Under each listed setting, we estimate the DTRs based on samples of size $n = 250, 2500, 5000$. For each estimated DTR $\widehat{d}$, we estimate the value function $V(\widehat{d}_1, \widehat{d}_2)$ by the empirical value function based on an independent sample of size 10,000. We estimate the expectation and the standard deviation of these value function estimates using 500 Monte Carlo replications. We also estimate the optimal value function $V^* = V(d_1^*(H_1), d_2^*(H_2))$ for each setting using these 500 Monte Carlo replications. Figures 1 and 2 compare the estimated expected value functions of the different methods under consideration. In these figures, we use the neural network DTRESLO as the nonlinear DTRESLO because this method is comparable to neural network Q-learning. The average value functions corresponding to the other nonlinear DTRESLO methods can be found in Table E.2 in Section E of the Supplementary Material (Laha et al. (2024)). The overall performance of all the nonlinear DTRESLO methods is quite similar, although NN DTRESLO is slightly better than the rest.

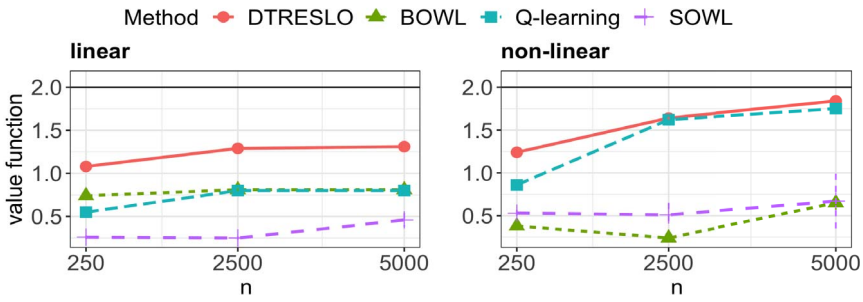
First of all, Figures 1 and 2 entail that DTRESLO consistently performs better or at least as good as the other methods under all our settings and all sample sizes. No other method has reliable performance across all settings. First, we will investigate the five settings in more detail. Then we will look more closely into the comparison between DTRESLO and the other methods. Finally, we will compare the run-time of different methods.

Figure 1a underscores that in the simple setting 1, DTRESLO outperforms all other methods under both linear and nonlinear treatment policies. Figures 1b and 1c show that in the nonlinear settings 2 and 4, as expected, the nonlinear versions of DTRESLO, BOWL and Q-learning perform better than the linear counterparts. The only exception is the case of SOWL, which we will discuss later in more detail. We also observe that setting 4 is quite hard in that the expected value function of all methods is noticeably lower than the optimal value function. In settings 2 and 4, nonlinear DTRESLO performs noticeably better than nonlinear Q-learning in a small sample ($n = 250$). As the sample size increases, the difference decreases. SOWL has poor performance under both settings. Although BOWL has better performance than SOWL, its performance improves rather slowly with n when compared to DTRESLO. This difference is most noticeable for the nonlinear treatment policies under large samples.

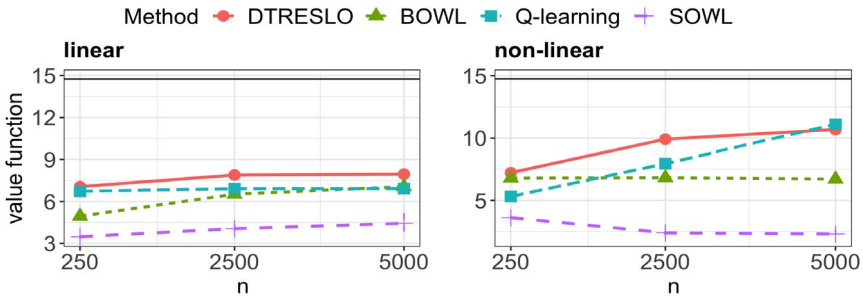
Figure 2a implies that under the linear setting 3, value function estimates of the linear treatment policies are as large as the nonlinear policies for all methods except SOWL. Setting 5, which has a larger number of variables, is a relatively more complicated setting. Although the second-stage outcome models are nonlinear in this setting, Figure 2b underscores that linear DTRESLO performs quite comparably to nonlinear DTRESLO in large samples under



(a) Setting 1.



(b) Setting 2.

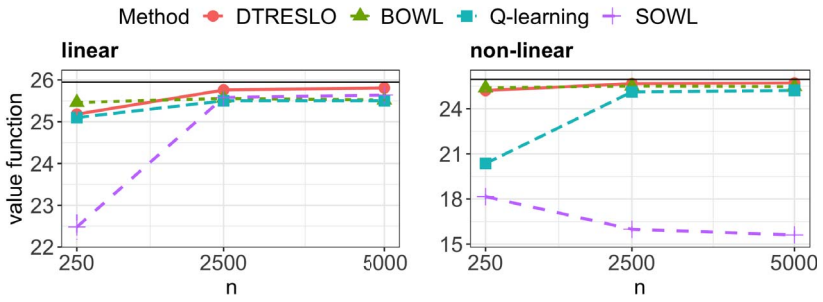


(c) Setting 4.

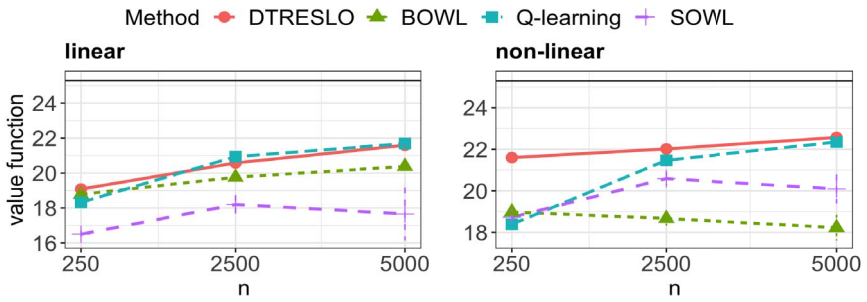
FIG. 1. *Plot of the estimated average value functions for the settings* 1, 2 *and* 4. *Here, the black horizontal line corresponds to the true value function. The left and right panels correspond to the linear and nonlinear treatment policies, respectively. Here, the nonlinear DTRESLO corresponds to the neural network classifier. The error bars are given by* ± 2 *SD.*

this setting. Table E.2 in the Supplementary Material (Laha et al. (2024)) implies that the situation with the other nonlinear DTRESLO methods is similar. Similar to settings 2, in this case, nonlinear DTRESLO has a noticeable edge over all other methods when the sample size is 250.

Under all settings, nonlinear DTRESLO and Q-learning exhibit one particular pattern, which merits some discussion. Nonlinear DTRESLO performs better than nonlinear Q learning in small samples, but their performance becomes almost similar when the sample size increases to 5000. The relative underperformance of nonparametric Q-learning in small samples may be due to its heavy reliance on the correct estimation of Q-functions. Nonparametric estimation of functions is harder unless the sample size is sufficiently large. In contrast, DTRESLO only needs to estimate the sign of the blip functions, which is easier than the estimation of the whole function. Finally, this difference may be the manifestation of the speculated faster regret decay of neural network DTRESLO (see Section C of the Supplement, Laha et al. (2024)). Thus our simulation study complements the theoretical comparison of the regrets between nonparametric Q-learning and DTRESLO.



(a) Setting 3.



(b) Setting 5.

FIG. 2. *Plot of the estimated average value functions for the settings* 3 *and* 5. *Here, the black horizontal line corresponds to the true value function. The left and right panels correspond to the linear and nonlinear treatment policies, respectively. Here, the nonlinear DTRESLO corresponds to the neural network classifier. The error bars are given by* ± 2 *SD.*

DTRESLO outperforms the other direct search methods, BOWL and SOWL, under all settings except the linear setting, that is, setting 3, where BOWL and DTRESLO have comparable performance. The difference is most pronounced for nonlinear treatment policies in large samples. DTRESLO's advantage over BOWL may be attributed to DTRESLO's simultaneous optimization approach as opposed to BOWL's stagewise approach. The latter reduces the effective sample size in the first stage. In general, SOWL's average value function stays quite below the optimal value function. Its performance is comparable to other methods only in setting 3, where classification is comparatively easy.

Figures 1 and 2 entail that the estimated value function of nonlinear SOWL, a nonparametric method by design, either does not improve with the sample size or exhibits a much slower increase compared to the other competing methods we consider. The last observation raises the question of whether the approximation error of SOWL at all decays to zero as the sample size increases. Indeed, this observation does not refute our Theorem 2, which establishes that the hinge loss, the surrogate employed in SOWL, requires the fulfilment of (13) for $\tilde{d}_1(H_1)$ to align with $d_1^*(H_1)$. Moreover, in Section A.2.2 of the Supplementary Material (Laha et al. (2024)), we demonstrate that (13) is not a pathological condition, as it fails in numerous nontrivial scenarios. To elaborate further, we focus on Setting 3 as an illustrative case. In this case, the nonparametric version of SOWL exhibits a decaying value with respect to n. For this setting, the outcome models are linear, and $H_1 \in \mathbb{R}^3$ follows a centered multivariate Gaussian distribution with an identity covariance matrix. Consequently, H_1 lies inside a ball of radius 5 centered at the origin with a high probability (specifically, greater than 0.999). However, we empirically evaluated that (13) holds nowhere inside this ball. Moreover, if a location transformation is required to ensure the positivity of outcomes for certain samples, as discussed in Section 3.2, (13) becomes more difficult to satisfy. Therefore, the suboptimal

TABLE 2
Run-time for estimating DTR for our smooth surrogates (DTRESLO), Zhao et al. (2015)'s BOWL and SOWL, and Q-learning under settings 1–5

		DTRESLO				BOWL		SOWL		Q-learning	
Setting	n	Linear	Wavelet	Spline	NeuNet	Linear	RBF	Linear	RBF	Linear	NeuNet
1	250	0.04	0.05	0.04	0.1	1.24	21.01	0.1	0.16	0.07	0.18
	2500	0.42	0.49	0.43	0.94	13.11	655.43	54.19	80.82	0.69	2.12
	5000	0.89	1.02	0.8	2.15	77.45	3913.85	400.32	534.36	1.4	3.94
2	250	0.04	0.06	0.04	0.16	1.36	3.48	1.3	1.33	0.09	0.19
	2500	0.54	0.6	0.42	1.01	27.75	271.08	773.79	822.42	0.73	1.83
	5000	0.88	1.25	0.91	3.7	136.88	5139.53	5901.54	5755.75	1.49	4.15
3	250	0.05	0.05	0.04	0.15	12.59	24.39	0.08	0.13	0.08	0.2
	2500	0.41	0.71	0.42	1.03	25.75	704.16	46.55	106.67	0.73	2.04
	5000	1.22	1.04	0.84	2.19	107.99	4063.34	345.83	859.37	1.46	5.36
4	250	0.04	0.06	0.04	0.1	1.66	3.21	1.28	1.33	0.09	0.19
	2500	0.59	0.49	0.42	1.04	20.12	424.81	806.54	833.64	0.73	1.88
	5000	0.86	1.32	0.91	1.5	70.86	3317.64	5674.96	5778.79	1.47	3.61
5	250	0.04	0.05	0.04	0.09	10.97	18.05	1.26	1.32	0.07	0.18
	2500	0.6	0.48	0.42	1.53	33.46	222.42	810.59	833.1	0.72	1.92
	5000	1.19	1.26	1.21	2.1	169.31	1012.86	5645.25	6010.88	1.38	3.79

performance of nonlinear SOWL may be attributable to the potential failure of (13) in this case.

Table 2 tabulates the run-time of the DTR estimation methods. Run times for DTRESLO with linear, wavelets, and spline-based treatment policies are relatively similar. The run-time doubles for neural network treatment policies. Nonetheless, they are all less than three seconds. Both linear and neural network Q-learning methods are slightly slower than their DTRESLO counterparts, but the difference in run-time is negligible. This is not surprising because DTRESLO and Q-learning methods are trained in a similar way. They all use stochastic gradient descent with RMSprop for optimization of the respective loss functions. All these methods are trained for 20 epochs and use a batch size of 128. As expected, BOWL and SOWL have a much larger run-time, which also increases sharply with n. This larger run-time is expected because SVMs utilize the dual space for optimization. The time cost is especially high in settings 2 and 4, which have highly nonlinear decision boundaries, and setting 5, which has over 32 features.

To summarize, DTRESLO improves the scalability of existing direct search methods, achieving run-time as small as Q-learning. We also observe that within the same class of treatment regimes, that is, linear or neural network, DTRESLO outperforms regression-based Q-learning in small samples. This observation aligns with the existing claim in the literature that classification is easier than regression in the context of DTR especially in small samples (Kosorok and Laber (2019), Zhao et al. (2015)). This may happen because regression-based methods focus on minimizing the squared error loss, where the estimation of optimal rules only requires minimization of the zero-one loss. This mismatch of loss has previously been discussed in literature (Murphy (2005), Qian and Murphy (2011)). Our observation thus hints that bypassing regression may result in better-quality treatment regimes, at least in small samples.

8. Discussion. Our work *is the first step* toward a unified understanding of general surrogate losses in the simultaneous optimization context. Our work leaves ample room for modification and generalization to complex real-world scenarios. We list below some important open questions.

Regarding the optimization error, we have analyzed linear-type treatment policies under conditions with a primary focus on landscape analysis. However, our simulations in Section 7 indicates that DTRESLO performs competitively to popular DTR methods, regardless of whether the policies are linear or nonlinear. Therefore, a more comprehensive analysis of the optimization error is required to gain deeper insight into the performance of DTRESLO.

The theoretical results in this paper consider the propensity scores to be known. They may be available in SMART studies, but they need to be estimated for observational studies. At best, we may be able to estimate the propensity scores at $n^{-1/2}$-rate. Therefore, it is possible that in this situation, our regret-decay rate will slow down. We also do not know if it is at all possible to push the regret decay to $O(1/n)$ in this situation because we do not know the minimax rate of regret-decay in this context. However, there is a more pressing issue with the use of inverse propensity score weighting. The weight will grow smaller as the number of stages increases, leading to a highly volatile method (Kosorok and Laber (2019)). However, there are strategies (Kallus (2018)) that can be incorporated to ensure robustness. Research in this direction is needed to increase the stability of our DTRESLO method.

Also, there are many choices of ϕ's that satisfy Condition 2, and hence can be used for DTRESLO. In this paper, we have not considered the problem of selecting a ϕ. We fixed a particular ϕ in our empirical study but the performance may be improved by a more careful tuning of ϕ.

The DTRESLO method easily extends to $K > 2$ by using a surrogate $\psi(x_1, \ldots, x_K) = \phi(x_1) \ldots \phi(x_K)$. Although we do not yet know whether Fisher consistency still holds, our proof techniques are readily extendable to the higher stages via mathematical induction. If our DTRESLO method is Fisher consistent for general K stages, the pattern of error accumulation over stages will be an immediate interest. For Q-learning, the regret grows exponentially with the number of stages (Murphy (2005)). In view of Wang, Foster and Kakade (2020), exponential error accumulation may sometimes be inevitable under very general conditions. However, we wonder whether our simultaneous maximization procedure escapes the exponential error accumulation in the presence of noise conditions.

Despite being of immense practical interest, this area greatly lacks direct search method with rigorous guarantees in multistage settings. Direct search method with more than two levels of treatment requires integration of multicategory classification with the sequential setting of DTR, and hence is conceptually more challenging than the regression-based counterparts. However, we expect that DTRESLO can be extended to identify optimal DTRs under this more complex setting. Detailed strategies for identifying the surrogate loss and implementing algorithms to estimate DTRs in practice warrant future research.

Acknowledgments. The third and the fourth authors are equal contributors.

Funding. Rajarshi Mukherjee and Nilanjana Laha's research was supported by National Institutes of Health Grant P42ES030990.

Nilanjana Laha's research was also supported by National Science Foundation Grant DMS-2311098.

Tianxi Cai's research was supported by National Institutes of Health Grant R01LM013614.

Aaron Sonabend's research was supported by the Boehringer–Ingelheim Fellowship at Harvard.

SUPPLEMENTARY MATERIAL

Supplement to "Finding the optimal dynamic treatment regimes using smooth Fisher consistent surrogate loss" (DOI: 10.1214/24-AOS2363SUPP; .pdf). The supplementary material contains discussions on optimization error, an application of DTRESLO to Electronic

Health Record (EHR) data, additional details on concave surrogates (particularly hinge loss), further elaboration on the regret decay of DRESLO, a comparison of DTRESLO with related literature, additional details regarding the simulation settings in Section 7, and the proofs of the theorems and lemmas.

REFERENCES

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032 https://doi.org/10.1198/016214505000000907

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026 https://doi.org/10.1162/153244303321897690

BLANCHARD, G., BOUSQUET, O. and MASSART, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36** 489–531. MR2396805 https://doi.org/10.1214/009053607000000839

CALAUZENES, C., USUNIER, N. and GALLINARI, P. (2012). On the (non-) existence of convex, calibrated surrogate losses for ranking. *Adv. Neural Inf. Process. Syst.* **25** 197–205.

CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. Statistics for Biology and Health*. Springer, New York. MR3112454 https://doi.org/10.1007/978-1-4614-7428-9

CHEN, G., ZENG, D. and KOSOROK, M. R. (2016). Personalized dose finding using outcome weighted learning. *J. Amer. Statist. Assoc.* **111** 1509–1521. MR3601705 https://doi.org/10.1080/01621459.2016.1148611

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534 https://doi.org/10.1111/biom.12676

CUI, Y. and TCHETGEN TCHETGEN, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *J. Amer. Statist. Assoc.* **116** 162–173. MR4227683 https://doi.org/10.1080/01621459.2020.1783272

DEMBCZYŃSKI, K., WAEGEMAN, W., CHENG, W. and HÜLLERMEIER, E. (2012). On label dependence and loss minimization in multi-label classification. *Mach. Learn.* **88** 5–45. MR2942603 https://doi.org/10.1007/s10994-012-5285-8

DUCHI, J., KHOSRAVI, K. and RUAN, F. (2018). Multiclass classification, information, divergence and surrogate risk. *Ann. Statist.* **46** 3246–3275. MR3852651 https://doi.org/10.1214/17-AOS1657

DUCHI, J. C., MACKEY, L. W. and JORDAN, M. I. (2010). On the consistency of ranking algorithms. In *ICML*.

FENG, H., NING, Y. and ZHAO, J. (2022). Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. *Ann. Statist.* **50** 2284–2305. MR4474491 https://doi.org/10.1214/22-aos2188

GAO, W. and ZHOU, Z.-H. (2011). On the consistency of multi-label learning. In *Proceedings of the* 24*th Annual Conference on Learning Theory* 341–358.

HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2001). *Fundamentals of Convex Analysis. Grundlehren Text Editions*. Springer, Berlin. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)]. MR1865628 https://doi.org/10.1007/978-3-642-56468-0

HOROWITZ, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60** 505–531. MR1162997 https://doi.org/10.2307/2951582

JIANG, B., SONG, R., LI, J. and ZENG, D. (2019). Entropy learning for dynamic treatment regimes. *Statist. Sinica* **29** 1633–1656. MR3970323

KALLUS, N. (2018). Balanced policy evaluation and learning. *Adv. Neural Inf. Process. Syst.* **31**.

KALLUS, N. (2019). Discussion: "Entropy learning for dynamic treatment regimes" [MR3970323]. *Statist. Sinica* **29** 1697–1705. MR3970330

KARP, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations* (*Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y.*, 1972). *The IBM Research Symposia Series* 85–103. Plenum, New York. MR0378476

KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

LABER, E. B. and DAVIDIAN, M. (2017). Dynamic treatment regimes, past, present, and future: A conversation with experts. *Stat. Methods Med. Res.* **26** 1605–1610. MR3687166 https://doi.org/10.1177/0962280217708661

LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **8** 1225–1272. MR3263118 https://doi.org/10.1214/14-EJS920

LABER, E. B. and ZHAO, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102** 501–514. MR3394271 https://doi.org/10.1093/biomet/asv028

LAHA, N., SONABEND-W, A., MUKHERJEE, R. and CAI, T. (2024). Supplement to "Finding the optimal dynamic treatment regimes using smooth Fisher consistent surrogate loss." https://doi.org/10.1214/24-AOS2363SUPP

LIN, Y. (2004). A note on margin-based loss functions in classification. *Statist. Probab. Lett.* **68** 73–82. MR2064687 https://doi.org/10.1016/j.spl.2004.03.002

LIU, Y. (2007). Fisher consistency of multicategory support vector machines. In *Artificial Intelligence and Statistics* 291–298.

LIU, Y. and SHEN, X. (2006). Multicategory ψ-learning. *J. Amer. Statist. Assoc.* **101** 500–509. MR2256170 https://doi.org/10.1198/016214505000000781

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

MOODIE, E. E., DEAN, N. and SUN, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Stat. Biosci.* **6** 223–243.

MUKHERJEE, D., BANERJEE, M. and RITOV, Y. (2021). Optimal linear discriminators for the discrete choice model in growing dimensions. *Ann. Statist.* **49** 3324–3357. MR4352532 https://doi.org/10.1214/21-aos2085

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

MURPHY, S. A., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586 https://doi.org/10.1198/016214501753382327

NEYKOV, M., LIU, J. S. and CAI, T. (2016). On the characterization of a class of Fisher-consistent loss functions and its application to boosting. *J. Mach. Learn. Res.* **17** Paper No. 70, 32. MR3517093

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *Int. J. Biostat.* **6** Art. 8, 49. MR2602551 https://doi.org/10.2202/1557-4679.1200

PEDREGOSA, F., BACH, F. and GRAMFORT, A. (2017). On the consistency of ordinal regression methods. *J. Mach. Learn. Res.* **18** Paper No. 55, 35. MR3687598

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). *Q*- and *A*-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. MR3300363 https://doi.org/10.1214/13-STS450

SONABEND-W, A. (2022). DTR estimation via surrogate loss. Available at https://github.com/asonabend?tab=repositories. Github code for "Finding the optimal dynamic treatment regimes using smooth Fisher consistent surrogate loss".

SONABEND-W, A., LAHA, N., ANANTHAKRISHNAN, A. N., CAI, T. and MUKHERJEE, R. (2023). Semi-supervised off-policy reinforcement learning and value estimation for dynamic treatment regimes. *J. Mach. Learn. Res.* **24** 86. MR4690272

SONG, R., KOSOROK, M., ZENG, D., ZHAO, Y., LABER, E. and YUAN, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat* **4** 59–68. MR3405390 https://doi.org/10.1002/sta4.78

STEINWART, I. and SCOVEL, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607. MR2336860 https://doi.org/10.1214/009053606000001226

SUN, Y. and WANG, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **116** 421–432. MR4227704 https://doi.org/10.1080/01621459.2020.1819294

TEWARI, A. and BARTLETT, P. L. (2007). On the consistency of multiclass classification methods. *J. Mach. Learn. Res.* **8** 1007–1025. MR2320680 https://doi.org/10.1007/11503415_10

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

WANG, R., FOSTER, D. P. and KAKADE, S. M. (2020). What are the statistical limits of offline RL with linear function approximation? arXiv preprint. Available at arXiv:2010.11895.

WATKINS, C. J. C. H. (1989). Learning from delayed rewards.

XU, T., WANG, J. and FANG, Y. (2014). A model-free estimation for the covariate-adjusted Youden index and its associated cut-point. *Stat. Med.* **33** 4963–4974. MR3276512 https://doi.org/10.1002/sim.6290

XU, Y., MÜLLER, P., WAHED, A. S. and THALL, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J. Amer. Statist. Assoc.* **111** 921–950. MR3561917 https://doi.org/10.1080/01621459.2015.1086353

ZAJONC, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *J. Amer. Statist. Assoc.* **107** 80–92. MR2949343 https://doi.org/10.1080/01621459.2011.643747

ZHANG, J., LIU, T. and TAO, D. (2022). On the rates of convergence from surrogate risk minimizers to the Bayes optimal classifier. *IEEE Trans. Neural Netw. Learn. Syst.* **33** 5766–5774. MR4497102

ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

ZHANG, Y., LABER, E. B., DAVIDIAN, M. and TSIATIS, A. A. (2018). Estimation of optimal treatment regimes using lists. *J. Amer. Statist. Assoc.* **113** 1541–1549. MR3902228 https://doi.org/10.1080/01621459.2017.1345743

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488

ZHOU, X. and KOSOROK, M. R. (2017). Augmented outcome-weighted learning for optimal treatment regimes. arXiv preprint. Available at arXiv:1711.10654.