Electronic Journal of Statistics
Vol. 16 (2022) 737–784
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1948

Estimating individualized treatment rules for treatments with hierarchical structure[*]

Yiwei Fan and Xiaoling Lu

Center for Applied Statistics, School of Statistics, Renmin University of China, China

Junlong Zhao

School of Statistics, Beijing Normal University, China

Haoda Fu

Advanced Analytics and Data Sciences, Eli Lilly and Company, U.S.A.

Yufeng Liu[†]

*Department of Statistics and Operations Research, Department of Genetics,
Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger
Comprehensive Cancer Center, University of North Carolina at Chapel Hill, U.S.A.*
e-mail: yfliu@email.unc.edu

Abstract: Precision medicine is an increasingly important area of research. Due to the heterogeneity of individual characteristics, patients may respond differently to treatments. One of the most important goals for precision medicine is to develop individualized treatment rules (ITRs) involving patients' characteristics directly. As an interesting topic in clinical research, many statistical methods have been developed in recent years to find optimal ITRs. For binary treatments, outcome weighted learning (OWL) was proposed to find a decision function of patient characteristics maximizing the expected clinical outcome. Treatments with hierarchical structure are commonly seen in practice. In hierarchical scenarios, how to estimate ITRs is still unclear. We propose a new framework named hierarchical outcome-weighted angle-based learning (HOAL) to estimate ITRs for treatments with hierarchical structure. Statistical properties including Fisher consistency and convergence rates of the proposed method are presented. Simulations and an application to a type 2 diabetes study under linear and nonlinear learning show the highly competitive performance of our proposed procedure in both numerical accuracy and computational efficiency.

Keywords and phrases: Heterogeneity, hierarchical angle-based classifi-

[*] Xiaoling Lu's research was supported in part by the National Natural Science Foundation of China (No. 72171229) and fund for building world-class universities (disciplines) of Renmin University of China. Junlong Zhao's research was supported in part by National Science Foundation of China, No. 11871104 and No. 12131006. Yufeng Liu's research was supported in part by NSF grants DMS-1821231 and DMS-2100729, and NIH grant R01GM126550.

[†] Corresponding author

cation, precision medicine, support vector machine.

Received October 2020.

Contents

1. Introduction

Precision medicine has received a lot of attention in recent years, originated from the fact that treatment effects manifest heterogeneously among patients due to individual characteristics. Specifically, a treatment that is effective for some patients may fail for others. For instance, in lung cancer, only people having a mutation in the gene EGFR respond to the treatment with tyrosine kinase inhibitors [20]. Similarly, in heart thickening, only people with mutations in the gene GLA respond positively to the enzyme replacement therapy [16]. Thus, one of the most important goals for precision medicine is to develop individualized treatment rules (ITRs) involving patients' characteristics directly.

Many methods have been proposed to design ITRs using statistical tools in the literature, mainly focusing on binary treatments. There is a large body of literature in developing ITRs, by first learning a regression model of outcomes using covariates and then assigning the treatment with the best estimated outcome for a patient given covariates based on this regression model [18, 24]. Instead of directly optimizing the decision rule, these methods obtain ITRs indirectly through the estimated regression model. Qian and Murphy [23] proposed to first estimate the conditional expectation of the response containing a 0–1 loss function and then maximize it to build ITRs. This approach highly depends on whether or not the assumed model is correctly specified.

Besides the indirect methods, there exist direct methods in estimating ITRs. Zhao et al. [37] developed outcome weighted learning (OWL) by treating the ITR problem as a weighted classification problem, where misclassification errors are weighted by clinical outcomes. The 0–1 loss function in Qian and Murphy [23] is replaced by a surrogate hinge loss, and thus the corresponding optimization problem becomes feasible. This approach presented an important idea to use statistical machine learning tools to directly estimate ITRs by maximizing clinical outcomes. Zhou et al. [38] and Liu et al. [15] proposed residual weighted learning, weighting misclassification errors by residuals of the outcome from a regression model on clinical covariates to improve finite sample performance of Zhao et al. [37]. By estimating residuals with generalized linear models, it can deal with different types of outcomes, such as continuous, binary and count data. Under the OWL framework, Chen et al. [5] proposed a data duplication technique with a piecewise convex loss function to estimate ITRs with ordinal treatments, and Zhang et al. [36] estimated optimal ITRs for nominal multicategory treatments, together with variable selection via an l_1 penalty.

Treatments with hierarchical structure are commonly seen in practice. For instance, Pelletier [21] presented the classes of oral diabetic drugs in a tree structure, and Kasi, Ansell and Gertz [12] showed the hierarchy of treatment options for Waldenström Macroglobulinemia, a type of non-Hodgkin lymphoma. Despite the success of OWL in estimating ITRs, how to extend it to the hierarchical setting is still not fully explored. In this paper, we propose a statistical learning framework to deal with ITR estimation in hierarchical treatment scenarios. The hierarchy of treatments can be defined as a graph. A directed edge from node u to node u' means u is a parent of u' and u' is a child of u. A node without any child is referred to a leaf. Each leaf node represents a treatment and we assume any patient is assigned to one treatment. Moreover, we assume each node has at most one parent, where the hierarchy is called a tree structure, and each node either is a leaf or has at least two children.

In the literature of hierarchical classification, applying a flat classifier to the leaf nodes, which ignores the hierarchy, is the simplest method. One popular alternative approach is to sequentially train a multicategory classifier at each parent node [6] or a binary classifier at each node [4], and then predict labels by the top–down strategy. Such an approach suffers from a small training set for each classifier and can be suboptimal. Some other methods have been developed to incorporate hierarchical information in learning classification rules including imposing inequality constraints [30] and designing cost-sensitive learning [8]. A detailed survey of hierarchical classification can be found in Silla and Freitas [28].

Besides the methods mentioned above, several methods on label embedding have been developed for hierarchical classification [3, 29], which map nodes into a set of points in the Euclidean space, such that the Euclidean distances among these points mimic the dissimilarities among the nodes. Recently, Fan et al. [9] has pointed out that the classical label embedding fails to keep the hierarchy well and can be inefficient because of the high dimension of the embedded space. To overcome these drawbacks, Fan et al. [9] proposed a label embedding method, which keeps the hierarchy exactly and reduces the dimension of the embedded space to $m_{\text{leaf}} - 1$ with m_{leaf} being the number of leaf nodes. Despite this method has great advantages in hierarchical classification, how to utilize it in OWL to estimate ITRs for hierarchical treatments remains unclear.

In this paper, we propose a new framework named hierarchical outcome-weighted angle-based learning (HOAL) to solve this problem. We show that for hierarchical treatments, maximizing the expected clinical outcome is equivalent to minimizing a weighted piecewise hierarchical zero–one loss. To assign treatments, we first embed nodes on the hierarchical tree into a series of points in $\mathbb{R}^K$ where $K = m_{\text{leaf}} - 1$. Then we map the covariates of each patient into a vector in $\mathbb{R}^K$ by a learning function and follow the top–down strategy by the angle between this vector and the embedded points.

There are several key contributions in this paper. Firstly, we propose HOAL to estimate ITRs in hierarchical treatment scenarios. The top–down strategy is adopted to assign treatments and an associated hierarchy margin is defined to compare treatment paths on the tree structure. Secondly, we design a linear loss function, under which a closed form solution is derived for both linear and nonlinear learning. Thus, our method can be very computationally efficient. Thirdly, the theoretical properties of the estimators are established.

The remaining of this paper is organized as follows. In Section 2, we review OWL and explain how to extend OWL to hierarchical cases via label embedding. We then introduce a special linear loss function, under which a closed form solution is desired. In Section 3, statistical theories of Fisher consistency and convergence rate are presented. Simulations and real data analysis using both linear and nonlinear learners are conducted in Sections 4 and 5. Section 6 concludes the paper.

2. Methodology

In this section, we first briefly review the framework of OWL and then explain how to extend it to hierarchical treatment scenarios. Furthermore, to reduce computation, we design a linear loss, under which the estimator has a closed form for both linear and nonlinear learning.

2.1. Individualized treatment rule and outcome weighted learning

In individualized treatments, denote $\boldsymbol{X} \in \mathcal{X}$ as the p-dimensional covariate vector for a patient, $A \in \mathcal{A}$ as the corresponding treatment, and $R \in \mathbb{R}$ as the observed outcome, called the "reward". The ITR is a map $\mathcal{D} : \mathcal{X} \to \mathcal{A}$ which assigns treatment $\mathcal{D}(\boldsymbol{X})$ to a patient with covariates $\boldsymbol{X}$. Assuming a larger R is desirable, an optimal ITR is a rule that maximizes the expected reward if implemented. Denote the distribution of $(\boldsymbol{X}, A, R)$ as P and the expectation with respect to P is denoted by E. The likelihood of $(\boldsymbol{X}, A, R)$ under P is then

$$p_0(\boldsymbol{x}) \Pr(a|\boldsymbol{x}) p_1(r|\boldsymbol{x}, a), \tag{2.1}$$

where p_0 is the unknown density of $\boldsymbol{X}$, $\Pr(a|\boldsymbol{x})$ is the probability of receiving treatment a for a patient with covariates $\boldsymbol{x}$, and p_1 is the unknown density of R conditional on $(\boldsymbol{X}, A)$. For any given ITR $\mathcal{D}$, denote $P^{\mathcal{D}}$ as the distribution of $(\boldsymbol{X}, A, R)$ given that $A = \mathcal{D}(\boldsymbol{X})$, that is, the treatments are chosen according to the rule $\mathcal{D}$, and denote $E^{\mathcal{D}}$ as the expectation with respect to $P^{\mathcal{D}}$. Then the likelihood of $(\boldsymbol{X}, A, R)$ under $P^{\mathcal{D}}$ is

$$p_0(\boldsymbol{x}) I(a = \mathcal{D}(\boldsymbol{x})) p_1(r|\boldsymbol{x}, a), \tag{2.2}$$

where $I(\cdot)$ is the indicator function. Under the assumption that $\Pr(a|\boldsymbol{x}) > 0$ for any $a \in \mathcal{A}$, by (2.1) and (2.2), for any subset $\mathscr{S} \subset \mathcal{X} \times \mathcal{A} \times \mathbb{R}$, we have $P^{\mathcal{D}}(\mathscr{S}) = 0$ when $P(\mathscr{S}) = 0$. Thus, $P^{\mathcal{D}}$ is absolutely continuous with respect to P. By the Radon–Nikodym theorem, the Radon–Nikodym derivative $dP^{\mathcal{D}}/dP$ exists and $dP^{\mathcal{D}}/dP = I(a = \mathcal{D}(\boldsymbol{x}))/\Pr(a|\boldsymbol{x})$ by (2.1) and (2.2). Thus, the expected reward given ITR $\mathcal{D}$ is [23]

$$\mathcal{V}(\mathcal{D}) \triangleq E^{\mathcal{D}}(R) = \int R dP^{\mathcal{D}} = \int R \frac{dP^{\mathcal{D}}}{dP} dP = \int R \frac{I(A = \mathcal{D}(\boldsymbol{X}))}{\Pr(A|\boldsymbol{X})} dP.$$

This expectation is called the value function associated with $\mathcal{D}$. One important goal of ITR is to find the optimal $\mathcal{D}^*$ that maximizes $\mathcal{V}(\mathcal{D})$, which is equivalent to defining $\mathcal{D}^*$ as

$$\mathcal{D}^*(\boldsymbol{X}) = \underset{\mathcal{D}}{\operatorname{argmin}} \left\{ E\left(\frac{R \cdot I(A \neq \mathcal{D}(\boldsymbol{X}))}{\Pr(A|\boldsymbol{X})} \right) \right\}. \tag{2.3}$$

For nonnegative rewards, Zhao et al. [37] proposed OWL, utilizing the hinge loss as a convex surrogate loss for the 0–1 loss $I(\cdot)$. Then, (2.3) can be viewed as a weighted classification problem. In practice, when there are negative rewards, one can replace R by $R + \rho$ for any constant ρ to ensure nonnegativeness, while such a constant shift process for the rewards may lead to suboptimal estimates

[5]. To better handle the case with negative rewards, Chen et al. [5] proposed the following formulation,

$$\mathcal{D}^*(\boldsymbol{X}) \\ = \underset{\mathcal{D}}{\operatorname{argmin}} E\left[\frac{|R|}{\Pr(A|\boldsymbol{X})}\{I(R \geq 0)I(A \neq \mathcal{D}(\boldsymbol{X})) + I(R < 0)I(A = \mathcal{D}(\boldsymbol{X}))\}\right]. \tag{2.4}$$

Note that (2.4) is equivalent to (2.3) as the term $R \cdot I(R < 0)/\Pr(A|\boldsymbol{X})$ is free of $\mathcal{D}(\boldsymbol{X})$ [5]. The loss in (2.4) has two parts by the sign of R. For nonnegative rewards, we penalize the inequality to A. For negative rewards, we encourage the optimal ITR to move away from A. Interestingly, (2.4) can also be viewed as a weighted misclassification error, for which we weigh each misclassification event by $|R|/\Pr(A|\boldsymbol{X})$.

2.2. Outcome weighted learning for hierarchical treatments

In this subsection, we discuss how to extend OWL to hierarchical treatment scenarios. The hierarchy of treatments is described by a tree, where each leaf represents a treatment. There are m_{leaf} leaf nodes accordingly to m_{leaf} treatments with a hierarchical structure among them.

We first introduce some notations. For a node, denote its parent, children, ancestors, offsprings, and siblings respectively as $\text{Par}(\cdot)$, $\text{Chi}(\cdot)$, $\text{Anc}(\cdot)$, $\text{Off}(\cdot)$, and $\text{Sib}(\cdot)$. The total number of layers of the tree is denoted as k. The root node at the first layer is denoted as T_1, which is meaningless. Moreover, T_{1,j_2} is the child of T_1 with index $j_2 = 1, 2, \ldots, N_1$ at the second layer from left to right, where N_1 is the number of children for T_1. In general, for $3 \leq m \leq k$, $T_{1,j_2,\ldots,j_{m-1},j_m}$ is the child of $T_{1,j_2,\ldots,j_{m-1}}$ with index $j_m = 1, 2, \ldots, N_{1,j_2,\ldots,j_{m-1}}$ at the m-th layer from left to right, where $N_{1,j_2,\ldots,j_{m-1}}$ is the number of children for $T_{1,j_2,\ldots,j_{m-1}}$. For example, FIG 1 (left panel) presents a hierarchical structure with three layers. At the first layer, the root node is denoted by T_1. The number of children for T_1 is denoted by N_1, which takes the value 2. These two children for T_1, which are located at the second layer, are denoted by $T_{1,1}$ and $T_{1,2}$. Furthermore, the number of children for $T_{1,1}$ is denoted by $N_{1,1}$, taking the value 2. In particular, $T_{1,1}$ has two children at the third layer denoted by $T_{1,1,1}$ and $T_{1,1,2}$.

Denote the collection of all nodes except for the root as

$$\mathcal{T} = \bigcup_{m=2}^{k} \mathcal{T}_m = \bigcup_{m=2}^{k} \{T_{j_1,j_2,\ldots,j_m} : j_1 \equiv 1, j_s = 1, \ldots, N_{j_1,\ldots,j_{s-1}}, s = 2, \ldots, m\},$$

where $\mathcal{T}_m$ is the set of nodes at the m-th layer. In hierarchical treatment scenarios, the treatment space $\mathcal{A}$ is a set of paths, where each path, denoted by A, is from the root to a leaf on the tree. Specifically, $A = \{A^{(1)}, \ldots, A^{(\mathcal{L}(A))}\} \in \mathcal{A}$, where $A^{(m)} \in \text{Chi}(A^{(m-1)}) \subset \mathcal{T}_m$ indicates the node at the m-th layer for $m =$

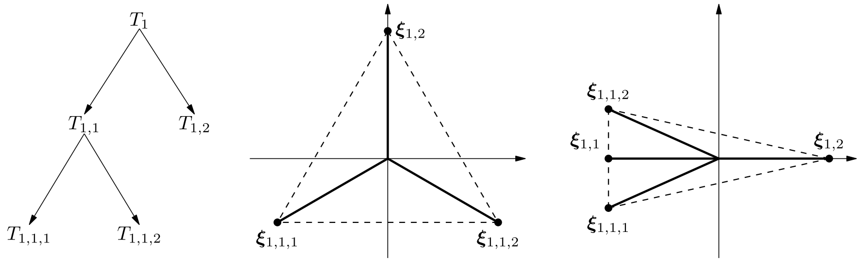

FIG 1. *The hierarchical structure (left panel), the embedded points for the standard multicategory classification considering only leaf nodes (middle panel) and the embedded points for hierarchical classification (right panel) of an illustrative example.*

$2, \dots, \mathcal{L}(A)$ and $\mathcal{L}(A)$ is the layer where the leaf locates. The ITR $\mathcal{D} : \mathcal{X} \to \mathcal{A}$ assigns a patient with covariates $\boldsymbol{X}$ to a treatment path $\mathcal{D}(\boldsymbol{X})$ on the tree with length $\mathcal{L}(\mathcal{D})$. Note that $A = \mathcal{D}(\boldsymbol{X})$ if and only if $\mathcal{L}(A) = \mathcal{L}(\mathcal{D})$ and $A^{(m)} = \mathcal{D}^{(m)}$ for $m = 2, \dots, \mathcal{L}(A)$.

Although there are many methods proposed to estimate ITR for binary and multicategory treatments, the extension to hierarchical treatments is nontrivial. To obtain the optimal ITR for hierarchical treatments, we propose hierarchical outcome-weighted angle-based learning. We adopt angle-based hierarchical classification to incorporate the hierarchical structure among treatments. Angle-based hierarchical classification first maps nodes into a set of points in the Euclidean space. Let $\boldsymbol{\xi}_{j_1,j_2,\dots,j_m}$ be the embedded point associated with $T_{j_1,j_2,\dots,j_m}$. Denote

$$\mathcal{E} = \bigcup_{m=2}^{k} \mathcal{E}_m = \bigcup_{m=2}^{k} \{\boldsymbol{\xi}_{j_1,j_2,\dots,j_m} : j_1 \equiv 1, j_s = 1, \dots, N_{j_1,j_2,\dots,j_{s-1}}, s = 2, \dots, m\},$$

where $\mathcal{E}_m$ is the set of points at the m-th layer. A desired embedding approach is to embed nodes into points in a low-dimensional space while keeping the hierarchical properties exactly. To achieve these two goals, Fan et al. [9] proposed a new label embedding method, which is summarized as follows.

To construct points in hierarchical classification, Fan et al. [9] first introduced an algorithm to embed nodes in a standard q-class multicategory classification problem. Algorithm S1 in the Appendix gives the detailed procedure. The embedded points form a simplex in $\mathbb{R}^{q-1}$, which is centered at the origin. Since the sum of the embedded points is a zero-vector, there is no need to include the explicit sum-to-zero constraint, which is required in the regular simultaneous multicategory classification [34]. Hence, the computational costs can be greatly reduced. Take FIG 1 as an illustrative example. If we consider only leaf nodes $T_{1,2}, T_{1,1,1}$ and $T_{1,1,2}$, it is a standard 3-class multicategory classification problem. We start from two points -1 and 1 in $\mathbb{R}$. Then we extend these two points into $\mathbb{R}^2$ as $(-1, 0)^\top$ and $(1, 0)^\top$. Furthermore, we construct the third point $(0, \sqrt{3})^\top$ satisfying the equal pairwise distance requirement. So far, these three points form an equilateral triangle centered at $(0, \sqrt{3}/3)^\top$. After that, we centralize these points as $(-1, -\sqrt{3}/3)^\top$, $(1, -\sqrt{3}/3)^\top$ and $(0, 2\sqrt{3}/3)^\top$. Finally, given length L, we scale them and obtain $\boldsymbol{\xi}_{1,1,1} = (-L\sqrt{3}/2, -L/2)^\top$, $\boldsymbol{\xi}_{1,1,2} = (L\sqrt{3}/2, -L/2)^\top$ and $\boldsymbol{\xi}_{1,2} = (0, L)^\top$. The embedded points are shown in the middle panel of FIG 1. However, the distance between each pair of points is same, violating the rule that $\boldsymbol{\xi}_{1,1,1}$ and $\boldsymbol{\xi}_{1,1,2}$ should be closer.

In hierarchical scenarios, the embedded points are located in $\mathbb{R}^K$ with $K = m_{\text{leaf}} - 1$. We begin from constructing points for the nodes at the second layer by applying Algorithm S1 with a given length in a subspace of $\mathbb{R}^K$. For children nodes at the m-th ($m \geq 3$) layer, we inherit the coordinates from their parents as the first part and construct the points by applying Algorithm S1 in another subspace of $\mathbb{R}^K$ as the second part. The two parts are then concatenated. The details are referred to Algorithm S2 in the Appendix. As shown in Proposition S1 in the Appendix, the embedded points satisfy hierarchical and symmetric (H. S.) properties such that the Euclidean distance between the embedded points can exactly mimic the dissimilarities between the nodes. Another advantage is that the dimension of embedded points is $K = m_{\text{leaf}} - 1$, which is the same as the one required for standard multicategory classification considering only leaf nodes [13, 34]. Thus, this method enjoys the advantages of both aspects, keeping the hierarchy exactly and involving a low-dimensional label space. Denoting the length of the embedded points for the nodes at the $(m+1)$-th layer by $L^{(m)}$, we set $L^{(m+1)} = L^{(m)}/\delta$ in Algorithm S2, where $\delta > 1$ is the down-scaling constant. To satisfy H. S. properties, it is required that $\delta^2 \geq 2\sqrt{2} + 2$ in Proposition S1. We set $L^{(1)} = 1$ and $\delta = \sqrt{5}$ as suggested by Fan et al. [9]. For the example shown in FIG 1, we begin from the second layer and construct two points -1 and 1 by Algorithm S1 given $L^{(1)} = 1$. Then we extend these two points into $\mathbb{R}^2$ by setting $\boldsymbol{\xi}_{1,1} = (-1, 0)^\top$ and $\boldsymbol{\xi}_{1,2} = (1, 0)^\top$ for $T_{1,1}$ and $T_{1,2}$. Furthermore, for $T_{1,1}$, we construct points for its children $T_{1,1,1}$ and $T_{1,1,2}$. We inherit the first coordinate of $\boldsymbol{\xi}_{1,1}$, and apply Algorithm S1 to construct $-1/\sqrt{5}$ and $1/\sqrt{5}$ given $L^{(2)} = L^{(1)}/\delta = 1/\sqrt{5}$ as the second coordinates of $\boldsymbol{\xi}_{1,1,1}$ and $\boldsymbol{\xi}_{1,1,2}$. Therefore, we have $\boldsymbol{\xi}_{1,1,1} = (-1, -1/\sqrt{5})^\top$ and $\boldsymbol{\xi}_{1,1,2} = (-1, 1/\sqrt{5})^\top$. The embedded points are shown in the right panel of FIG 1, which keep the hierarchical structure. The distance between $\boldsymbol{\xi}_{1,1,1}$ and $\boldsymbol{\xi}_{1,1,2}$ is much smaller than the distance between $\boldsymbol{\xi}_{1,1,1}$ and $\boldsymbol{\xi}_{1,2}$. Moreover, the distance between $\boldsymbol{\xi}_{1,1,1}$ and $\boldsymbol{\xi}_{1,2}$ is equal to the distance between $\boldsymbol{\xi}_{1,1,2}$ and $\boldsymbol{\xi}_{1,2}$. Indeed, $\boldsymbol{\xi}_{1,1,1}, \boldsymbol{\xi}_{1,1,2}$ and $\boldsymbol{\xi}_{1,2}$ form an isosceles triangle.

After label embedding, we map the covariates of each patient into a vector in $\mathbb{R}^K$ by some decision function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^K$. Denote $\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}) \in \mathcal{A}$ as the ITR associated with $\boldsymbol{f}$. We determine $\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})$ by the following top–down strategy, the most commonly used strategy in hierarchical scenarios [4, 31]. For any treatment path $A \in \mathcal{A}$ and $m = 2, \ldots, \mathcal{L}(A)$, let $\boldsymbol{\xi}_m(A)$ be the embedded point corresponding to A at the m-th layer and $\mathscr{E}_m(A)$ be the set of paths, where the m-th element of each path is one of the siblings of the m-th element of A. For any $m = 2, \ldots, k$, assuming $\boldsymbol{x}$ has been assigned to a node at the $(m-1)$-th layer, the top–down strategy assigns $\boldsymbol{x}$ to its child node, of which the corresponding embedded point has the largest inner product with $\boldsymbol{f}(\boldsymbol{x})$ among all children.

Definition 1 (top–down strategy)**.** *Let $\mathcal{D}_{\boldsymbol{f}}^{(1)}(\boldsymbol{x}) \equiv T_1$. Suppose $\boldsymbol{x}$ has been assigned with $\mathcal{D}_{\boldsymbol{f}}^{(m-1)}(\boldsymbol{x})$ at the $(m-1)$-th layer, then $\boldsymbol{x}$ is assigned to $\mathcal{D}_{\boldsymbol{f}}^{(m)}(\boldsymbol{x})$ at the m-th layer, if $\mathcal{D}_{\boldsymbol{f}}^{(m)}(\boldsymbol{x}) \in Chi(\mathcal{D}_{\boldsymbol{f}}^{(m-1)}(\boldsymbol{x}))$ and*

$$\langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\xi}_m(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))\rangle > \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\xi}_m(\tilde{A})\rangle, \tag{2.5}$$

for any $\tilde{A} \in \mathscr{E}_m(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})) = \{A : A^{(m)} \neq \mathcal{D}_{\boldsymbol{f}}^{(m)}(\boldsymbol{x}), A^{(m)} \in Chi(\mathcal{D}_{\boldsymbol{f}}^{(m-1)}(\boldsymbol{x}))\}$, where $\langle\cdot,\cdot\rangle$ is the inner product between two vectors.

As $\boldsymbol{\xi}_m(\tilde{A})$ depends only on $\tilde{A}^{(m)}$, it is possible that there are multiple $\tilde{A}$ with the same node at the m-th layer. If this is the case, by taking only one of them as the representative and denoting the set of representatives as $[\mathscr{E}_m(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))]$, we only need to require that (2.5) holds for any $\tilde{A} \in [\mathscr{E}_m(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))]$. Taking the hierarchical tree shown in FIG 1 as an example, for a learning function $\boldsymbol{f}$, let $\mathcal{D}_{\boldsymbol{f}}^{(1)}(\boldsymbol{x}) \equiv T_1$. At the second layer, $\boldsymbol{x}$ is assigned to $T_{1,2}$ if $\langle \boldsymbol{f}, \boldsymbol{\xi}_{1,2}\rangle > \langle \boldsymbol{f}, \boldsymbol{\xi}_2(\tilde{A})\rangle$ for any $\tilde{A} \in \mathscr{E}_2(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))$ with $\mathscr{E}_2(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})) = \{\{T_1, T_{1,1}, T_{1,1,1}\}, \{T_1, T_{1,1}, T_{1,1,2}\}\}$. Note that $\{T_1, T_{1,1}, T_{1,1,1}\}$ and $\{T_1, T_{1,1}, T_{1,1,2}\}$ have the same node $T_{1,1}$ at the second layer, thus it is sufficient to take only one of them as the representative. Let $[\mathscr{E}_2(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))] = \{\{T_1, T_{1,1}, T_{1,1,1}\}\}$ or $[\mathscr{E}_2(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))] = \{\{T_1, T_{1,1}, T_{1,1,2}\}\}$. To assign $\boldsymbol{x}$ to $T_{1,2}$, we only need to require $\langle \boldsymbol{f}, \boldsymbol{\xi}_{1,2}\rangle > \langle \boldsymbol{f}, \boldsymbol{\xi}_2(\tilde{A})\rangle$ for any $\tilde{A} \in [\mathscr{E}_2(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))]$. Otherwise, $\boldsymbol{x}$ is assigned to $T_{1,1}$. Suppose that $\boldsymbol{x}$ has been assigned to $T_{1,1}$, then at the third layer, $\boldsymbol{x}$ is assigned to $T_{1,1,1}$ if $\langle \boldsymbol{f}, \boldsymbol{\xi}_{1,1,1}\rangle > \langle \boldsymbol{f}, \boldsymbol{\xi}_3(\tilde{A})\rangle$ for any $\tilde{A} \in \mathscr{E}_3(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}))$ with $\mathscr{E}_3(\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})) = \{\{T_1, T_{1,1}, T_{1,1,2}\}\}$. Otherwise, $\boldsymbol{x}$ is assigned to $T_{1,1,2}$.

By the top–down strategy, any ITR $\mathcal{D}(\boldsymbol{x})$ can always be represented as $\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})$ for some decision function $\boldsymbol{f}$. For example, let $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\xi}_{\mathcal{L}(\mathcal{D})}(\mathcal{D}(\boldsymbol{x}))$, which is the embedded point corresponding to the leaf node of $\mathcal{D}(\boldsymbol{x})$. One can verify that $\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x}) = \mathcal{D}(\boldsymbol{x})$ according to the top–down strategy. Moreover, note that $A = \mathcal{D}_{\boldsymbol{f}}$ is equivalent to a series of inequalities

$$\langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(A)\rangle - \langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(\tilde{A})\rangle \geq 0, \quad \tilde{A} \in [\mathscr{E}_m(A)], m = 2, \ldots, \mathcal{L}(A). \tag{2.6}$$

Define the hierarchy margin $M(\boldsymbol{f}(\boldsymbol{X}), A)$ as

$$\begin{aligned} M(\boldsymbol{f}(\boldsymbol{X}), A) &= \min_{m=2,\ldots,\mathcal{L}(A)} \{\langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(A)\rangle - \max_{\tilde{A}\in[\mathscr{E}_m(A)]} \langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(\tilde{A})\rangle\} \\ &= \min_{m,\tilde{A}\in[\mathscr{E}_m(A)]} \{\langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(A)\rangle - \langle \boldsymbol{f}(\boldsymbol{X}), \boldsymbol{\xi}_m(\tilde{A})\rangle\}. \end{aligned}$$

It can be seen that (2.6) holds if and only if $M(\boldsymbol{f}(\boldsymbol{X}), A) \geq 0$. Therefore, we have $I(A = \mathcal{D}_{\boldsymbol{f}}(\boldsymbol{X})) = I(M(\boldsymbol{f}(\boldsymbol{X}), A) \geq 0)$ and $I(A \neq \mathcal{D}_{\boldsymbol{f}}(\boldsymbol{X})) = I(M(\boldsymbol{f}(\boldsymbol{X}), A) < 0)$. For any ITR $\mathcal{D} = \mathcal{D}_{\boldsymbol{f}}$ associated with some decision function $\boldsymbol{f}$, based on

(2.4), we have the following optimization problem

$$\begin{aligned}\bar{\boldsymbol{f}}(\boldsymbol{X}) = \operatorname*{arginf}_{\boldsymbol{f}} E\bigg[\frac{|R|}{\Pr(A|\boldsymbol{X})}\{&I(R \geq 0)I(M(\boldsymbol{f}(\boldsymbol{X}), A) < 0)+ \\ &I(R < 0)I(M(\boldsymbol{f}(\boldsymbol{X}), A) \geq 0)\}\bigg],\end{aligned} \tag{2.7}$$

where $\bar{\boldsymbol{f}}$ is the Bayes rule. Recall the optimization problem (2.4), which aims to find the optimal ITR. For simplicity, let

$$\mathcal{C}(\mathcal{D}) = E\left[\frac{|R|}{\Pr(A|\boldsymbol{X})}\{I(R \geq 0)I(A \neq \mathcal{D}(\boldsymbol{X})) + I(R < 0)I(A = \mathcal{D}(\boldsymbol{X}))\}\right].$$

Proposition 1 shows that the associated ITR $\mathcal{D}_{\bar{\boldsymbol{f}}}$ yielded by $\bar{\boldsymbol{f}}$ is a minimizer of $\mathcal{C}(\mathcal{D})$. The proof is given in the Appendix. Therefore, solving the optimization problem (2.4) is equivalent to solving (2.7).

Proposition 1. *The associated ITR $\mathcal{D}_{\bar{\boldsymbol{f}}}$ yielded by $\bar{\boldsymbol{f}}$ according to the top–down strategy is a minimizer of $\mathcal{C}(\mathcal{D})$.*

Remark 1. *Since $\langle \boldsymbol{f}, \boldsymbol{\xi}_m(A)\rangle \geq \langle \boldsymbol{f}, \boldsymbol{\xi}_m(\tilde{A})\rangle \Leftrightarrow \langle \rho\boldsymbol{f}, \boldsymbol{\xi}_m(A)\rangle \geq \langle \rho\boldsymbol{f}, \boldsymbol{\xi}_m(\tilde{A})\rangle$ for any $\rho > 0$, the minimizer $\bar{\boldsymbol{f}}$ of $\mathcal{R}(\boldsymbol{f})$ is not unique. Here $\boldsymbol{f}$ is identifiable up to a scale. To learn the optimal ITR, only the direction of $\boldsymbol{f}$ is required. We emphasize that the aim is to learn the optimal ITR rather than the optimal decision function. To ensure identifiability of $\boldsymbol{f}$, we can impose a constraint on the norm of $\boldsymbol{f}$ (e.g., $\|\boldsymbol{f}\| = 1$ with $\|\cdot\|$ being the l_2 norm).*

Let $\{(\boldsymbol{x}_i, a_i, r_i)\}_{i=1}^n$ be the observations of $(\boldsymbol{X}, A, R)$ with n being the sample size, the empirical loss is

$$\frac{1}{n}\sum_{i=1}^n \frac{|r_i|}{\Pr(a_i|\boldsymbol{x}_i)}\{I(r_i \geq 0)I(M(\boldsymbol{f}(\boldsymbol{x}_i), a_i) < 0) + I(r_i < 0)I(M(\boldsymbol{f}(\boldsymbol{x}_i), a_i) \geq 0)\},$$

which is hard to minimize because of the discontinuity of the 0–1 loss. Applying a surrogate loss ℓ, the objective function can be formulated as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \frac{|r_i|}{\Pr(a_i|\boldsymbol{x}_i)}\{I(r_i \geq 0)\ell(M(\boldsymbol{f}(\boldsymbol{x}_i), a_i)) + I(r_i < 0)\ell(-M(\boldsymbol{f}(\boldsymbol{x}_i), a_i))\} \\ =&\frac{1}{n}\sum_{i=1}^n \frac{|r_i|}{\Pr(a_i|\boldsymbol{x}_i)}\ell_{r_i}(M(\boldsymbol{f}(\boldsymbol{x}_i), a_i)),\end{aligned} \tag{2.8}$$

where $\ell_{r_i}(u) = \ell(u)$ if $r_i \geq 0$ and $\ell_{r_i}(u) = \ell(-u)$ if $r_i < 0$. As the partial derivative of the hierarchy margin $M(\boldsymbol{f}(\boldsymbol{x}_i, a_i))$ is discontinuous, solving (2.8) is still difficult. We overcome this challenge by replacing the term

$$|r_i|\ell_{r_i}(M(\boldsymbol{f}(\boldsymbol{x}_i), a_i))/\Pr(a_i|\boldsymbol{x}_i),$$

with its additive formulation

$$\sum_{m=2}^{\mathcal{L}(a_i)} \sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} |r_i|\ell_{r_i}((\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top \boldsymbol{f}(\boldsymbol{x}_i))/\Pr(a_i^{(m)}|\boldsymbol{x}_i),$$

where $\Pr(a_i^{(m)}|\boldsymbol{x}_i)$ is the propensity score, defined as $\sum_{A\in\mathcal{A}:A^{(m)}=a_i^{(m)}}\Pr(A|\boldsymbol{x}_i)$. For each observation, at the m-th layer, we compare $\boldsymbol{\xi}_m(a_i)^\top \boldsymbol{f}(\boldsymbol{x}_i)$ with the remaining components $\boldsymbol{\xi}_m(\tilde{a})^\top \boldsymbol{f}(\boldsymbol{x}_i)$. The hierarchy margin $M(\boldsymbol{f}(\boldsymbol{x}_i), a_i)$ considers the minimum pairwise difference, while after replacement, the additive formulation considers the sum of all pairwise differences. The spirit is similar to that in Zhang [33]. In Subsection 3.1, we show that after replacement, the theoretical minimizer leads to the same ITR as the Bayes rule $\bar{\boldsymbol{f}}$, which guarantees the rationality of this replacement. Furthermore, to avoid overfitting, we add a penalty term to control the complexity of the learning function. Specifically, instead of minimizing (2.8), we aim to solve the following optimization problem,

$$\operatorname*{argmin}_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]}\frac{|r_i|}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)}\ell_{r_i}((\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top \boldsymbol{f}(\boldsymbol{x}_i)) + \lambda J(\boldsymbol{f}), \tag{2.9}$$

or equivalently,

$$\begin{aligned}&\operatorname*{argmin}_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]}\frac{|r_i|}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)}\ell_{r_i}((\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top \boldsymbol{f}(\boldsymbol{x}_i))\\ &s.t. \quad J(\boldsymbol{f}) \le s_\lambda,\end{aligned}$$

where $\mathcal{F}$ is the functional space (e.g. the set of linear functions for linear learning), $J(\boldsymbol{f})$ is the penalty term, and λ, s_λ are tuning parameters. The minimizer of (2.9) is denoted as $\hat{\boldsymbol{f}}_\lambda$.

Remark 2. *In the next subsection, we introduce the special linear loss $\ell(u) = -u$ to reduce computation. For the linear loss, a restriction $E(\|\boldsymbol{f}\|^2) \le 1$ is required in (2.9) to keep consistency with the theoretical assumptions. In this case, we restrict $\mathcal{F}$ on the set $\{\boldsymbol{f} : E(\|\boldsymbol{f}\|^2) \le 1\}$. The restricted optimization problem is difficult to solve and has the same solution as the unrestricted one when λ is large. Therefore, for more efficient computation, we consider only the unrestricted optimization problem in implementation.*

2.3. Linear loss functions

To reduce computation, we adopt the linear loss $\ell(u) = -u$ [25, 9], where a closed form estimator can be derived for both linear and nonlinear learning.

We first consider the linear learning function $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{b}$, where $\boldsymbol{C} \in \mathbb{R}^{K\times p}, \boldsymbol{b} \in \mathbb{R}^K$. The regularization term is chosen as $J(\boldsymbol{f}) = \|\boldsymbol{C}\|_F^2 + \lambda'\|\boldsymbol{b}\|^2$, where $\|\cdot\|_F$ represents the Frobenius norm and $\|\cdot\|$ is the l_2 norm [7, 35]. As shown in the Appendix, the problem (2.9) has a closed form solution,

$$\hat{\boldsymbol{C}}_{\text{lin},\lambda} = -\boldsymbol{B}_1/(2\lambda), \quad \hat{\boldsymbol{b}}_{\text{lin},\lambda,\lambda'} = -\tilde{\boldsymbol{b}}_1/(2\lambda\lambda'), \tag{2.10}$$

where $\boldsymbol{B}_1 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} w_{i,m}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))\boldsymbol{x}_i^\top$ and

$$\tilde{\boldsymbol{b}}_1 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} w_{i,m}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i)), \tag{2.11}$$

with $w_{i,m} = r_i/\Pr(a_i^{(m)}|\boldsymbol{x}_i)$. Note that for two siblings $\boldsymbol{\xi}_{j_1,j_2,\ldots,j_m}$, $\boldsymbol{\xi}_{j_1,j_2,\ldots,j'_m}$ and any $\rho > 0$, we have

$$\langle \boldsymbol{\xi}_{j_1,j_2,\ldots,j_m}, \boldsymbol{f}\rangle \le \langle \boldsymbol{\xi}_{j_1,j_2,\ldots,j'_m}, \boldsymbol{f}\rangle \Leftrightarrow \langle \boldsymbol{\xi}_{j_1,j_2,\ldots,j_m}, \rho\boldsymbol{f}\rangle \le \langle \boldsymbol{\xi}_{j_1,j_2,\ldots,j'_m}, \rho\boldsymbol{f}\rangle.$$

Thus, the value of ρ does not affect the assigned nodes following the top–down strategy. Since the estimator under the linear loss is proportional to λ^{-1}, we simply set $\lambda = 1$ and denote the corresponding solution as $\hat{\boldsymbol{C}}_{\text{lin}}$ and $\hat{\boldsymbol{b}}_{\text{lin},\lambda'}$. Thus, only the tuning parameter λ' needs to be tuned.

The linear loss uses the same weight for all observations. It awards the correctly assigned instances and is unbounded, thus may be not robust. The plot of the linear loss is shown in FIG 2. For any observation with $(\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f} < 0$, the predicted treatment is away from its current received treatment a_i. The smaller $(\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f}$ is, the larger distant it is from a_i and the more likely it is an outlier. On the other hand, for observations with $(\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f} > 0$, it is assigned to $a_i^{(m)}$ by $\boldsymbol{f}$. The linear loss awards these instances and when $(\boldsymbol{\xi}_m(a_i) - \boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f}$ is large, the classifier tends to be strongly affected by them because the linear loss is unbounded. To alleviate the impact of outliers and restrict awards of correctly assigned instances, we assign them smaller weights [32]. Specifically, we design an adaptive weighted linear loss and the optimization problem is

$$\begin{aligned}\text{argmin}\frac{1}{n}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} \frac{|r_i|}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)}\mu_{i,m,\tilde{a}}\ell_{r_i}((\boldsymbol{\xi}_m(a_i) - \boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f}(\boldsymbol{x}_i))+\\ \lambda\|\boldsymbol{C}\|_F^2 + \lambda\lambda'\|\boldsymbol{b}\|^2,\end{aligned} \tag{2.12}$$

where $\mu_{i,m,\tilde{a}} = 1/(1 + |(\boldsymbol{\xi}_m(a_i) - \boldsymbol{\xi}_m(\tilde{a}))^\top(\hat{\boldsymbol{C}}_{\text{lin}}\boldsymbol{x}_i + \hat{\boldsymbol{b}}_{\text{lin},\lambda'})|^\gamma)$ is the adaptive weight and γ can be chosen via validation. Our motivation of the adaptive weight comes from the form of the linear loss $\ell(u) = -u$. We consider the weight function $1/(1 + |u|)$, which is decreasing when $|u|$ is increasing [32]. The plot of the weight function is shown in FIG 2. Using the estimators from the linear loss, $\mu_{i,m,\tilde{a}}$ assigns a smaller weight for any observation with a larger $|(\boldsymbol{\xi}_m(a_i) - \boldsymbol{\xi}_m(\tilde{a}))^\top(\hat{\boldsymbol{C}}_{\text{lin}}\boldsymbol{x}_i + \hat{\boldsymbol{b}}_{\text{lin},\lambda'})|$, which matches with our goal. To further

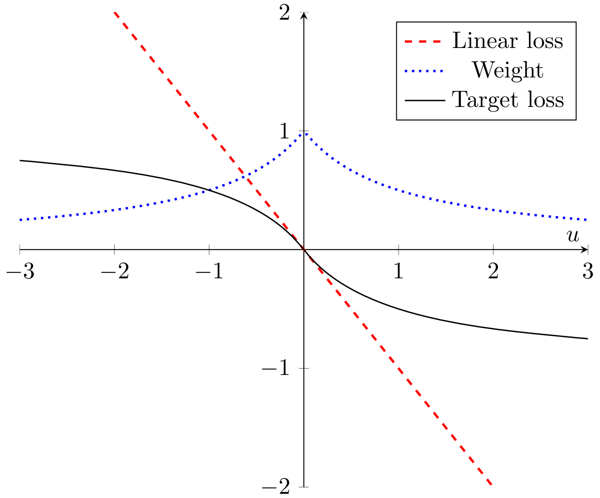

FIG 2. *Plot of the linear loss* $-u$*, the weight function* $1/(1+|u|)$*, and the target function* $-u/(1+|u|)$*.*

understand the weight function, we display the function $-u/(1+|u|)$ in FIG 2, which can be viewed as the target loss through our weighted learning. It can be seen that $-u/(1+|u|)$ is bounded between -1 and 1.

The estimator under the weighted linear loss also has a closed form. The proof is similar to that of the linear loss and is omitted. Setting $\lambda = 1$, the solution of (2.12) is $\hat{\boldsymbol{C}}_{\text{ada}} = -\boldsymbol{B}_2/2$, $\hat{\boldsymbol{b}}_{\text{ada},\lambda'} = -\tilde{\boldsymbol{b}}_2/(2\lambda')$, where

$$\boldsymbol{B}_2 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} \widetilde{w}_{i,m,\tilde{a}}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))\boldsymbol{x}_i^\top,$$

and

$$\tilde{\boldsymbol{b}}_2 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} \widetilde{w}_{i,m,\tilde{a}}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i)),$$

with $\widetilde{w}_{i,m,\tilde{a}} = w_{i,m}\mu_{i,m,\tilde{a}}$.

Next, we consider nonlinear learning. Define a kernel $h(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ that is continuous, symmetric, and positive semidefinite. Let

$$\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{Z}(h(\boldsymbol{X},\boldsymbol{x}_1),\ldots,h(\boldsymbol{X},\boldsymbol{x}_n))^\top + \boldsymbol{b},$$

where $\boldsymbol{Z} \in \mathbb{R}^{K\times n}$ and $\boldsymbol{b} \in \mathbb{R}^K$. Set $J(\boldsymbol{f}) = tr(\boldsymbol{Z}\boldsymbol{\mathcal{H}}\boldsymbol{Z}^\top) + \lambda'\|\boldsymbol{b}\|^2$ with $\boldsymbol{\mathcal{H}} = (h(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1}^n$.

Similarly, the kernel estimator under the linear loss has a closed form. For $\lambda = 1$, we have $\hat{\boldsymbol{Z}}_{\text{lin}} = -\boldsymbol{B}_3\boldsymbol{\mathcal{H}}^{-1}/2, \hat{\boldsymbol{b}}_{\text{lin},\lambda'} = -\tilde{\boldsymbol{b}}_1/(2\lambda')$, where

$$\boldsymbol{B}_3 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} w_{i,m}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))(h(\boldsymbol{x}_i,\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_i,\boldsymbol{x}_n)),$$

and $\tilde{\boldsymbol{b}}_1$ is defined in (2.11).

For $\lambda = 1$, the solution under the adaptive weighted linear loss in the kernel space is $\hat{\boldsymbol{Z}}_{\mathrm{ada}} = -\boldsymbol{B}_4\boldsymbol{\mathcal{H}}^{-1}/2, \hat{\boldsymbol{b}}_{\mathrm{ada},\lambda'} = -\tilde{\boldsymbol{b}}_3/(2\lambda')$, where

$$\boldsymbol{B}_4 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} \underset{\sim}{w}_{i,m,\tilde{a}}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))(h(\boldsymbol{x}_i,\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_i,\boldsymbol{x}_n)),$$

and

$$\tilde{\boldsymbol{b}}_3 = n^{-1}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]} \underset{\sim}{w}_{i,m,\tilde{a}}(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i)),$$

with $\underset{\sim}{w}_{i,m,\tilde{a}} = w_{i,m}/(1 + |(\boldsymbol{\xi}_m(a_i) - \boldsymbol{\xi}_m(\tilde{a}))^\top\{\hat{\boldsymbol{Z}}_{\mathrm{lin}}(h(\boldsymbol{x}_i,\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_i,\boldsymbol{x}_n))^\top + \hat{\boldsymbol{b}}_{\mathrm{lin},\lambda'}\}|^\gamma)$.

As a comparison, we also apply the hinge loss, $\ell(u) = (1-u)_+$ for both linear and nonlinear learning functions. The optimization problem can be written as

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sum_{m=2}^{\mathcal{L}(a_i)}\sum_{\tilde{a}\in[\mathscr{E}_m(a_i)]}\frac{|r_i|}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)}\Big\{&I(r_i\geq 0)(1-(\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f}(\boldsymbol{x}_i))_+\\ &+I(r_i<0)(1+(\boldsymbol{\xi}_m(a_i)-\boldsymbol{\xi}_m(\tilde{a}))^\top\boldsymbol{f}(\boldsymbol{x}_i))_+\Big\}+\lambda J(\boldsymbol{f}),\end{aligned}$$

where $J(\boldsymbol{f}) = \|\boldsymbol{C}\|_F^2 + \lambda'\|\boldsymbol{b}\|^2$ in linear learning and $J(\boldsymbol{f}) = tr(\boldsymbol{Z}\boldsymbol{\mathcal{H}}\boldsymbol{Z}^\top) + \lambda'\|\boldsymbol{b}\|^2$ in kernel learning. This optimization problem can be solved by the regular dual quadratic program.

3. Statistical theory

In this section, we first establish Fisher consistency of the estimated ITRs by our proposed method, and then study the convergence rate of the excess risk.

3.1. Fisher consistency

Given $\boldsymbol{x} \in \mathcal{X}$ and $a \in \mathcal{A}$, denote the expected reward as $R(\boldsymbol{x}, a) = E[R|\boldsymbol{X} = \boldsymbol{x}, A = a]$. Define the positive part as $R^+(\boldsymbol{x}, a) = E[R\cdot I(R \geq 0)|\boldsymbol{X} = \boldsymbol{x}, A = a]$ and the negative part as $R^-(\boldsymbol{x}, a) = E[R \cdot I(R < 0)|\boldsymbol{X} = \boldsymbol{x}, A = a]$. Note that $R(\boldsymbol{x}, a) = R^+(\boldsymbol{x}, a) + R^-(\boldsymbol{x}, a)$. Before proceeding, we give the following proposition to specify the optimal ITR $\mathcal{D}^*(\boldsymbol{x})$.

Proposition 2. *The optimal ITR $\mathcal{D}^*(\boldsymbol{x})$ defined in (2.4) satisfies $\mathcal{D}^*(\boldsymbol{x}) = \operatorname{argmax}_{a\in\mathcal{A}} R(\boldsymbol{x}, a)$.*

By (2.7), define the generalization error as

$$\mathcal{R}(\boldsymbol{f}) = E\Bigg[\frac{|R|}{\Pr(A|\boldsymbol{X})}\{I(R\geq 0)I(M(\boldsymbol{f}(\boldsymbol{X}),A)<0)+$$

$$I(R<0)I(M(\boldsymbol{f}(\boldsymbol{X}),A)\geq 0)\}\Bigg].$$

The minimizer of $\mathcal{R}(\boldsymbol{f})$, also called the Bayes rule, is denoted by $\bar{\boldsymbol{f}}$ yielding the optimal ITR $\mathcal{D}^*(\boldsymbol{x})$ under the top–down strategy as shown in Proposition 1, that is, $\mathcal{D}^*(\boldsymbol{x}) = \mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{x})$. Let $Z = (\boldsymbol{X}, A, R)$. For a surrogate loss ℓ, denote

$$V(\boldsymbol{f}, Z) = \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})} \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})),$$

where $\ell_R(u) = \ell(u)$ if $R \geq 0$ and $\ell_R(u) = \ell(-u)$ if $R < 0$. Define the risk as

$$\mathcal{R}_V(\boldsymbol{f}) = E\left[\sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})} \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X}))\right].$$

The theoretical minimizer of $\mathcal{R}_V(\boldsymbol{f})$ is denoted as $\boldsymbol{f}^* = \operatorname{arginf}_{\boldsymbol{f}} \mathcal{R}_V(\boldsymbol{f})$.

Fisher consistency, which is also referred as classification-calibration [2] or infinite-sample consistency [33], is a fundamental requirement of a (weighted) classification method. It requires that the ITRs corresponding to $\boldsymbol{f}^*$ and $\bar{\boldsymbol{f}}$ following the top–down strategy are same, that is, $\mathcal{D}_{\boldsymbol{f}^*}(\boldsymbol{x}) = \mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{x})$. Denote the expected reward given $\boldsymbol{x}$ and $a^{(m)}$ at the m-th layer as $R_m(\boldsymbol{x}, a^{(m)}) = E[R|\boldsymbol{X} = \boldsymbol{x}, A^{(m)} = a^{(m)}]$. Correspondingly, the positive part is $R_m^+(\boldsymbol{x}, a^{(m)}) = E[R \cdot I(R \geq 0)|\boldsymbol{X} = \boldsymbol{x}, A^{(m)} = a^{(m)}]$ and the negative part is $R_m^-(\boldsymbol{x}, a^{(m)}) = E[R \cdot I(R < 0)|\boldsymbol{X} = \boldsymbol{x}, A^{(m)} = a^{(m)}]$. We first give the following assumptions.

Assumption 1. *For a patient with covariates $\boldsymbol{x}$, denote by $\mathcal{D}^*(\boldsymbol{x})$ the optimal path. Assume that $\mathcal{D}^*(\boldsymbol{x})$ is the dominating path on the tree satisfying for any $m = 2, \ldots, \mathcal{L}(\mathcal{D}^*(\boldsymbol{x}))$,*

$$\mathcal{D}^{*(m)}(\boldsymbol{x}) = \operatorname*{argmax}_{T_{j_1,\ldots,j_m}\in Chi(\mathcal{D}^{*(m-1)}(\boldsymbol{x}))} R_m(\boldsymbol{x}, T_{j_1,\ldots,j_m}).$$

Assumption 2. *For a patient with covariates $\boldsymbol{x}$, denote by $\mathcal{D}^*(\boldsymbol{x})$ the optimal path. Assume that*

$$\mathcal{D}^{*(m)}(\boldsymbol{x}) = \operatorname*{argmax}_{T_{j_1,\ldots,j_m}\in Chi(\mathcal{D}^{*(m-1)}(\boldsymbol{x}))} R_m^-(\boldsymbol{x}, T_{j_1,\ldots,j_m}), \quad m = 2, \ldots, \mathcal{L}(\mathcal{D}^*(\boldsymbol{x})).$$

Assumption 1 states that at any layer, the corresponding node $\mathcal{D}^{*(m)}(\boldsymbol{x})$ is optimal among all candidates, the children of $\mathcal{D}^{*(m-1)}(\boldsymbol{x})$. Assumption 1 is natural in order to utilize hierarchical information. Assumption 2 requires that at any layer, $R_m^-(\boldsymbol{x}, \mathcal{D}^{*(m)}(\boldsymbol{x}))$ is the largest among the children of $\mathcal{D}^{*(m-1)}(\boldsymbol{x})$. It is reasonable in practice that adverse effects of the optimal treatment should be small on average.

Theorem 1. *Under Assumptions 1 and 2, it holds that $\mathcal{D}_{\boldsymbol{f}^*}(\boldsymbol{x}) = \mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{x})$ if (i) $\ell(u)$ is differentiable with $\ell'(u) < 0$ for any u; (ii) $\ell'(u)$ is nondecreasing in u.*

Theorem 1 gives the conditions of a surrogate loss to achieve Fisher consistency. The linear loss and commonly used loss functions such as the exponential loss $\ell(u) = e^{-u}$ and the deviance loss $\ell(u) = \log(1+\exp(-u))$ satisfy conditions in Theorem 1, and thus lead to Fisher consistency. In this case, the hinge loss does not satisfy the conditions in Theorem 1. Liu, Zhang and Wu [14] proposed a set of large-margin unified machine loss functions satisfying the conditions in Theorem 1, which take the hinge loss as a limit.

Remark 3. *As shown in the Appendix, for the linear loss, we define* $\boldsymbol{f}^* = \operatorname{arginf}_{\boldsymbol{f}:E(\|\boldsymbol{f}\|^2)\le 1} R_V(\boldsymbol{f})$ *to avoid the infinite case. We prove in the Appendix that under this definition, it also holds that* $\mathcal{D}_{\boldsymbol{f}^*}(\boldsymbol{x}) = \mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{x})$ *for the linear loss.*

Remark 4. *For a surrogate loss function* $\ell(u)$*, denote the minimizer of*

$$E\left[|R|/\Pr(A|\boldsymbol{X})\ell_R(M(\boldsymbol{f}(\boldsymbol{X}), A))\right],$$

by $\boldsymbol{f}_M$*. We also establish Fisher consistency for* $\boldsymbol{f}_M$ *in Theorem S1 in the Appendix. However, as argued in Section 2, for efficient computation, we use the additive formulation.*

3.2. Convergence rate of excess risk

In this subsection, we show that our method can achieve a fast convergence rate of excess risk under mild conditions. Recall that the expectation risk is

$$\mathcal{R}_V(\boldsymbol{f}) = E\left[\sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})} \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X}))\right],$$

and $\boldsymbol{f}^* = \operatorname{arginf}_{\boldsymbol{f}} \mathcal{R}_V(\boldsymbol{f})$. In addition, recall the Bayes rule $\bar{\boldsymbol{f}} = \operatorname{arginf}_{\boldsymbol{f}} \mathcal{R}(\boldsymbol{f})$, which yields the optimal ITR $\mathcal{D}^*(\boldsymbol{X})$ under the top–down strategy. By Fisher consistency, $\boldsymbol{f}^*$ leads to the same decision rule as $\bar{\boldsymbol{f}}$. Thus, $\mathcal{R}(\bar{\boldsymbol{f}}) = \mathcal{R}(\boldsymbol{f}^*)$. To quantify the performance of any $\boldsymbol{f} \in \mathcal{F}$ with $\mathcal{F}$ being the set of candidate functions, define the excess ℓ-risk as $e_V(\boldsymbol{f}, \boldsymbol{f}^*) = \mathcal{R}_V(\boldsymbol{f}) - \mathcal{R}_V(\boldsymbol{f}^*)$, measuring the difference in terms of the expectation risk. As shown in Subsection 2.2, we have $\mathcal{C}(\mathcal{D}_{\boldsymbol{f}}) - \mathcal{C}(\mathcal{D}^*) = \mathcal{R}(\boldsymbol{f}) - \mathcal{R}(\boldsymbol{f}^*)$, where $\mathcal{D}_{\boldsymbol{f}}$ is the ITR associated with $\boldsymbol{f}$ by the top–down strategy. Hence, it is sufficient to define the excess risk as the difference of the generalization error, that is, $e(\boldsymbol{f}, \boldsymbol{f}^*) = \mathcal{R}(\boldsymbol{f}) - \mathcal{R}(\boldsymbol{f}^*)$.

Note that $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^K$, where $K = m_{\text{leaf}} - 1$. For any loss ℓ, recall

$$V(\boldsymbol{f}, Z) = \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})} \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})),$$

with $Z = (\boldsymbol{X}, A, R)$. Denote the truncated version as $V^{\widetilde{T}}(\boldsymbol{f}, Z) = \widetilde{T} \wedge V(\boldsymbol{f}, Z)$, where $\widetilde{T}$ is a truncation constant. Let $e_{V^{\tilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) = E[V^{\widetilde{T}}(\boldsymbol{f}, Z) - V(\boldsymbol{f}^*, Z)]$. We introduce the following assumptions similar to Wang, Shen and Pan [31] and give the result of the convergence rate.

Assumption 3. *There exist constants $0 < \alpha \leq \infty$ and $c_1 > 0$ such that for any small $\epsilon > 0$,*

$$\sup_{\{\boldsymbol{f}\in\mathcal{F}: e_{V^{\widetilde{T}}}(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} |e(\boldsymbol{f},\boldsymbol{f}^*)| \leq c_1\epsilon^{\alpha}.$$

Assumption 4. *There exist constants $\beta \geq 0$ and $c_2 > 0$ such that for any small $\epsilon > 0$,*

$$\sup_{\{\boldsymbol{f}\in\mathcal{F}: e_{V^{\widetilde{T}}}(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} Var(V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}^*,Z)) \leq c_2\epsilon^{\beta}.$$

Assumption 3 controls the excess risk $|e(\boldsymbol{f},\boldsymbol{f}^*)|$ in the neighborhood of $\boldsymbol{f}^*$ and describes a first moment relationship of the excess risk between $\boldsymbol{f}$ and $\boldsymbol{f}^*$. Assumption 4 describes a variance condition. Now, we define a complexity measure of a function space $\mathcal{G}$. Given $\epsilon > 0$, denote $\{(g_i^l, g_i^u)\}$ as an ϵ-bracketing set of $\mathcal{G}$ if for any $g \in \mathcal{G}$, there exists an i such that $g_i^l \leq g \leq g_i^u$ and $[E(\|g_i^u - g_i^l\|^2)]^{1/2} \leq \epsilon$. Define the metric entropy with bracketing $\mathcal{H}_B(\epsilon,\mathcal{G})$ as the logarithm of the cardinality of the smallest ϵ-bracketing set for $\mathcal{G}$.

Let $\boldsymbol{f}_0 = \boldsymbol{f}^*$ if $\boldsymbol{f}^* \in \mathcal{F}$, otherwise $\boldsymbol{f}_0 \in \mathcal{F}$ is an approximation to $\boldsymbol{f}^*$ such that $e_V(\boldsymbol{f}_0,\boldsymbol{f}^*) \leq \varepsilon_n^2/4$, where ε_n is defined in the following Assumption 5. Define $\mathcal{F}^V(t) = \{V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}_0,Z) : \boldsymbol{f} \in \mathcal{F}, J(\boldsymbol{f}) \leq J_0 t\}$, where $J(\boldsymbol{f}) = \|\boldsymbol{f}\|^2$ for the linear learner or $J(\boldsymbol{f}) = \langle \boldsymbol{f},\boldsymbol{f}\rangle_h$ for the nonlinear learner with $\langle\cdot,\cdot\rangle_h$ being the inner product of h, and $J_0 = \max\{J(\boldsymbol{f}_0), 1\}$. The following assumption measures the complexity of the function space $\mathcal{F}^V(t)$.

Assumption 5. *For some constants $c_i > 0, i = 3,4,5$, there exists $\varepsilon_n > 0$ such that $\sup_{t\geq 1}\phi(\varepsilon_n,t) \leq c_3 n^{1/2}$, where*

$$\phi(\varepsilon_n,t) = \int_{c_5\tilde{L}}^{c_4^{1/2}\tilde{L}^{\beta/2}} \mathcal{H}_B^{1/2}(u,\mathcal{F}^V(t))du/\tilde{L},$$

with $\tilde{L} = \tilde{L}(\varepsilon_n,\lambda,t) = \min\{\varepsilon_n^2 + \lambda J_0(t/2-1), 1\}$.

The following Theorem 2 gives the bound of the excess risk $e(\hat{\boldsymbol{f}}_\lambda,\boldsymbol{f}^*)$ between $\hat{\boldsymbol{f}}_\lambda$ and $\boldsymbol{f}^*$, where $\hat{\boldsymbol{f}}_\lambda \in \mathcal{F}$ is the minimizer of the optimization problem (2.9).

Theorem 2. *Under Assumptions 3–5, there exists a constant c_6 such that,*

$$P(e(\hat{\boldsymbol{f}}_\lambda,\boldsymbol{f}^*) \geq c_1\delta_n^{2\alpha}) \leq 3.5\exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)}),$$

provided that $\lambda^{-1} \geq 2\delta_n^{-2}J_0$, where $\delta_n^2 = \min\{\varepsilon_n^2 + 2e_V(\boldsymbol{f}_0,\boldsymbol{f}^*), 1\}$.

Corollary 1. *Under the assumptions in Theorem 2, $|e(\hat{\boldsymbol{f}}_\lambda,\boldsymbol{f}^*)| = O_p(\delta_n^{2\alpha})$ provided that $n(\lambda J_0)^{2-\min\{\beta,1\}}$ is bounded away from 0 when $n \to \infty$.*

The conclusion of Theorem 2 is similar to that of Wang, Shen and Pan [31]. In the Appendix, we verify Assumptions 3–5 under mild conditions. It is shown that $\beta = 1$ in Assumption 4 for the linear loss and loss functions that are twice continuously differentiable with $\ell''(u) > 0$. For Assumption 5, the explicit expression of $\mathcal{H}_B(\epsilon, \mathcal{F}^V(t))$ is given. Finally, it should be pointed out that there is no general result about α in Assumption 3 as it depends on $\boldsymbol{f}^*$ and the distribution of $(\boldsymbol{X}, A, R)$. We give a specific example in the Appendix, showing the value of α and deriving the convergence rate by Corollary 1.

Theorem 2 and Corollary 1 are established when the propensity score is known. When the propensity score is unknown, let $\hat{\pi}(A^{(m)}|\boldsymbol{X})$ be an estimator of $\Pr(A^{(m)}|\boldsymbol{X})$. Define

$$V^{\hat{\pi}}(\boldsymbol{f}, Z) = \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{|R|}{\hat{\pi}(A^{(m)}|\boldsymbol{X})} \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})),$$

and $\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}} = \operatorname{argmin}_{\boldsymbol{f} \in \mathcal{F}} n^{-1} \sum_{i=1}^n V^{\hat{\pi}}(\boldsymbol{f}, z_i) + \lambda J(\boldsymbol{f})$. To establish the asymptotic theory for $\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}$, the following assumptions are required.

Assumption 6. *Assume the following conditions are satisfied,*

(1) $|R|$ is bounded by a constant $M_R > 0$.

(2) There exists $M_A, M_\pi > 0$ such that $\Pr(A^{(m)}|\boldsymbol{X}) \geq M_A$ and $\hat{\pi}(A^{(m)}|\boldsymbol{X}) \geq M_\pi$ for any $A \in \mathcal{A}$, $m = 2, \dots, \mathcal{L}(A)$ and $\boldsymbol{X} \in \mathcal{X}$.

(3) The loss function $\ell(u)$ is Lipschitz with a constant $\tilde{\gamma}$. Specifically, for any u_1, u_2 that are bounded by a finite value, it holds that $|\ell(u_1) - \ell(u_2)| \leq \tilde{\gamma}|u_1 - u_2|$.

Assumption 7. *There exists $s_n > 0$ such that $\delta_n^{-2} s_n = o(1)$ and $|\hat{\pi}(A^{(m)}|\boldsymbol{X}) - \Pr(A^{(m)}|\boldsymbol{X})| = O_p(s_n)$ with $\delta_n^2 = \min\{\varepsilon_n^2 + 2e_V(\boldsymbol{f}_0, \boldsymbol{f}^*), 1\}$.*

Assumption 6 assumes that the outcomes are bounded, which is reasonable in practice. We also assume that the propensity scores are bounded away from 0, which is a key assumption to build connections between observed and potential data in causal inference. Moreover, the loss function is assumed to be Lipschitz. Assumption 7 clarifies the convergence rate of the estimator $\hat{\pi}$. Theorem 3 below shows that the convergence rate in Corollary 1 still holds for $\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}$, where the proof is given in the Appendix.

Theorem 3. *Under Assumptions 3–7, it holds that $|e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)| = O_p(\delta_n^{2\alpha})$ provided that $n(\lambda J_0)^{2-\min\{\beta,1\}}$ is bounded away from 0 as $n \to \infty$ with $\delta_n^2 = \min\{\varepsilon_n^2 + 2e_V(\boldsymbol{f}_0, \boldsymbol{f}^*), 1\}$.*

4. Simulation study

In this section, we conduct simulations under hierarchical treatment scenarios using both linear and nonlinear learners to evaluate the performance of our method. Denote the proposed HOAL under three loss functions as HOAL$_{\text{lin}}$ (linear loss), HOAL$_{\text{wl}}$ (weighted linear loss), and HOAL$_{\text{hinge}}$ (hinge loss). For comparisons, we adapt several existing methods to hierarchical scenarios. Specifically, we consider (1) the multicategory outcome-weighted margin-based learning (MOML) [36], which is a flat approach involving only leaf nodes and ignoring the hierarchical structure; (2) the sequential multicategory outcome-weighted margin-based learning (SMOML), sequentially applying MOML for each parent node; (3) the sequential binary outcome-weighted learning (SBOWL), which sequentially applies binary OWL [37] for each node in the hierarchical tree.

4.1. Evaluation metrics

We first introduce several evaluation metrics. Two types of criteria are considered. The first is to evaluate the misclassification rates of the estimated optimal ITRs $\hat{\mathcal{D}}(\boldsymbol{X})$ from the true optimal ITRs, including the 0–1 loss, symmetric loss, and two hierarchical losses. The second is to evaluate the value function using the estimated optimal ITRs. Smaller values are preferred for the first type of criteria and larger values are preferred for the second type.

Note that paths on the tree may have different lengths, depending on the layer where the leaf node locates. For simplicity of comparison, we introduce the following notations. Let $|\mathcal{T}| = q$, representing the total number of nodes on the tree except for the root. Sort the nodes in $\mathcal{T}$ by layers from top to bottom, and the nodes at the same layer from left to right. Denote the sorted nodes as $\{T_{(1)}, T_{(2)}, \ldots, T_{(q)}\}$ and the root node is denoted as $T_{(0)}$. For a path from the root to a leaf on the tree, we transform it into a binary vector, denoted as $\boldsymbol{\mathcal{Q}}(\cdot) \in \boldsymbol{R}^q$, of which the j-th coordinate indicates whether the j-th node is on the path.

The 0–1 misclassification rate [3] is defined as

$$\ell_{0\text{–}1} = I(\boldsymbol{\mathcal{Q}}(\hat{\mathcal{D}}(\boldsymbol{X})) \neq \boldsymbol{\mathcal{Q}}(\mathcal{D}(\boldsymbol{X}))),$$

which equals to 0 if the whole path is the same, and 1 otherwise. The symmetric misclassification rate [30] penalizes errors at each node, defined as

$$\ell_{\Delta} = q^{-1} \sum_{j=1}^{q} I(\boldsymbol{\mathcal{Q}}(\hat{\mathcal{D}}(\boldsymbol{X}))_j \neq \boldsymbol{\mathcal{Q}}(\mathcal{D}(\boldsymbol{X}))_j),$$

where $\boldsymbol{\mathcal{Q}}(\cdot)_j$ is the j-th coordinate of $\boldsymbol{\mathcal{Q}}(\cdot)$. The hierarchical misclassification rate [30] views the mistakes made at higher layers being more important than those at lower layers, defined as

$$\ell_{\mathrm{H}} = \sum_{j=1}^{q} v_{T_{(j)}} I(\{\boldsymbol{\mathcal{Q}}(\hat{\mathcal{D}}(\boldsymbol{X}))_j \neq \boldsymbol{\mathcal{Q}}(\mathcal{D}(\boldsymbol{X}))_j\} \wedge \{\boldsymbol{\mathcal{Q}}(\hat{\mathcal{D}}(\boldsymbol{X}))_s = \boldsymbol{\mathcal{Q}}(\mathcal{D}(\boldsymbol{X}))_s,\ \forall s < j\}).$$

The coefficients $0 \leq v_{T_{(j)}} \leq 1, j = 1, \ldots, q$ are used for down-scaling. There are two popular choices for $v_{T_{(j)}}$. Specifically, denote ℓ_{H} as $\ell_{\mathrm{H(sib)}}$ when $v_{T_{(0)}} = 1, v_{T_{(j)}} = v_{\mathrm{Par}(T_{(j)})}/|\mathrm{Sib}(T_{(j)})|, j = 1, \ldots, q$ with $|\mathrm{Sib}(T_{(j)})|$ being the number of siblings of $T_{(j)}$, and denote ℓ_{H} as $\ell_{\mathrm{H(sub)}}$ when $v_{T_{(j)}} = q^{-1}|\mathrm{subtree}(T_{(j)})|, j = 1, \ldots, q$ with $|\mathrm{subtree}(T_{(j)})|$ being the size of the subtree rooted by $T_{(j)}$.

Murphy, Der Laan and Robins [19] proposed the estimated value function for $\hat{\mathcal{D}}(\boldsymbol{X})$. According to our formulation in hierarchical scenarios, the empirical value function is defined as

$$\mathbb{P}^* \left[\sum_{m=2}^{\mathcal{L}(A)} \frac{R \cdot I(A^{(m)} = \hat{\mathcal{D}}^{(m)}(\boldsymbol{X}))}{\Pr(A^{(m)}|\boldsymbol{X})} \right] / \mathbb{P}^* \left[\sum_{m=2}^{\mathcal{L}(A)} \frac{I(A^{(m)} = \hat{\mathcal{D}}^{(m)}(\boldsymbol{X}))}{\Pr(A^{(m)}|\boldsymbol{X})} \right],$$

where $\mathbb{P}^*$ denotes the empirical average of the testing dataset.

4.2. Simulations

In our simulations, we consider four examples, where the first two utilize linear learner and the last two apply nonlinear learner. As for nonlinear scenarios, we utilize Gaussian kernels, that is, $h(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\tau_n \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, where τ_n is the kernel bandwidth. In Examples 1 and 3, we simulate data from a randomized trial. That is, treatment assignments are independent of any patient's covariates. In Examples 2 and 4, we consider observational studies, where the assignment of the treatment depends on covariates. We use multinomial logistic regression to estimate the propensity score.

In each example, we generate samples and split them into the training set, the validation set and the test set with their sizes denoted as n, n_{vl} and n_{te}, respectively. In the following simulations, we set the ratio $n : n_{\mathrm{vl}} : n_{\mathrm{te}} = 1 : 1 : 2$. We first learn $\boldsymbol{f}$ on the training set and choose the best tuning parameter based on the validation set. Then we apply the estimated learner on the test set to compute the evaluation metrics. The tuning parameters λ, λ' and γ are chosen from $\{10^{-2+4k/10}, k = 0, 1, \ldots, 10\}$. Let $\tau_n = 1/(2\sigma_n^2)$, where σ_n is chosen from $\{10^{-2+2k/5}, k = 0, 1, \ldots, 5\}$. Moreover, to reduce the computational cost, we set $\lambda' = 1$ for $\mathrm{HOAL}_{\mathrm{hinge}}$.

Example 1 (linear). We consider a tree of k layers. There are three nodes at the second layer and each nonleaf node at the m-th layer ($m \geq 2$) has two children. The tree is shown in FIG 3 (left), where the digits stand for the labels. Note that all leaves locate at the k-th layer. The optimal path $\mathcal{D}^*$ is generated by a discrete uniform distribution in the set of all paths on the tree. Moreover, $\boldsymbol{X}|\mathcal{D}^* \sim N(\boldsymbol{t}(\mathcal{D}^*), 0.1\boldsymbol{I}_{p\times p})$, where $\boldsymbol{t}(\mathcal{D}^*)$ is a zero vector of length $p = q$ except for the $\mathcal{D}^{*(m)}$-th element being $1/\sqrt{m-1}, m = 2, \ldots, k$. For example, if $\mathcal{D}^* = \{0, 1, 4\}$, $\boldsymbol{t}(\mathcal{D}^*) = (1, 0, 0, 1/\sqrt{2}, 0, \ldots, 0)^\top$. The assigned A is generated similar to $\mathcal{D}^*$ by following a discrete uniform distribution in the set of all paths on the tree. The reward $R \sim N(\mu(\boldsymbol{X}, A, \mathcal{D}^*), 1)$, where $\mu(\boldsymbol{X}, A, \mathcal{D}^*) = \boldsymbol{X}^\top(\mathbf{1}_{\lfloor p/2 \rfloor}^\top, -\mathbf{1}_{p-\lfloor p/2 \rfloor}^\top)^\top + 5 \cdot I(A = \mathcal{D}^*)$. We set $n = 500$ and $k = 3, 4$, respectively. Note that $p = 9$ for $k = 3$ and $p = 21$ for $k = 4$.

Example 2 (linear). We consider a tree of 3 layers with leaf nodes locating at different layers. The tree is shown in FIG 3 (right), where the digits stand for the labels. The optimal path $\mathcal{D}^*$ is generated by a discrete uniform distribution in the set of all paths on the tree. Moreover, $\boldsymbol{X}|\mathcal{D}^* \sim N(\boldsymbol{t}(\mathcal{D}^*), 0.1\boldsymbol{I}_{p\times p})$, where $\boldsymbol{t}(\mathcal{D}^*)$ is a zero vector of length $p = q + \tilde{p}$ except for the $\mathcal{D}^{*(m)}$-th element being $1/\sqrt{m-1}$ for $m = 2, \ldots, \mathcal{L}(\mathcal{D}^*)$. Note that we add additional $\tilde{p}$ covariates as random noises. The assignment of the treatment is based on a multinomial

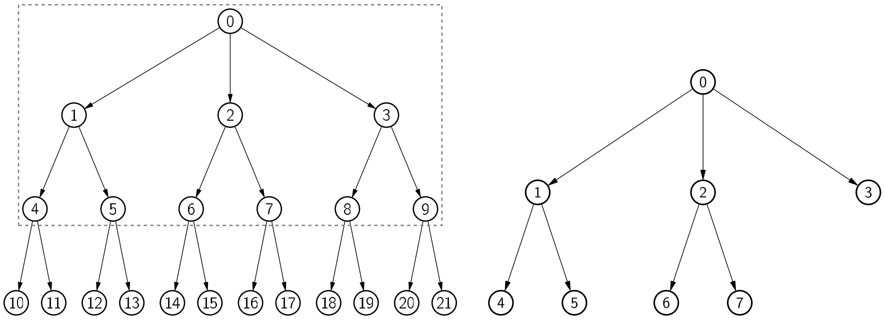

FIG 3. *Structures of Examples 1 (left) and 2 (right) where digits represent labels of nodes.*

distribution with

$$\begin{aligned}\Pr(A=\{0,3\}|\boldsymbol{X})=1/\widetilde{M},\quad &\Pr(A=\{0,1,4\}|\boldsymbol{X})=\exp\left(0.1X_1-0.2X_2\right)/\widetilde{M},\\ &\Pr(A=\{0,1,5\}|\boldsymbol{X})=\exp\left(0.1X_1+0.2X_2\right)/\widetilde{M},\\ &\Pr(A=\{0,2,6\}|\boldsymbol{X})=\exp\left(0.1X_1-0.1X_2\right)/\widetilde{M},\\ &\Pr(A=\{0,2,7\}|\boldsymbol{X})=\exp\left(0.1X_1+0.1X_2\right)/\widetilde{M},\end{aligned}$$

and $\widetilde{M}=1+\exp\left(0.1X_1-0.2X_2\right)+\exp\left(0.1X_1+0.2X_2\right)+\exp\left(0.1X_1-0.1X_2\right)+\exp\left(0.1X_1+0.1X_2\right)$. The reward $R\sim N(\mu(\boldsymbol{X},A,\mathcal{D}^*),1)$, where $\mu(\boldsymbol{X},A,\mathcal{D}^*)=\boldsymbol{X}^\top(\mathbf{1}^\top_{\lfloor p/2\rfloor},-\mathbf{1}^\top_{p-\lfloor p/2\rfloor},\mathbf{0}_{\tilde{p}})^\top+5\cdot I(A=\mathcal{D}^*)$. We set $n=500$ and $\tilde{p}=5,10$, respectively. Note that $p=12$ for $\tilde{p}=5$ and $p=17$ for $\tilde{p}=10$.

Example 3 (nonlinear). The setup is the same as that of Example 1, except for $\boldsymbol{t}(\mathcal{D}^*)=2\sin(\xi_k(\mathcal{D}^*))$, where $\xi_k(\mathcal{D}^*)$ is the embedded point for the leaf node $\mathcal{D}^{*(k)}$. Furthermore, we randomly choose 50% samples with $\mathcal{D}^{*(2)}=1$ to set $\boldsymbol{t}(\mathcal{D}^*)=-2\sin(\xi_k(\mathcal{D}^*))$. Let $n=300$ and $k=3,4$, respectively. Note that $p=9$ for $k=3$ and $p=21$ for $k=4$.

Example 4 (nonlinear). The setup is the same as that of Example 2, except for $\boldsymbol{t}(\mathcal{D}^*)=2\sin(\xi_{\mathcal{L}(\mathcal{D}^*)}(\mathcal{D}^*))$, where $\xi_{\mathcal{L}(\mathcal{D}^*)}(\mathcal{D}^*)$ is the embedded point for the leaf node $\mathcal{D}^{*(\mathcal{L}(\mathcal{D}^*))}$. Furthermore, we randomly choose 50% samples with $\mathcal{D}^{*(2)}=1$ to set $\boldsymbol{t}(\mathcal{D}^*)=-2\sin(\xi_{\mathcal{L}(\mathcal{D}^*)}(\mathcal{D}^*))$. Let $n=300$ and $\tilde{p}=5,10$, respectively. Note that $p=12$ for $\tilde{p}=5$ and $p=17$ for $\tilde{p}=10$.

Tables 1 and 2 display the average results of the five evaluation metrics and the running time (seconds) over 100 replications for Examples 1 and 2 using linear learners. In Example 1, we consider a randomized trial with a homogeneous tree. All hierarchical methods, SMOML, SBOWL, HOAL_{lin}, HOAL_{wl} and $\text{HOAL}_{\text{hinge}}$ outperform the flat approach, MOML, which considers only leaf nodes. Furthermore, the proposed HOAL under three different loss functions has advantages over other existing methods in all evaluation metrics. HOAL_{wl} provides more robust improvements, while HOAL_{lin} achieves comparable results. When the hierarchical tree goes deeper as the number of layers k increases from 3 to 4, the hierarchical structure becomes more complex and the advantages of

Table 1

Average results as well as standard deviations for Example 1 over 100 replications. The best value in each column is boldfaced.

	$\ell_{\text{0-1}}$	ℓ_{Δ}	$\ell_{\text{H(sib)}}$	$\ell_{\text{H(sub)}}$	Value	Time (s)
Example 1 $k = 3, n = 500$						
MOML	$0.234_{0.006}$	$0.075_{0.002}$	$0.056_{0.002}$	$0.049_{0.002}$	$3.712_{0.03}$	2.036
SMOML	$0.164_{0.004}$	$0.049_{0.001}$	$0.037_{0.001}$	$0.030_{0.001}$	$3.987_{0.024}$	1.709
SBOWL	$0.168_{0.004}$	$0.050_{0.001}$	$0.038_{0.001}$	$0.031_{0.001}$	$3.967_{0.022}$	3.639
HOAL$_{\text{lin}}$	$0.162_{0.005}$	$0.043_{0.001}$	$0.032_{0.001}$	$0.025_{0.001}$	$3.999_{0.024}$	0.702
HOAL$_{\text{wl}}$	$\mathbf{0.141}_{0.007}$	$\mathbf{0.036}_{0.001}$	$\mathbf{0.027}_{0.001}$	$\mathbf{0.021}_{0.001}$	$\mathbf{4.082}_{0.02}$	1.032
HOAL$_{\text{hinge}}$	$\mathbf{0.141}_{0.004}$	$0.039_{0.001}$	$0.029_{0.001}$	$0.023_{0.001}$	$4.063_{0.022}$	24.798
Example 1 $k = 4, n = 500$						
MOML	$0.744_{0.006}$	$0.167_{0.002}$	$0.180_{0.003}$	$0.170_{0.003}$	$2.255_{0.035}$	5.375
SMOML	$0.720_{0.007}$	$0.153_{0.002}$	$0.164_{0.003}$	$0.153_{0.003}$	$2.415_{0.033}$	4.036
SBOWL	$0.716_{0.007}$	$0.151_{0.003}$	$0.160_{0.003}$	$0.149_{0.003}$	$2.448_{0.037}$	8.342
HOAL$_{\text{lin}}$	$\mathbf{0.605}_{0.006}$	$\mathbf{0.110}_{0.002}$	$\mathbf{0.112}_{0.002}$	$\mathbf{0.099}_{0.002}$	$\mathbf{2.845}_{0.034}$	0.999
HOAL$_{\text{wl}}$	$0.606_{0.006}$	$0.111_{0.002}$	$0.113_{0.002}$	$0.100_{0.002}$	$2.830_{0.035}$	1.529
HOAL$_{\text{hinge}}$	$0.617_{0.006}$	$0.113_{0.002}$	$0.115_{0.002}$	$0.102_{0.002}$	$2.757_{0.038}$	58.456

Table 2

Average results as well as standard deviations for Example 2 over 100 replications. The best value in each column is boldfaced.

	$\ell_{\text{0-1}}$	ℓ_{Δ}	$\ell_{\text{H(sib)}}$	$\ell_{\text{H(sub)}}$	Value	Time (s)
Example 2 $\tilde{p} = 5, n = 500$						
MOML	$0.183_{0.005}$	$0.069_{0.002}$	$0.045_{0.001}$	$0.050_{0.002}$	$3.766_{0.024}$	1.958
SMOML	$0.143_{0.004}$	$0.053_{0.002}$	$0.034_{0.001}$	$0.038_{0.001}$	$3.956_{0.017}$	1.758
SBOWL	$0.144_{0.003}$	$0.052_{0.001}$	$0.034_{0.001}$	$0.037_{0.001}$	$3.923_{0.016}$	3.678
HOAL$_{\text{lin}}$	$0.137_{0.003}$	$0.048_{0.001}$	$0.032_{0.001}$	$0.035_{0.001}$	$3.998_{0.015}$	0.672
HOAL$_{\text{wl}}$	$0.121_{0.002}$	$0.043_{0.001}$	$0.028_{0.001}$	$0.031_{0.001}$	$\mathbf{4.049}_{0.016}$	1.015
HOAL$_{\text{hinge}}$	$\mathbf{0.118}_{0.002}$	$\mathbf{0.041}_{0.001}$	$\mathbf{0.027}_{0.001}$	$\mathbf{0.029}_{0.001}$	$4.04_{0.015}$	25.336
Example 2 $\tilde{p} = 10, n = 500$						
MOML	$0.221_{0.005}$	$0.084_{0.002}$	$0.055_{0.001}$	$0.062_{0.002}$	$3.645_{0.022}$	2.054
SMOML	$0.171_{0.003}$	$0.063_{0.001}$	$0.041_{0.001}$	$0.046_{0.001}$	$3.878_{0.018}$	2.074
SBOWL	$0.177_{0.003}$	$0.064_{0.001}$	$0.042_{0.001}$	$0.047_{0.001}$	$3.823_{0.016}$	4.231
HOAL$_{\text{lin}}$	$0.165_{0.003}$	$0.059_{0.001}$	$0.039_{0.001}$	$0.043_{0.001}$	$3.921_{0.017}$	0.67
HOAL$_{\text{wl}}$	$0.153_{0.003}$	$\mathbf{0.054}_{0.001}$	$\mathbf{0.036}_{0.001}$	$0.040_{0.001}$	$\mathbf{3.964}_{0.016}$	1.001
HOAL$_{\text{hinge}}$	$\mathbf{0.152}_{0.003}$	$\mathbf{0.054}_{0.001}$	$\mathbf{0.036}_{0.001}$	$\mathbf{0.039}_{0.001}$	$3.939_{0.016}$	31.441

our method become clearer. This implies that HOAL can produce stable estimation in complex hierarchical scenarios. As for the running time, HOAL$_{\text{lin}}$ is the fastest, followed by HOAL$_{\text{wl}}$, SMOML, MOML, SBOWL and HOAL$_{\text{hinge}}$. Thus, our method achieves great improvements in prediction accuracy and is computationally efficient under the (weighted) linear loss, where a closed form solution is available. In Example 2, we consider an observational study with a heterogeneous tree and add additional random noises. Again, all hierarchical methods perform better than the flat approach and our method outperforms other existing methods in most evaluation metrics. HOAL$_{\text{wl}}$ and HOAL$_{\text{hinge}}$ achieve the best performance and HOAL$_{\text{lin}}$ provides comparable results. The results imply that the proposed method is robust in an observational study when random noises are included. Furthermore, the pattern of the running time is similar to that of Example 1. When nonlinear learning is applied, the average results of the five evaluation metrics and the running time (seconds) over 100 replications for Examples 3 and 4 are shown in Tables 3 and 4. In Example 3, our method outperforms other methods in all cases. All the methods become worse when the tree gets bigger, while HOAL can still produce better results. $\mathrm{HOAL}_{\mathrm{wl}}$ is the best one in all evaluation metrics. As for the running time, $\mathrm{HOAL}_{\mathrm{lin}}$ is the fastest, followed by $\mathrm{HOAL}_{\mathrm{wl}}$, SMOML, SBOWL, MOML and $\mathrm{HOAL}_{\mathrm{hinge}}$. In Example 4, we consider the effect of additional random noises in an observational study. Our method under the weighted linear loss and the hinge loss still has great advantages over other methods, while $\mathrm{HOAL}_{\mathrm{lin}}$ is not good since the linear loss is not that robust to outliers. Note that in these two examples, for nonlinear learning, our method under the (weighted) linear loss shows great advantages in computation.

Table 3

Average results as well as standard deviations for Example 3 over 100 replications. The best value in each column is boldfaced.

	$\ell_{\text{0-1}}$	ℓ_Δ	$\ell_{\mathrm{H(sib)}}$	$\ell_{\mathrm{H(sub)}}$	Value	Time (s)
			Example 3 $k = 3, n = 300$			
MOML	$0.263_{0.013}$	$0.095_{0.003}$	$0.070_{0.002}$	$0.066_{0.002}$	$3.851_{0.036}$	163.619
SMOML	$0.203_{0.008}$	$0.081_{0.003}$	$0.060_{0.003}$	$0.058_{0.003}$	$4.001_{0.036}$	54.061
SBOWL	$0.191_{0.008}$	$0.074_{0.003}$	$0.055_{0.003}$	$0.053_{0.003}$	$4.050_{0.033}$	86.863
$\mathrm{HOAL}_{\mathrm{lin}}$	$0.191_{0.017}$	$0.056_{0.003}$	$0.042_{0.002}$	$0.040_{0.002}$	$4.226_{0.033}$	2.026
$\mathrm{HOAL}_{\mathrm{wl}}$	$\mathbf{0.169}_{0.011}$	$\mathbf{0.050}_{0.003}$	$\mathbf{0.037}_{0.002}$	$\mathbf{0.035}_{0.002}$	$\mathbf{4.277}_{0.032}$	3.160
$\mathrm{HOAL}_{\mathrm{hinge}}$	$0.173_{0.012}$	$0.061_{0.003}$	$0.045_{0.002}$	$0.042_{0.002}$	$4.129_{0.035}$	254.995
			Example 3 $k = 4, n = 300$			
MOML	$0.777_{0.005}$	$0.185_{0.002}$	$0.205_{0.002}$	$0.198_{0.002}$	$1.600_{0.035}$	551.337
SMOML	$0.768_{0.006}$	$0.181_{0.002}$	$0.200_{0.003}$	$0.192_{0.003}$	$1.713_{0.038}$	58.672
SBOWL	$0.775_{0.007}$	$0.183_{0.003}$	$0.202_{0.003}$	$0.193_{0.003}$	$1.665_{0.039}$	100.524
$\mathrm{HOAL}_{\mathrm{lin}}$	$0.725_{0.007}$	$0.168_{0.003}$	$0.184_{0.003}$	$0.177_{0.003}$	$1.789_{0.043}$	2.787
$\mathrm{HOAL}_{\mathrm{wl}}$	$\mathbf{0.713}_{0.007}$	$\mathbf{0.157}_{0.002}$	$\mathbf{0.172}_{0.003}$	$\mathbf{0.164}_{0.003}$	$\mathbf{1.826}_{0.041}$	4.519
$\mathrm{HOAL}_{\mathrm{hinge}}$	$0.741_{0.006}$	$0.166_{0.002}$	$0.181_{0.003}$	$0.170_{0.003}$	$1.775_{0.041}$	543.331

4.3. Sensitive study

For embedding nodes in Algorithm S2, we define a down-scaling constant δ such that $L^{(m+1)} = L^{(m)}/\delta$, where $L^{(m)}$ is the length of the embedded points at the m-th layer. As suggested by Fan et al. [9], we use $\delta = \sqrt{5}$ in above simulations. Next we perform a sensitive study to investigate the performance of our method HOAL relative to the value of δ. For illustration, we report the results of $\mathrm{HOAL}_{\mathrm{hinge}}$ with δ being $0.1, 1, 2, \sqrt{5}, 2.5, 3$.

FIG 4 shows the plot of average results of ℓ_{0-1} over 100 replications in Examples 1–4 with $K = 3$ and $\tilde{p} = 5$, respectively, versus different values of δ using $\mathrm{HOAL}_{\mathrm{hinge}}$. The other evaluation metrics show the similar pattern and thus are omitted here. It can be seen that when $\delta < 1$, the performance of $\mathrm{HOAL}_{\mathrm{hinge}}$ is relatively poor since it violates the basic requirement that the length of the embedded points should be decreasing along the tree. For $\delta = 1$, the result is still

Table 4

Average results as well as standard deviations for Example 4 over 100 replications. The best value in each column is boldfaced.

	$\ell_{\text{0-1}}$	ℓ_{Δ}	$\ell_{\text{H(sib)}}$	$\ell_{\text{H(sub)}}$	Value	Time (s)
			Example 4 $\tilde{p} = 5, n = 300$			
MOML	$0.379_{0.009}$	$0.166_{0.003}$	$0.106_{0.002}$	$0.131_{0.002}$	$3.036_{0.034}$	102.037
SMOML	$0.333_{0.007}$	$0.150_{0.003}$	$0.099_{0.002}$	$0.122_{0.003}$	$3.217_{0.036}$	47.171
SBOWL	$0.328_{0.007}$	$0.143_{0.003}$	$0.094_{0.002}$	$0.115_{0.003}$	$3.248_{0.036}$	76.654
HOAL_{lin}	$0.345_{0.009}$	$0.149_{0.003}$	$0.102_{0.002}$	$0.128_{0.003}$	$3.287_{0.034}$	1.865
HOAL_{wl}	$0.331_{0.018}$	$\mathbf{0.121}_{0.003}$	$\mathbf{0.080}_{0.002}$	$\mathbf{0.099}_{0.002}$	$\mathbf{3.450}_{0.029}$	3.053
$\text{HOAL}_{\text{hinge}}$	$\mathbf{0.318}_{0.009}$	$0.135_{0.003}$	$0.090_{0.002}$	$0.109_{0.002}$	$3.331_{0.032}$	197.455
			Example 4 $\tilde{p} = 10, n = 300$			
MOML	$0.463_{0.006}$	$0.209_{0.003}$	$0.135_{0.002}$	$0.166_{0.002}$	$2.631_{0.035}$	99.149
SMOML	$0.443_{0.007}$	$0.200_{0.004}$	$0.131_{0.002}$	$0.161_{0.003}$	$2.783_{0.041}$	43.238
SBOWL	$0.438_{0.008}$	$0.198_{0.004}$	$0.129_{0.002}$	$0.158_{0.003}$	$2.805_{0.037}$	68.957
HOAL_{lin}	$0.463_{0.008}$	$0.210_{0.005}$	$0.141_{0.003}$	$0.176_{0.004}$	$2.746_{0.048}$	2.045
HOAL_{wl}	$\mathbf{0.412}_{0.009}$	$\mathbf{0.181}_{0.003}$	$\mathbf{0.119}_{0.002}$	$\mathbf{0.146}_{0.003}$	$\mathbf{2.960}_{0.04}$	3.311
$\text{HOAL}_{\text{hinge}}$	$0.416_{0.006}$	$0.184_{0.003}$	$0.122_{0.002}$	$0.149_{0.002}$	$2.900_{0.041}$	200.726

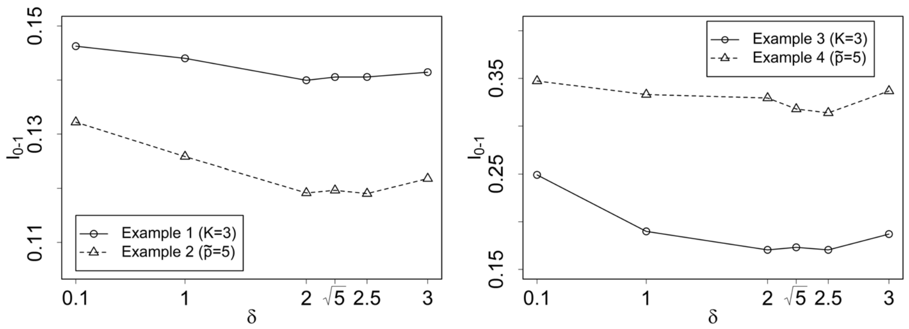

Fig 4. *Plot of average results of ℓ_{0-1} over 100 replications versus different values of δ using $HOAL_{hinge}$.*

somewhat inferior since it equally treats the nodes at different layers and does not utilize the hierarchical structure sufficiently. Moreover, for $\delta = 2, \sqrt{5}, 2.5$, the performance of $\text{HOAL}_{\text{hinge}}$ has a low volatility and is not sensitive to the value of δ. When δ becomes larger, that is, $\delta = 3$, ℓ_{0-1} increases slightly. Note that the theoretical analysis requires $\delta^2 \geq 2\sqrt{2} + 2$ to satisfy the hierarchical properties. Based on this sensitive study, we recommend to set $\delta = \sqrt{5}$.

5. Application to a type 2 diabetes study

We apply HOAL to a type 2 diabetes mellitus (T2DM) observational study to evaluate its performance in real data applications. The study population comprises T2DM patients during 2012-2013, from clinical practice research datalink (CPRD) [10]. Treatment exposures will focus on first-line injectables, which are first categorized to two groups and further into subgroups, naturally following

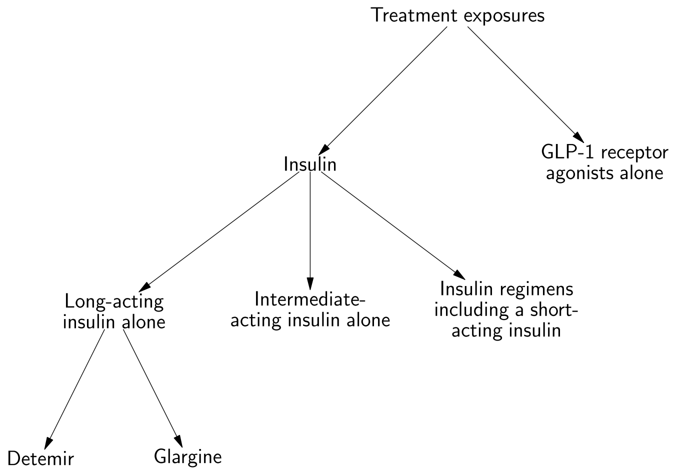
Treatment exposures
Insulin
GLP-1 receptor agonists alone
Long-acting insulin alone
Intermediate-acting insulin alone
Insulin regimens including a short-acting insulin
Detemir
Glargine

FIG 5. *Hierarchical structure of the treatments in T2DM.*

a hierarchical structure as shown in FIG 5. The tree has four layers. There are two nodes, insulin and glucagon-like peptide-1 (GLP-1) receptor agonists alone at the second layer, which are two important medications for patients with type 2 diabetes [17]. Moreover, there are different types of insulin. The node insulin then has three children, long-acting insulin alone, intermediate-acting insulin along, and insulin regimens including a short-acting insulin at the third layer. The node, long-acting insulin alone, has two children at the fourth layer, Detemir and Glargine.

As the goal of the treatment is to decrease HbA1c (%), the negative value of HbA1c change is chosen as the reward. Several covariates are considered, including demographics such as gender, age, and ethnicity, clinical factors such as high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), baseline HbA1c and smoking status. There are 1,138 patients satisfying aforementioned requirements and around 17% patients have complete observations. We use the same procedure in Chen et al. [5] to handle the missing data. After data preprocessing, we have 229 observations and involve five covariates including gender, age, HDL, LDL and baseline HbA1c.

To compute the propensity score, we fit a multinomial logistic regression model between the assigned path and the covariates. After estimating $\Pr(A|\boldsymbol{X})$, let $\Pr(A^{(m)}|\boldsymbol{X}) = \sum_{A\in\mathcal{A}:A^{(m)}\in A}\Pr(A|\boldsymbol{X})$. We apply HOAL under three losses to this study and compare with MOML, SMOML, and SBOWL. Linear learning and nonlinear learning using Gaussian kernels are considered. The tuning parameters λ, λ' and γ are chosen from $\{10^{-2+4k/10}, k = 0, 1, \ldots, 10\}$, and σ_n is chosen from $\{10^{-2+2k/5}, k = 0, 1, \ldots, 5\}$. Moreover, to reduce the computational cost, we set $\lambda' = 1$ for $\text{HOAL}_{\text{hinge}}$. We use the 5-fold cross-validation to estimate the empirical value function. Specifically, we repeat the 5-fold cross-validation 100 times. In each replication, we randomly split the whole data into 5 folds.

Each fold is used exactly once as a testing set to compute the empirical value function with the remaining 4 folds as training data. The results of the empirical value function based on five different folds are then averaged to produce a single estimation. The average results of the estimations and running time (seconds) over 100 replications are shown in Table 5.

As a comparison, we also compute the value function with the originally assigned treatments, which is 1.815. From Table 5, our method under the weighted linear loss and the hinge loss outperforms other methods in both linear and nonlinear cases, which implies that our method benefits from the hierarchical information in this study. HOAL_{wl} using linear learning is the best. Considering the running time (seconds), HOAL_{lin} is the fastest and HOAL_{wl} is comparable in computational efficiency. As for the optimal ITRs following the top–down strategy, we present the average proportions of treatment assignments over 100 replications by HOAL_{wl} using linear learning. It assigns around 60% patients into the insulin group and the rest 40% patients into the GLP-1 receptor agonists alone group at the second layer. At the third layer, around 46% patients are assigned into the insulin regimens including a short acting insulin group and around 10% into the intermediate-acting insulin alone group. The rest 4% patients are assigned into the long-acting insulin alone group, where 3% patients are assigned to the Glargine group and only 1% patients are assigned into the Detemir group. This result is consistent with the literature, which shows that the short-acting insulin and the GLP-1 have the benefit of reducing HbA1c [11, 1]. We also note that the prandial insulin is associated with greater risks of hypoglycemia and weight gain, while GLP-1 has slightly added benefits regarding bodyweight, hypoglycemia, and lipoproteins. Though the primary goal in this study here is the reduction of HbA1c, more composite metrics including HbA1c change, hypo events, and weight gain can be evaluated.

Table 5

Average empirical value function results as well as standard deviations and running time using 5-fold cross-validation over 100 replications for the T2DM dataset. The best value in each column is boldfaced.

	Linear		Gaussian	
	Value	Time (s)	Value	Time (s)
MOML	$2.711_{0.011}$	0.619	$2.721_{0.013}$	56.615
SMOML	$2.889_{0.017}$	0.622	$2.888_{0.018}$	27.862
SBOWL	$2.580_{0.014}$	1.265	$2.706_{0.015}$	50.565
HOAL_{lin}	$3.077_{0.016}$	0.163	$2.800_{0.018}$	1.166
HOAL_{wl}	$\mathbf{3.349}_{0.016}$	0.218	$2.948_{0.017}$	1.575
$\text{HOAL}_{\text{hinge}}$	$3.016_{0.017}$	4.111	$\mathbf{3.054}_{0.017}$	89.696

6. Discussion

In this paper, we study the estimation of optimal ITRs for treatments with hierarchical structure. We propose a new framework named hierarchical outcome-weighted angle-based learning (HOAL) to utilize the angle-based hierarchical classification method in OWL. By designing a linear loss function, a closed form solution can be derived, thus our method is computationally efficient. Statistical properties including Fisher consistency and convergence rate are studied. Simulations and a real data application demonstrate the advantage of our proposed method in both accuracy and computational efficiency.

There are several extensions can be considered. For example, sparse penalties can be used to select important variables. Furthermore, the proposed method can be integrated with the residual weighted learning [38, 15]. Our current focus is on continuous responses, it will be interesting to extend the method to handle other types of outcome such as binary and censored outcomes [22].

Appendix A: Conclusions and algorithms in Section 2

A.1. H. S. properties

For $T_{i_1,\dots,i_m}$ at the m-th layer and $T_{j_1,\dots,j_l}$ at the l-th layer, where $i_1 = j_1 \equiv 1$ and $1 \leq m, l \leq k$, define $I_{i_1,\dots,i_m;j_1,\dots,j_l}$ as the layer at which the latest common ancestor of $T_{i_1,\dots,i_m}$ and $T_{j_1,\dots,j_l}$ locates, that is,

$$I_{i_1,\dots,i_m;j_1,\dots,j_l} = \max\left\{t : (i_1,\dots,i_t) = (j_1,\dots,j_t), 1 \leq t \leq \min\{m,l\}\right\}.$$

Proposition S1. *For $\delta^2 \geq 2\sqrt{2}+2$, it holds that*

(1) (Hierarchical property) For any two pairs of points $\{\boldsymbol{\xi}_{i_1,\dots,i_m}, \boldsymbol{\xi}_{i'_1,\dots,i'_{\tilde{m}}}\}$ and $\{\boldsymbol{\xi}_{j_1,\dots,j_l}, \boldsymbol{\xi}_{j'_1,\dots,j'_{\tilde{l}}}\}$, if $I_{i_1,\dots,i_m;i'_1,\dots,i'_{\tilde{m}}} < I_{j_1,\dots,j_l;j'_1,\dots,j'_{\tilde{l}}}$, then

$$\|\boldsymbol{\xi}_{i_1,\dots,i_m} - \boldsymbol{\xi}_{i'_1,\dots,i'_{\tilde{m}}}\| > \|\boldsymbol{\xi}_{j_1,\dots,j_l} - \boldsymbol{\xi}_{j'_1,\dots,j'_{\tilde{l}}}\|.$$

(2) (Symmetric property) For any two pairs of points $\{\boldsymbol{\xi}_{i_1,\dots,i_m}, \boldsymbol{\xi}_{i'_1,\dots,i'_{\tilde{m}}}\}$ and $\{\boldsymbol{\xi}_{i_1,\dots,i_m}, \boldsymbol{\xi}_{j'_1,\dots,j'_{\tilde{m}}}\}$, if $I_{i_1,\dots,i_m;i'_1,\dots,i'_{\tilde{m}}} = I_{i_1,\dots,i_m;j'_1,\dots,j'_{\tilde{m}}}$, then

$$\|\boldsymbol{\xi}_{i_1,\dots,i_m} - \boldsymbol{\xi}_{i'_1,\dots,i'_{\tilde{m}}}\| = \|\boldsymbol{\xi}_{i_1,\dots,i_m} - \boldsymbol{\xi}_{j'_1,\dots,j'_{\tilde{m}}}\|.$$

A.2. Algorithm to embed nodes in a standard q-class multicategory classification problem

Denote $\boldsymbol{e}_i$ as the coordinate bases in $\mathbb{R}^m$, which is a zero vector except for the i-th coordinate being 1. For $\boldsymbol{U} = (u_1,\dots,u_m)^\top$, denote $\boldsymbol{U}^{(s)}$ as the subvector consisting of the first $s \leq m$ coordinates of $\boldsymbol{U}$, that is, $\boldsymbol{U}^{(s)} = (u_1,\dots,u_s)^\top$. Denote the embedded q points for a standard q-class multicategory classification problem in $\mathbb{R}^{q-1}$ with the given length L as $\boldsymbol{\xi}_i, i = 1,\dots,q$.

Algorithm S1 : Label embedding for the standard q-class multicategory classification

```
Initialization: Set $\xi_1^{(1)} = 1, \xi_2^{(1)} = -1$.
Iteration: For $m = 2, \dots, q-1$, repeat the following Steps (1) and (2).
    (1) Set $\boldsymbol{\xi}_i^{(m)} = ((\boldsymbol{\xi}_i^{(m-1)})^\top, 0)^\top \in \mathbb{R}^m, i = 1, \dots, m$.
    (2) $\boldsymbol{\xi}_{m+1}^{(m)} = m^{-1}\sum_{i=1}^m \boldsymbol{\xi}_i^{(m)} + a_m \boldsymbol{e}_m^{(m)}$, where $a_m = \sqrt{2^2 - d_{m-1}^2}$ with $d_{m-1} = \|m^{-1}\sum_{i=1}^m \boldsymbol{\xi}_i^{(m-1)} - \boldsymbol{\xi}_m^{(m-1)}\|$ and $\boldsymbol{e}_m$ is the coordinate base in $\mathbb{R}^m$.
Centralization: Let $\boldsymbol{\xi}_i \leftarrow \boldsymbol{\xi}_i - q^{-1}\sum_{j=1}^q \boldsymbol{\xi}_j, i = 1, \dots, q$.
Scaling: $\boldsymbol{\xi}_i \leftarrow L\|\boldsymbol{\xi}_i\|^{-1}\boldsymbol{\xi}_i$ for $i = 1, \dots, q$.
```

A.3. Algorithm to embed nodes in hierarchical classification

Algorithm S2 : Label embedding for hierarchical classification

```
Initialization: Initialize any $\boldsymbol{\xi}_{j_1,j_2,\dots,j_m} \in \mathbb{R}^K$ being a zero vector. For $m = 2$, let $D_2 = N_{j_1} - 1$. Construct subvectors $\{\boldsymbol{\xi}_{j_1,j_2}^{(D_2)} \in \mathbb{R}^{D_2}, j_2 = 1, \dots, N_{j_1}\}$ based on Algorithm S1 with a given norm $L^{(1)}$.
Iteration: Repeat the following Steps (1)–(3) for $m = 3, \dots, k$.
    (1) Sort all nonleaf nodes in $\mathcal{T}_{m-1}$ from left to right and rename them as $T_1^{(m-1)}, \dots, T_{n_{m-1}}^{(m-1)}$, where $n_{m-1}$ is the number of nonleaf nodes at the $(m-1)$-th layer. There exists some $(j_2', \dots, j_{m-1}')$ such that $T_i^{(m-1)} = T_{j_1',\dots,j_{m-1}'}$ for any $1 \le i \le n_{m-1}$ with children $\{T_{j_1',\dots,j_{m-1}',j_m}, j_m = 1, \dots, N_{j_1',\dots,j_{m-1}'}\}$ at the $m$-th layer where $N_{j_1',\dots,j_{m-1}'} \ge 2$ according to the assumption that each node either is a leaf or has at least two children. Let $d_{m,i} = N_{j_1',\dots,j_{m-1}'} - 1$.
    (2) Let $L^{(m-1)} = L^{(m-2)}/\delta$ with $\delta$ being the down-scaling constant given in advance. Given some $1 \le i \le n_{m-1}$, for $\mathrm{Chi}(T_i^{(m-1)})$, construct $N_{j_1',\dots,j_{m-1}'}$ points denoted as $\{\boldsymbol{\eta}_{j_1',\dots,j_{m-1}',j_m}, j_m = 1, \dots, N_{j_1',\dots,j_{m-1}'}\}$ based on Algorithm S1 with the given norm $L^{(m-1)}$ in the subspace span$\left\{\boldsymbol{e}_j \in \mathbb{R}^K : D_{m-1} + 1 + \sum_{s=0}^{i-1} d_{m,s} \le j \le D_{m-1} + \sum_{s=0}^{i} d_{m,s}\right\}$, where $d_{m,0} = 0$ and $\boldsymbol{e}_j$'s are the coordinate bases in $\mathbb{R}^K$. Let $\boldsymbol{\xi}_{j_1',\dots,j_{m-1}',j_m} = \boldsymbol{\xi}_{j_1',\dots,j_{m-1}'} + \boldsymbol{\eta}_{j_1',\dots,j_{m-1}',j_m}$, $j_m = 1, \dots, N_{j_1',\dots,j_{m-1}'}$.
    (3) Repeat Step (2) for all $n_{m-1}$ nonleaf nodes in $\mathcal{T}_{m-1}$. Let $D_m = D_{m-1} + \sum_{i=1}^{n_{m-1}} d_{m,i}$.
```

Appendix B: Proof of conclusions in Section 2

B.1. Proof of Proposition 1.

Proof of Proposition 1. We prove this by contradiction. Let

$$\mathcal{R}(\boldsymbol{f}) = E\left[\frac{|R|}{\Pr(A|\boldsymbol{X})}\{I(R \ge 0)I(M(\boldsymbol{f}(\boldsymbol{X}), A) < 0) + I(R < 0)I(M(\boldsymbol{f}(\boldsymbol{X}), A) \ge 0)\}\right].$$

Suppose $\mathcal{C}(\mathcal{D}^*) < \mathcal{C}(\mathcal{D}_{\bar{\boldsymbol{f}}})$. For $\mathcal{D}^*$, let $\boldsymbol{f}'(\boldsymbol{X}) = \boldsymbol{\xi}_{\mathcal{L}(\mathcal{D}^*)}(\mathcal{D}^*(\boldsymbol{X}))$, which is the embedded point corresponding to the leaf node of $\mathcal{D}^*(\boldsymbol{X})$. It can be seen that for any $m \geq 2$, the embedded point $\boldsymbol{\xi}_m(\mathcal{D}^*)$ of $(\mathcal{D}^*)^{(m)}$ has the largest inner product with $\boldsymbol{f}'$ among all child nodes of $(\mathcal{D}^*)^{(m-1)}$. Thus, by the top–down strategy, we have $\mathcal{D}_{\boldsymbol{f}'} = \mathcal{D}^*$. Moreover, since $I(A = \mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})) = I(M(\boldsymbol{f}(\boldsymbol{X}), A) \geq 0)$ and $I(A \neq \mathcal{D}_{\boldsymbol{f}}(\boldsymbol{x})) = I(M(\boldsymbol{f}(\boldsymbol{X}), A) < 0)$, for any $\boldsymbol{f}$, we have $\mathcal{C}(\mathcal{D}_{\boldsymbol{f}}) = \mathcal{R}(\boldsymbol{f})$. It holds that

$$\mathcal{R}(\bar{\boldsymbol{f}}) = \mathcal{C}(\mathcal{D}_{\bar{\boldsymbol{f}}}) \overset{(i)}{>} \mathcal{C}(\mathcal{D}^*) \overset{(ii)}{=} \mathcal{C}(\mathcal{D}_{\boldsymbol{f}'}) = \mathcal{R}(\boldsymbol{f}'), \tag{SB.1}$$

where (i) is derived by the assumption $\mathcal{C}(\mathcal{D}^*) < \mathcal{C}(\mathcal{D}_{\bar{\boldsymbol{f}}})$ and (ii) is from $\mathcal{D}_{\boldsymbol{f}'} = \mathcal{D}^*$. Since $\bar{\boldsymbol{f}}$ is the minimizer of $\mathcal{R}(\boldsymbol{f})$, (SB.1) leads to a contradiction. Hence, we have $\mathcal{C}(\mathcal{D}^*) = \mathcal{C}(\mathcal{D}_{\bar{\boldsymbol{f}}})$. That is, $\mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{X})$ is a minimizer of $\mathcal{C}(\mathcal{D})$. This completes the proof. □

B.2. Proof of (2.10)

Proof of (2.10). We prove that for linear learning, the estimator under the linear loss has a closed form.

It holds that

$$\begin{aligned}
&\hat{\boldsymbol{C}}_{\text{lin},\lambda}, \hat{\boldsymbol{b}}_{\text{lin},\lambda,\lambda'} \\
&= \text{argmin} \frac{1}{n} \sum_{i=1}^{n} \sum_{m=2}^{\mathcal{L}(a_i)} \sum_{\tilde{a} \in [\mathscr{E}_m(a_i)]} \frac{r_i}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)} (\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))^\top (\boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{b}) + \\
&\quad \lambda \|\boldsymbol{C}\|_F^2 + \lambda\lambda' \|\boldsymbol{b}\|^2 \\
&= \text{argmin} \frac{1}{n} \sum_{i=1}^{n} \sum_{m=2}^{\mathcal{L}(a_i)} \sum_{\tilde{a} \in [\mathscr{E}_m(a_i)]} tr \left(\frac{r_i}{\Pr(a_i^{(m)}|\boldsymbol{x}_i)} (\boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{b})(\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i))^\top \right) + \\
&\quad \lambda \|\boldsymbol{C}\|_F^2 + \lambda\lambda' \|\boldsymbol{b}\|^2 \\
&= \text{argmin} tr(\boldsymbol{C}\boldsymbol{B}_1^\top + \boldsymbol{b}\tilde{\boldsymbol{b}}_1^\top) + \lambda \|\boldsymbol{C}\|_F^2 + \lambda\lambda' \|\boldsymbol{b}\|^2,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{B}_1 &= n^{-1} \sum_{i=1}^{n} \sum_{m=2}^{\mathcal{L}(a_i)} \sum_{\tilde{a} \in [\mathscr{E}_m(a_i)]} w_{i,m} (\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i)) \boldsymbol{x}_i^\top, \\
\tilde{\boldsymbol{b}}_1 &= n^{-1} \sum_{i=1}^{n} \sum_{m=2}^{\mathcal{L}(a_i)} \sum_{\tilde{a} \in [\mathscr{E}_m(a_i)]} w_{i,m} (\boldsymbol{\xi}_m(\tilde{a}) - \boldsymbol{\xi}_m(a_i)),
\end{aligned}$$

with $w_{i,m} = r_i / \Pr(a_i^{(m)}|\boldsymbol{x}_i)$. Thus, the solution of $\hat{\boldsymbol{C}}_{\text{lin},\lambda}, \hat{\boldsymbol{b}}_{\text{lin},\lambda,\lambda'}$ has the closed form,

$$\hat{\boldsymbol{C}}_{\text{lin},\lambda} = -\boldsymbol{B}_1/(2\lambda), \hat{\boldsymbol{b}}_{\text{lin},\lambda,\lambda'} = -\tilde{\boldsymbol{b}}_1/(2\lambda\lambda').$$

Appendix C: Proof of conclusions in Section 3

C.1. Proof of Proposition 2

Proof of Proposition 2. For each $\boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
&E\left[\frac{|R|}{\Pr(A|\boldsymbol{X}=\boldsymbol{x})}\{I(R\geq 0)I(A\neq \mathcal{D}(\boldsymbol{x}))+I(R<0)I(A=\mathcal{D}(\boldsymbol{x}))\}|\boldsymbol{X}=\boldsymbol{x}\right]\\
=&\sum_a E\left[|R|\{I(R\geq 0)I(a\neq \mathcal{D}(\boldsymbol{x}))+I(R<0)I(a=\mathcal{D}(\boldsymbol{x}))\}|\boldsymbol{X}=\boldsymbol{x},A=a\right]\\
=&\sum_a [R^+(\boldsymbol{x},a)I(a\neq \mathcal{D}(\boldsymbol{x}))-R^-(\boldsymbol{x},a)I(a=\mathcal{D}(\boldsymbol{x}))].
\end{aligned}
$$

We prove the conclusion by a contradiction. Suppose $\mathcal{D}^*(\boldsymbol{x}) \neq \operatorname{argmax}_a R(\boldsymbol{x},a)$ and denote $a^* = \operatorname{argmax}_a R(\boldsymbol{x},a)$. It holds that

$$
\begin{aligned}
&\sum_a [R^+(\boldsymbol{x},a)I(a\neq \mathcal{D}^*(\boldsymbol{x}))-R^-(\boldsymbol{x},a)I(a=\mathcal{D}^*(\boldsymbol{x}))]-\\
&\sum_a [R^+(\boldsymbol{x},a)I(a\neq a^*)-R^-(\boldsymbol{x},a)I(a=a^*)]\\
=&R^+(\boldsymbol{x},a^*)-R^-(\boldsymbol{x},\mathcal{D}^*(\boldsymbol{x}))-R^+(\boldsymbol{x},\mathcal{D}^*(\boldsymbol{x}))+R^-(\boldsymbol{x},a^*)\\
=&R(\boldsymbol{x},a^*)-R(\boldsymbol{x},\mathcal{D}^*(\boldsymbol{x}))>0,
\end{aligned}
$$

which leads to a contradiction that $\mathcal{D}^*(\boldsymbol{x})$ is optimal. Thus, we have $\mathcal{D}^*(\boldsymbol{x}) = \operatorname{argmax}_a R(\boldsymbol{x},a)$. This completes the proof. □

C.2. Proof of Theorem 1

Proof of Theorem 1. For any $a \in \mathcal{A}$ and $m = 2,\dots,\mathcal{L}(a)$, denote $\boldsymbol{\eta}_m(a) = \boldsymbol{\xi}_m(a) - \boldsymbol{\xi}_{m-1}(a)$. We have

$$
\begin{aligned}
\boldsymbol{f}^* =&\operatorname*{arginf}_{\boldsymbol{f}} E\left[\sum_{m=2}^{\mathcal{L}(A)}\sum_{\tilde{A}\in[\mathscr{E}_m(A)]}\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{x})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}\right]\\
=&\operatorname*{arginf}_{\boldsymbol{f}}\sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]} E\left[\frac{|R|\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}\ell_R((\boldsymbol{\xi}_m(a)-\boldsymbol{\xi}_m(\tilde{a}))^\top \boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{x},a\right]\\
=&\operatorname*{arginf}_{\boldsymbol{f}}\sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}[R^+(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top \boldsymbol{f}(\boldsymbol{x}))-\\
&R^-(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top \boldsymbol{f}(\boldsymbol{x}))]\\
=&\operatorname*{arginf}_{\boldsymbol{f}}\sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\dots,j_{m-1}}}\sum_{j'_m\neq j_m}[R_m^+(\boldsymbol{x},T_{j_1,\dots,j_m})\ell((\boldsymbol{\eta}_{j_1,\dots,j_m}-\\
&\boldsymbol{\eta}_{j_1,\dots,j'_m})^\top \boldsymbol{f}(\boldsymbol{x}))-R_m^-(\boldsymbol{x},T_{j_1,\dots,j_m})\ell((\boldsymbol{\eta}_{j_1,\dots,j'_m}-\boldsymbol{\eta}_{j_1,\dots,j_m})^\top \boldsymbol{f}(\boldsymbol{x}))].
\end{aligned}
$$

To show Fisher consistency, according to the top–down strategy defined in Definition 1, by Assumption 1, it is sufficient to show for any $m = 2, \ldots, \mathcal{L}(\mathcal{D}^*(\boldsymbol{x}))$ and any $\tilde{a} \in [\mathscr{E}_m(\mathcal{D}^*(\boldsymbol{x}))]$,

$$\langle \boldsymbol{\xi}_m(\mathcal{D}^*(\boldsymbol{x})), \boldsymbol{f}^*(\boldsymbol{x})\rangle > \langle \boldsymbol{\xi}_m(\tilde{a}), \boldsymbol{f}^*(\boldsymbol{x})\rangle$$

which equals to

$$\langle \boldsymbol{\eta}_m(\mathcal{D}^*(\boldsymbol{x})), \boldsymbol{f}^*(\boldsymbol{x})\rangle > \langle \boldsymbol{\eta}_m(\tilde{a}), \boldsymbol{f}^*(\boldsymbol{x})\rangle,$$

based on the construction of points.

We prove this by contradiction. Suppose that there exists some l such that

$$a' = \underset{\tilde{a}\in[\mathscr{E}_l(\mathcal{D}^*(\boldsymbol{x}))]}{\operatorname{argmax}} \langle \boldsymbol{\eta}_l(\tilde{a}), \boldsymbol{f}^*(\boldsymbol{x})\rangle,$$

and

$$\langle \boldsymbol{\eta}_l(\mathcal{D}^*(\boldsymbol{x})), \boldsymbol{f}^*(\boldsymbol{x})\rangle < \langle \boldsymbol{\eta}_l(a'), \boldsymbol{f}^*(\boldsymbol{x})\rangle. \tag{SC.1}$$

Denote $D^{*(l)}(\boldsymbol{x}) = T_{i_1,\ldots,i_l}$ and $a'^{(l)} = T_{i_1,\ldots,i_{l-1},i'_l}$. By Lemma S3 in the Appendix of Fan et al. [9], for any $\epsilon > 0$, there exists some $\kappa > 0$ such that

$$\begin{aligned}\langle \kappa(\boldsymbol{\eta}_{i_1,\ldots,i_l} - \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l}), \boldsymbol{\eta}_{i_1,\ldots,i_l}))\rangle &= \epsilon,\\ \langle \kappa(\boldsymbol{\eta}_{i_1,\ldots,i_l} - \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l}), \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l}\rangle &= -\epsilon.\end{aligned}$$

Let $\widetilde{\boldsymbol{f}^*} = \boldsymbol{f}^* + \kappa(\boldsymbol{\eta}_{i_1,\ldots,i_l} - \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l})$. It holds that for any $\boldsymbol{\eta}_{j_1,\ldots,j_m}$,

$$\langle \boldsymbol{\eta}_{j_1,\ldots,j_m}, \widetilde{\boldsymbol{f}^*} - \boldsymbol{f}^*\rangle = \begin{cases}\epsilon, & m = l, (j_1,\ldots,j_m) = (i_1,\ldots,i_l),\\ -\epsilon, & m = l, (j_1,\ldots,j_m) = (i_1,\ldots,i_{l-1},i'_l),\\ 0, & \text{otherwise.}\end{cases}$$

For simplicity, denote $\boldsymbol{\Lambda}_{j_1,\ldots,j_m,j'_m} = \boldsymbol{\eta}_{j_1,\ldots,j_m} - \boldsymbol{\eta}_{j_1,\ldots,j'_m}$ Then we have

$$\begin{aligned}&\mathcal{R}_V(\widetilde{\boldsymbol{f}^*}) - \mathcal{R}_V(\boldsymbol{f}^*)\\ &= \sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\ldots,j_{m-1}}}\sum_{j'_m\neq j_m}[R_m^+(\boldsymbol{x}, T_{j_1,\ldots,j_m})\ell(\boldsymbol{\Lambda}_{j_1,\ldots,j_m,j'_m}^\top\widetilde{\boldsymbol{f}^*}(\boldsymbol{x})) -\\ &\quad R_m^-(\boldsymbol{x}, T_{j_1,\ldots,j_m})\ell(\boldsymbol{\Lambda}_{j_1,\ldots,j'_m,j_m}^\top\widetilde{\boldsymbol{f}^*}(\boldsymbol{x})) - R_m^+(\boldsymbol{x}, T_{j_1,\ldots,j_m})\ell(\boldsymbol{\Lambda}_{j_1,\ldots,j_m,j'_m}^\top\boldsymbol{f}^*(\boldsymbol{x})) +\\ &\quad R_m^-(\boldsymbol{x}, T_{j_1,\ldots,j_m})\ell(\boldsymbol{\Lambda}_{j_1,\ldots,j'_m,j_m}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\ &= \sum_{j_2=1}^{N_1}\cdots\sum_{j_l=1}^{N_{j_1,\ldots,j_{l-1}}}\sum_{j'_l\neq j_l} R_l^+(\boldsymbol{x}, T_{j_1,\ldots,j_l})[\boldsymbol{\Lambda}_{j_1,\ldots,j_l,j'_l}^\top(\widetilde{\boldsymbol{f}^*} - \boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j_l,j'_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\ &\quad - R_l^-(\boldsymbol{x}, T_{j_1,\ldots,j_l})[\boldsymbol{\Lambda}_{j_1,\ldots,j'_l,j_l}^\top(\widetilde{\boldsymbol{f}^*} - \boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j'_l,j_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))] + o(2\epsilon).\end{aligned}$$

Define

$$
\begin{aligned}
&B_1\\
&=\sum_{j_2=1}^{N_1}\cdots\sum_{j_l=1}^{N_{j_1,\ldots,j_{l-1}}}\sum_{j_l'\neq j_l} R_l^+(\boldsymbol{x},T_{j_1,\ldots,j_l})\boldsymbol{\eta}_{j_1,\ldots,j_l}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&=\sum_{j_l'\neq i_l} R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\boldsymbol{\eta}_{i_1,\ldots,i_l}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+\\
&\quad\sum_{j_l'\neq i_l'} R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l'}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&=\sum_{j_l'\neq i_l} \epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\\
&\quad\sum_{j_l'\neq i_l'} \epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x})),
\end{aligned}
$$

and

$$
\begin{aligned}
&B_2\\
&=\sum_{j_2=1}^{N_1}\cdots\sum_{j_l=1}^{N_{j_1,\ldots,j_{l-1}}}\sum_{j_l'\neq j_l} R_l^+(\boldsymbol{x},T_{j_1,\ldots,j_l})\boldsymbol{\eta}_{j_1,\ldots,j_l'}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&=\epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&\quad+\sum_{j_l\neq i_l,i_l'}\sum_{j_l'\neq j_l} R_l^+(\boldsymbol{x},T_{i_1,\ldots,j_l})\epsilon[\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l,i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\
&\geq\epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\epsilon R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&\geq\epsilon\ell'(0)[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})-R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})].
\end{aligned}
$$

The last two inequalities are derived from (SC.1),

$$\boldsymbol{\eta}_{i_1,\ldots,i_l}^\top\boldsymbol{f}^*<\boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l'}^\top\boldsymbol{f}^*,$$

and $\ell'(u)$ is nondecreasing.

Define

$$
\begin{aligned}
&B_3\\
&=\sum_{j_2=1}^{N_1}\cdots\sum_{j_l=1}^{N_{j_1,\ldots,j_{l-1}}}\sum_{j_l'\neq j_l} R_l^-(\boldsymbol{x},T_{j_1,\ldots,j_l})\boldsymbol{\eta}_{j_1,\ldots,j_l'}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j_l',j_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&=\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&\quad+\sum_{j_l\neq i_l,i_l'}\sum_{j_l'\neq j_l} R_l^-(\boldsymbol{x},T_{i_1,\ldots,j_l})\epsilon[\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,j_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\
&\geq\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))\\
&\geq\epsilon\ell'(0)[R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})].
\end{aligned}
$$

The last two inequalities are derived from (SC.1),

$$\boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{f}^* < \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l'}^\top \boldsymbol{f}^*,$$

and $\ell'(u)$ is nondecreasing.

Define

$$\begin{aligned}
&B_4\\
=&\sum_{j_2=1}^{N_1}\cdots\sum_{j_l=1}^{N_{j_1,\ldots,j_{l-1}}}\sum_{j_l'\neq j_l} R_l^-(\boldsymbol{x},T_{j_1,\ldots,j_l})\boldsymbol{\eta}_{j_1,\ldots,j_l}^\top(\widetilde{\boldsymbol{f}^*}-\boldsymbol{f}^*)\ell'(\boldsymbol{\Lambda}_{j_1,\ldots,j_l',j_l}^\top\boldsymbol{f}^*(\boldsymbol{x})))\\
=&\sum_{j_l'\neq i_l}\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\\
&\sum_{j_l'\neq i_l'}\epsilon R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x})).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&B_1+B_4\\
=&\sum_{j_l'\neq i_l}\epsilon[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))]-\\
&\sum_{j_l'\neq i_l'}\epsilon[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\
=&\sum_{j_l'\neq i_l,i_l'}\epsilon[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+\\
&R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))]+\\
&\epsilon[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))+\\
&R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',i_l}^\top\boldsymbol{f}^*(\boldsymbol{x}))-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l,i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))]\\
\overset{(i)}{\leq}&\sum_{j_l'\neq i_l,i_l'}\epsilon[(R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'}))\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+\\
&(R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'}))\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))]+\\
&\epsilon\ell'(0)[R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l^+(\boldsymbol{x},T_{i_1,\ldots,i_l'})+R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l'})]\\
=&\sum_{j_l'\neq i_l,i_l'}\epsilon[(R_l(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'}))\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))+\\
&(R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l^-(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'}))(\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,j_l',i_l'}^\top\boldsymbol{f}^*(\boldsymbol{x}))-\\
&\ell'(\boldsymbol{\Lambda}_{i_1,\ldots,i_l',j_l'}^\top\boldsymbol{f}^*(\boldsymbol{x})))]+\epsilon\ell'(0)[R_l(\boldsymbol{x},T_{i_1,\ldots,i_l})-R_l(\boldsymbol{x},T_{i_1,\ldots,i_{l-1},i_l'})]\\
&<0,
\end{aligned}$$

where the inequality (i) is derived from

$$\boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{f}^* < \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l'}^\top \boldsymbol{f}^*,$$

and $\ell'(u)$ is nondecreasing, and the last inequality is derived from Assumption 1, Assumption 2, the definition of a', $\ell'(u) < 0$ and $\ell'(u)$ is nondecreasing. Furthermore,

$$\begin{aligned}
&B_2 + B_3\\
\geq&\epsilon\ell'(0)[R_l^+(\boldsymbol{x}, T_{i_1,\ldots,i_l'}) - R_l^+(\boldsymbol{x}, T_{i_1,\ldots,i_l}) + R_l^-(\boldsymbol{x}, T_{i_1,\ldots,i_l'}) - R_l^-(\boldsymbol{x}, T_{i_1,\ldots,i_l})]\\
=&\epsilon\ell'(0)[R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},i_l'}) - R_l(\boldsymbol{x}, T_{i_1,\ldots,i_l})] > 0.
\end{aligned}$$

Hence, we have

$$B_1 - B_2 - B_3 + B_4 < 0,$$

which leads to a contradiction of the optimality of $\boldsymbol{f}^*$.

This completes the proof. □

C.3. Proof of the conclusion in Remark 3

Proof of Remark 3. For the linear loss, we have

$$\begin{aligned}
&E\left[\sum_{m=2}^{\mathcal{L}(A)}\sum_{\tilde{A}\in[\mathscr{E}_m(A)]}\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{x})}\ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}\right]\\
=&-\sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\ldots,j_{m-1}}}\sum_{j_m'\neq j_m}[R_m^+(\boldsymbol{x}, T_{j_1,\ldots,j_m})+\\
&R_m^-(\boldsymbol{x}, T_{j_1,\ldots,j_m})](\boldsymbol{\eta}_{j_1,\ldots,j_m} - \boldsymbol{\eta}_{j_1,\ldots,j_m'})^\top \boldsymbol{f}(\boldsymbol{x})\\
=&-\sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\ldots,j_{m-1}}}\sum_{j_m'\neq j_m}R_m(\boldsymbol{x}, T_{j_1,\ldots,j_m})(\boldsymbol{\eta}_{j_1,\ldots,j_m} - \boldsymbol{\eta}_{j_1,\ldots,j_m'})^\top \boldsymbol{f}(\boldsymbol{x})\\
=&-\sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\ldots,j_{m-1}}}|\mathrm{Chi}(\boldsymbol{\eta}_{j_1,\ldots,j_{m-1}})|R_m(\boldsymbol{x}, T_{j_1,\ldots,j_m})\boldsymbol{\eta}_{j_1,\ldots,j_m}^\top \boldsymbol{f}(\boldsymbol{x})\\
=&-\boldsymbol{V}_0^\top \boldsymbol{f}(\boldsymbol{x}),
\end{aligned}$$

where

$$\boldsymbol{V}_0 = \sum_{m=2}^{k}\sum_{j_2=1}^{N_1}\cdots\sum_{j_m=1}^{N_{j_1,\ldots,j_{m-1}}}|\mathrm{Chi}(\boldsymbol{\eta}_{j_1,\ldots,j_{m-1}})|R_m(\boldsymbol{x}, T_{j_1,\ldots,j_m})\boldsymbol{\eta}_{j_1,\ldots,j_m}.$$

It can be verified that $E(\|\boldsymbol{V}_0\|^2) = 0$ leads to the trivial case. Thus, it is reasonable to assume $E(\|\boldsymbol{V}_0\|^2) > 0$. The minimizer is of the form

$$\boldsymbol{f}^* = \lim_{\rho\to\infty}\rho\boldsymbol{V}_0.$$

Note that all learners in $\{\rho\boldsymbol{V}_0 : \rho > 0\}$ lead to the same result. Therefore, we define

$$\boldsymbol{f}^* = \operatorname*{arginf}_{\boldsymbol{f}:E(\|\boldsymbol{f}\|^2)\leq 1} R_V(\boldsymbol{f}) = \boldsymbol{V}_0/[E(\|\boldsymbol{V}_0\|^2)]^{1/2}$$

with $\rho = [E(\|\boldsymbol{V}_0\|^2)]^{-1/2} > 0$.

Then for any two siblings $T_{i_1,\ldots,i_{l-1},i_l}, T_{i_1,\ldots,i_{l-1},i'_l}$, if

$$R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},i_l}) > R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},i'_l}),$$

it holds that

$$\begin{aligned}
&\boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l}^\top \boldsymbol{f}^* \\
=&\frac{1}{[E(\|\boldsymbol{V}_0\|^2)]^{1/2}} \sum_{j_l=1}^{N_{i_1,\ldots,i_{l-1}}} |\mathrm{Chi}(\boldsymbol{\eta}_{i_1,\ldots,i_{l-1}})| R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},j_l}) \boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},j_l} \\
=&\frac{1}{[E(\|\boldsymbol{V}_0\|^2)]^{1/2}} |\mathrm{Chi}(\boldsymbol{\eta}_{i_1,\ldots,i_{l-1}})| \Bigg[\sum_{j_l \neq i_l, i'_l} R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},j_l}) \boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},j_l} \\
&+ R_l(\boldsymbol{x}, T_{i_1,\ldots,i_l}) \boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l} + R_l(\boldsymbol{x}, T_{i_1,\ldots,i'_l}) \boldsymbol{\eta}_{i_1,\ldots,i_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l} \Bigg] \\
>&\frac{1}{[E(\|\boldsymbol{V}_0\|^2)]^{1/2}} |\mathrm{Chi}(\boldsymbol{\eta}_{i_1,\ldots,i_{l-1}})| \Bigg[\sum_{j_l \neq i_l, i'_l} R_l(\boldsymbol{x}, T_{i_1,\ldots,i_{l-1},j_l}) \boldsymbol{\eta}_{i_1,\ldots,i'_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},j_l} \\
&+ R_l(\boldsymbol{x}, T_{i_1,\ldots,i_l}) \boldsymbol{\eta}_{i_1,\ldots,i'_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i_l} + R_l(\boldsymbol{x}, T_{i_1,\ldots,i'_l}) \boldsymbol{\eta}_{i_1,\ldots,i'_l}^\top \boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l} \Bigg] \\
=&\boldsymbol{\eta}_{i_1,\ldots,i_{l-1},i'_l}^\top \boldsymbol{f}^*.
\end{aligned}$$

Thus, it is also Fisher consistent.

This completes the proof. □

C.4. Proof of the conclusion in Remark 4

Theorem S1. *Assuming* $\max_{a \in \mathcal{A}} R(\boldsymbol{x}, a) \geq 0$*, it holds that* $\mathcal{D}_{\boldsymbol{f}_M}(\boldsymbol{x}) = \mathcal{D}_{\bar{\boldsymbol{f}}}(\boldsymbol{x})$ *for any nonincreasing and convex surrogate loss* $\ell(u)$*.*

Proof of Theorem S1. We prove that the minimizer $\boldsymbol{f}_M$ of

$$E\left[|R|/\Pr(A|\boldsymbol{X})\ell_R(M(\boldsymbol{f}(\boldsymbol{X}), A))\right]$$

satisfies $\mathcal{D}_{\boldsymbol{f}_M} = \mathcal{D}_{\bar{\boldsymbol{f}}}$. We prove this by contradiction. Given $\boldsymbol{x}$, let $\mathcal{D}_{\boldsymbol{f}_M} = a_M$ and $M(\boldsymbol{f}_M(\boldsymbol{x}), a_M) = \hat{M}$. By the top–down strategy, we have that $\hat{M} \geq 0$. For any $a \neq a_M$, there exists $2 \leq l \leq k$ such that $a^{(l)}$ and $a_M^{(l)}$ are siblings. Therefore, we have

$$M(\boldsymbol{f}_M(\boldsymbol{x}), a) \leq \langle \boldsymbol{f}_M(\boldsymbol{x}), \boldsymbol{\xi}_l(a) \rangle - \langle \boldsymbol{f}_M(\boldsymbol{x}), \boldsymbol{\xi}_l(a_M) \rangle \leq -M(\boldsymbol{f}_M(\boldsymbol{x}), a_M) = -\hat{M}.$$

Let $\mathcal{D}_{\bar{\boldsymbol{f}}} = \bar{a}$ and assume $a_M \neq \bar{a}$. For any $T_{i_1,\ldots,i_l} \in \bar{a}$ and $T_{i_1,\ldots,i'_l} \in \mathrm{Sib}(T_{i_1,\ldots,i_l})$, there exists $\kappa_{i_1,\ldots,i_l}$ such that $\langle \boldsymbol{\eta}_{i_1,\ldots,i_l} - \boldsymbol{\eta}_{i_1,\ldots,i'_l}, \kappa_{i_1,\ldots,i_l} \boldsymbol{\eta}_{i_1,\ldots,i_l} \rangle =$

$\hat{M}$. Let $\tilde{\boldsymbol{f}}(\boldsymbol{x}) = \sum_{T_{i_1,\ldots,i_l} \in \bar{a}} \kappa_{i_1,\ldots,i_l} \boldsymbol{\eta}_{i_1,\ldots,i_l}$. One can verify $M(\tilde{\boldsymbol{f}}(\boldsymbol{x}), \bar{a}) = \hat{M}$. For any $a \neq \bar{a}$, there exists l_a such that $a^{(l_a)}$ and $\bar{a}^{(l_a)}$ are siblings. Thus, we have

$$M(\tilde{\boldsymbol{f}}(\boldsymbol{x}), a) = \langle \tilde{\boldsymbol{f}}(\boldsymbol{x}), \boldsymbol{\xi}_{l_a}(a) \rangle - \langle \tilde{\boldsymbol{f}}(\boldsymbol{x}), \boldsymbol{\xi}_{l_a}(\bar{a}) \rangle = -\hat{M}.$$

For any $a \neq a_M, \bar{a}$, there exists $\kappa_a < 0$ such that $\langle \kappa_a \boldsymbol{\eta}_{l_a}(a), \boldsymbol{\xi}_{l_a}(a) - \boldsymbol{\xi}_{l_a}(\bar{a}) \rangle = M(\boldsymbol{f}_M(\boldsymbol{x}), a) + \hat{M}$. Let $\tilde{\boldsymbol{f}}'(\boldsymbol{x}) = \tilde{\boldsymbol{f}}(\boldsymbol{x}) + \sum_{a \neq a_M, \bar{a}} \kappa_a \boldsymbol{\eta}_{l_a}(a)$. It holds that

$$M(\tilde{\boldsymbol{f}}'(\boldsymbol{x}), a) = \begin{cases} -\hat{M}, & a = a_M, \\ \hat{M}, & a = \bar{a}, \\ M(\boldsymbol{f}_M(\boldsymbol{x}), a), & a \neq a_M, \bar{a}. \end{cases}$$

Hence, we have

$$\begin{aligned}
& E\left[\frac{|R|}{\Pr(A|\boldsymbol{x})} \ell_R(M(\boldsymbol{f}_M, A)) | \boldsymbol{X} = \boldsymbol{x}\right] - E\left[\frac{|R|}{\Pr(A|\boldsymbol{x})} \ell_R(M(\tilde{\boldsymbol{f}}', A)) | \boldsymbol{X} = \boldsymbol{x}\right] \\
= & \{R^+(\boldsymbol{x}, a_M)\ell(M(\boldsymbol{f}_M, a_M)) - R^-(\boldsymbol{x}, a_M)\ell(-M(\boldsymbol{f}_M, a_M)) \\
& + R^+(\boldsymbol{x}, \bar{a})\ell(M(\boldsymbol{f}_M, \bar{a})) - R^-(\boldsymbol{x}, \bar{a})\ell(-M(\boldsymbol{f}_M, \bar{a}))\} \\
& - \{R^+(\boldsymbol{x}, a_M)\ell(M(\tilde{\boldsymbol{f}}', a_M)) - R^-(\boldsymbol{x}, a_M)\ell(-M(\tilde{\boldsymbol{f}}', a_M)) \\
& + R^+(\boldsymbol{x}, \bar{a})\ell(M(\tilde{\boldsymbol{f}}', \bar{a})) - R^-(\boldsymbol{x}, \bar{a})\ell(-M(\tilde{\boldsymbol{f}}', \bar{a}))\} \\
= & \{R^+(\boldsymbol{x}, a_M)\ell(\hat{M}) - R^-(\boldsymbol{x}, a_M)\ell(-\hat{M}) \\
& + R^+(\boldsymbol{x}, \bar{a})\ell(M(\boldsymbol{f}_M, \bar{a})) - R^-(\boldsymbol{x}, \bar{a})\ell(-M(\boldsymbol{f}_M, \bar{a}))\} \\
& - \{R^+(\boldsymbol{x}, a_M)\ell(-\hat{M}) - R^-(\boldsymbol{x}, a_M)\ell(\hat{M}) + R^+(\boldsymbol{x}, \bar{a})\ell(\hat{M}) - R^-(\boldsymbol{x}, \bar{a})\ell(-\hat{M})\} \\
= & \{R^+(\boldsymbol{x}, a_M) - R^-(\boldsymbol{x}, \bar{a})\}\ell(\hat{M}) - \{R^-(\boldsymbol{x}, a_M) - R^+(\boldsymbol{x}, \bar{a})\}\ell(-\hat{M}) \\
& - \{R^+(\boldsymbol{x}, a_M) - R^-(\boldsymbol{x}, \bar{a})\}\ell(-\hat{M}) + \{R^-(\boldsymbol{x}, a_M) - R^+(\boldsymbol{x}, \bar{a})\}\ell(\hat{M}) \\
& + R^+(\boldsymbol{x}, \bar{a})\{\ell(M(\boldsymbol{f}_M, \bar{a})) - \ell(-\hat{M})\} - R^-(\boldsymbol{x}, \bar{a})\{\ell(-M(\boldsymbol{f}_M, \bar{a})) - \ell(\hat{M})\} \\
= & \{R^+(\boldsymbol{x}, a_M) - R^-(\boldsymbol{x}, \bar{a})\}\{\ell(\hat{M}) - \ell(-\hat{M})\} \\
& - \{R^-(\boldsymbol{x}, a_M) - R^+(\boldsymbol{x}, \bar{a})\}\{\ell(-\hat{M}) - \ell(\hat{M})\} \\
& + R^+(\boldsymbol{x}, \bar{a})\{\ell(M(\boldsymbol{f}_M, \bar{a})) - \ell(-\hat{M})\} - R^-(\boldsymbol{x}, \bar{a})\{\ell(-M(\boldsymbol{f}_M, \bar{a})) - \ell(\hat{M})\} \\
\overset{(i)}{>} & \{R^+(\boldsymbol{x}, \bar{a}) - R^-(\boldsymbol{x}, a_M)\}\{\ell(\hat{M}) - \ell(-\hat{M}) + \ell(-\hat{M}) - \ell(\hat{M})\} \\
& + R^+(\boldsymbol{x}, \bar{a})\{\ell(M(\boldsymbol{f}_M, \bar{a})) - \ell(-\hat{M})\} - R^-(\boldsymbol{x}, \bar{a})\{\ell(-M(\boldsymbol{f}_M, \bar{a})) - \ell(\hat{M})\} \\
\overset{(ii)}{\geq} & 0,
\end{aligned}$$

where (i) is derived from $R(\boldsymbol{x}, \bar{a}) > R(\boldsymbol{x}, a_M)$ and (ii) is derived from the assumption $R(\boldsymbol{x}, \bar{a}) \geq 0$ and the fact $\ell(-t-u) - \ell(-t) \geq \ell(t) - \ell(t+u)$ when ℓ is convex.

This completes the proof. □

C.5. Verification of Assumptions 3–5

In this subsection, we verify Assumptions 3–5 under Assumption 6, thus the specific convergence rate can be derived based on Corollary 1.

Assumption 3 depends on $\boldsymbol{f}^*$ and the distribution of $(\boldsymbol{X}, A, R)$. There is no general result about α. We give an example to show the value of α in that case later. As for Assumption 4, we have the following Proposition S2. Define the positive part of the reward R as $R^+ = R \cdot I(R \geq 0)$ and the negative part as $R^- = R \cdot I(R < 0)$.

Proposition S2. *Under Assumption 6, it holds that $\beta = 1$ for the linear loss with the restriction $E(\|\boldsymbol{f}\|^2) \leq 1$ and loss functions that are twice continuously differentiable with $\ell''(u) > 0$ in Assumption 4.*

Proof of Proposition S2. Note that

$$
\begin{aligned}
&|V^{\tilde{T}}(\boldsymbol{f}, Z) - V(\boldsymbol{f}^*, Z)| \\
&\leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})} |\ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})) - \\
&\quad \ell_R((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}^*(\boldsymbol{X}))| \\
&\leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{1}{\Pr(A^{(m)}|\boldsymbol{X})} |R^+ \ell((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})) - \\
&\quad R^- \ell((\boldsymbol{\xi}_m(\tilde{A}) - \boldsymbol{\xi}_m(A))^\top \boldsymbol{f}(\boldsymbol{X})) - \\
&\quad R^+ \ell((\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}^*(\boldsymbol{X})) + R^- \ell((\boldsymbol{\xi}_m(\tilde{A}) - \boldsymbol{\xi}_m(A))^\top \boldsymbol{f}^*(\boldsymbol{X}))| \\
&\leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{1}{\Pr(A^{(m)}|\boldsymbol{X})} [|R^+ \tilde{\gamma}(\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top (\boldsymbol{f} - \boldsymbol{f}^*)| + \\
&\quad |R^- \tilde{\gamma}(\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A}))^\top (\boldsymbol{f} - \boldsymbol{f}^*)|] \\
&\leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{\tilde{\gamma}}{\Pr(A^{(m)}|\boldsymbol{X})} (|R^+| + |R^-|) \|\boldsymbol{\xi}_m(A) - \boldsymbol{\xi}_m(\tilde{A})\| \|\boldsymbol{f} - \boldsymbol{f}^*\| \\
&\leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A} \in [\mathscr{E}_m(A)]} \frac{4\tilde{\gamma} M_R L^{(m)}}{M_A} \|\boldsymbol{f} - \boldsymbol{f}^*\| \leq \frac{4\tilde{\gamma}(k-1) N_{\text{Sib}} M_R L^{(1)}}{M_A} \|\boldsymbol{f} - \boldsymbol{f}^*\|,
\end{aligned}
$$

where N_{Sib} denotes the maximum number of sibling nodes on the tree. Define $M_1 = 4\tilde{\gamma}(k-1) N_{\text{Sib}} M_R L^{(1)}/M_A$, then we have $|V^{\tilde{T}}(\boldsymbol{f}, Z) - V(\boldsymbol{f}^*, Z)| \leq M_1 \|\boldsymbol{f} - \boldsymbol{f}^*\|$. Furthermore, it holds that

$$
\operatorname{Var}(V^{\tilde{T}}(\boldsymbol{f}, Z) - V(\boldsymbol{f}^*, Z)) \leq M_1^2 E(\|\boldsymbol{f} - \boldsymbol{f}^*\|^2). \tag{SC.2}
$$

(1) We first show that $\beta = 1$ for the linear loss. For $\ell(u) = -u$, by the proof of Remark 3, the minimizer $\boldsymbol{f}^* = \operatorname{arginf}_{\boldsymbol{f}: E(\|\boldsymbol{f}\|^2) \leq 1} R_V(\boldsymbol{f}) = \boldsymbol{V}_0 / [E(\|\boldsymbol{V}_0\|^2)]^{1/2}$.

Then

$$\begin{aligned} e_V(\boldsymbol{f},\boldsymbol{f}^*) =& E[e_{V|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{f},\boldsymbol{f}^*)] = E[\boldsymbol{V}_0^\top(\boldsymbol{f}^*-\boldsymbol{f})] \\ =& E\{[E(\|\boldsymbol{V}_0\|^2)]^{1/2}(\boldsymbol{f}^*)^\top(\boldsymbol{f}^*-\boldsymbol{f})\} \\ \overset{(i)}{\geq}& 2^{-1}[E(\|\boldsymbol{V}_0\|^2)]^{1/2}E[(\boldsymbol{f}^*-\boldsymbol{f})^\top(\boldsymbol{f}^*-\boldsymbol{f})] = 2^{-1}\kappa_1 E[\|\boldsymbol{f}^*-\boldsymbol{f}\|^2], \end{aligned}$$

where (i) is derived from $E(\|\boldsymbol{f}\|^2) \leq E(\|\boldsymbol{f}^*\|^2) = 1$ and $\kappa_1 = [E(\|\boldsymbol{V}_0\|^2)]^{1/2}$. From (SC.2), one can verify

$$\begin{aligned} & \sup_{\{\boldsymbol{f}\in\mathcal{F}:e_V(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z)-V(\boldsymbol{f}^*,Z)) \\ \leq & \sup_{\{\boldsymbol{f}\in\mathcal{F}:E[\|\boldsymbol{f}-\boldsymbol{f}^*\|^2]\leq 2\kappa_1^{-1}\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z)-V(\boldsymbol{f}^*,Z)) \leq 2\kappa_1^{-1}M_1^2\epsilon. \end{aligned}$$

Thus, we have

$$\sup_{\{\boldsymbol{f}\in\mathcal{F}:e_{V^{\widetilde{T}}}(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z)-V(\boldsymbol{f}^*,Z)) \leq c_2\epsilon^\beta \leq 2\kappa_1^{-1}M_1^2\epsilon.$$

Then Assumption 4 is satisfied with $\beta = 1$.

(2) For a loss function $\ell(u)$, which is twice continuously differentiable with $\ell''(u) > 0$,

$$\begin{aligned} & e_{V|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{f},\boldsymbol{f}^*) \\ =& E\left[\sum_{m=2}^{\mathcal{L}(A)}\sum_{\tilde{A}\in[\mathscr{E}_m(A)]}\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{x})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top\boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}\right] - \\ & E\left[\sum_{m=2}^{\mathcal{L}(A)}\sum_{\tilde{A}\in[\mathscr{E}_m(A)]}\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{x})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top\boldsymbol{f}^*(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}\right] \\ =& \sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}[R^+(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}(\boldsymbol{x}))- \\ & R^-(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}(\boldsymbol{x})) - R^+(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}^*(\boldsymbol{x}))+ \\ & R^-(\boldsymbol{x},a)\ell((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}^*(\boldsymbol{x}))] \\ \geq& \sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}[R^+(\boldsymbol{x},a)\{\ell'((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}^*)(\boldsymbol{\eta}_m(a)- \\ & \boldsymbol{\eta}_m(\tilde{a}))^\top(\boldsymbol{f}-\boldsymbol{f}^*) + \ell''(\kappa_2(\boldsymbol{x}))((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top(\boldsymbol{f}-\boldsymbol{f}^*))^2\}- \\ & R^-(\boldsymbol{x},a)\{\ell'((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}^*)(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top(\boldsymbol{f}-\boldsymbol{f}^*)+ \\ & \ell''(\kappa_3(\boldsymbol{x}))((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top(\boldsymbol{f}-\boldsymbol{f}^*))^2\}], \end{aligned}$$

where $\kappa_2(\boldsymbol{x})$ is bounded by $(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}$ and $(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}^*$, and $\kappa_3(\boldsymbol{x})$ is bounded by $(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}$ and $(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}^*$.

Now, let us consider the first order optimality. It holds that the partial derivative of

$$E\left[\sum_{m=2}^{\mathcal{L}(A)}\sum_{\tilde{A}\in[\mathscr{E}_m(A)]}(|R|/\Pr(A^{(m)}|\boldsymbol{x}))\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top\boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}\right]$$

with respect to $\boldsymbol{f}$ is a zero vector of length $K-1$ at $\boldsymbol{f}^*$, because we assume ℓ is differentiable and $\boldsymbol{f}^*$ is the minimizer. Thus, we have

$$\sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}\{R^+(\boldsymbol{x},a)\ell'((\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{f}^*)(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))-$$
$$R^-(\boldsymbol{x},a)\ell'((\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\boldsymbol{f}^*)(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))\}=\boldsymbol{0}_{K-1}.$$

Then

$$\begin{aligned}&e_{V|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{f},\boldsymbol{f}^*)\\
\geq&\sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})}{\Pr(a^{(m)}|\boldsymbol{x})}[R^+(\boldsymbol{x},a)\ell''(\kappa_2(\boldsymbol{x}))\{(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top(\boldsymbol{f}-\boldsymbol{f}^*)\}^2-\\
&R^-(\boldsymbol{x},a)\ell''(\kappa_3(\boldsymbol{x}))\{(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top(\boldsymbol{f}-\boldsymbol{f}^*)\}^2]\\
=&(\boldsymbol{f}-\boldsymbol{f}^*)^\top\left[\sum_{a\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(a)}\sum_{\tilde{a}\in[\mathscr{E}_m(a)]}\frac{\Pr(a|\boldsymbol{x})(R^+(\boldsymbol{x},a)\ell''(\kappa_2(\boldsymbol{x}))-R^-(\boldsymbol{x},a)\ell''(\kappa_3(\boldsymbol{x})))}{\Pr(a^{(m)}|\boldsymbol{x})}\right.\\
&\left.(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))(\boldsymbol{\eta}_m(\tilde{a})-\boldsymbol{\eta}_m(a))^\top\right](\boldsymbol{f}-\boldsymbol{f}^*)\\
=&(\boldsymbol{f}-\boldsymbol{f}^*)^\top\boldsymbol{\Sigma}(\boldsymbol{f}-\boldsymbol{f}^*),\end{aligned}$$

where $\boldsymbol{\Sigma}$ depending on $\boldsymbol{x}$ is defined accordingly. Taking expectations on both sides leads to

$$e_V(\boldsymbol{f},\boldsymbol{f}^*)=E[e_{V|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{f},\boldsymbol{f}^*)]=E[(\boldsymbol{f}-\boldsymbol{f}^*)^\top\boldsymbol{\Sigma}(\boldsymbol{f}-\boldsymbol{f}^*)].$$

We now prove $\boldsymbol{\Sigma}$ is positive definite. Note that for any $\boldsymbol{\eta}_m(a)$ and its sibling $\boldsymbol{\eta}_m(\tilde{a})$, $\{(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a})),\tilde{a}\in[\mathscr{E}_m(a)]\}$ are linearly independent in the corresponding subspace. Thus, for all nodes $\boldsymbol{\eta}_m(a)$ and its siblings $\boldsymbol{\eta}_m(\tilde{a})$, the following equations

$$\boldsymbol{\zeta}^\top(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))=0,\tilde{a}\in[\mathscr{E}_m(a)],m=2,\dots,\mathcal{L}(a),a\in\mathcal{A},$$

have exactly one solution $\boldsymbol{\zeta}=\boldsymbol{0}_{K-1}$. Then for any $\boldsymbol{\zeta}\in\mathbb{R}^K\backslash\{\boldsymbol{0}_{K-1}\}$, there exists at least one node $\boldsymbol{\eta}_m(a)$ and one of its siblings $\boldsymbol{\eta}_m(\tilde{a})$ such that

$$\boldsymbol{\zeta}^\top(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))(\boldsymbol{\eta}_m(a)-\boldsymbol{\eta}_m(\tilde{a}))^\top\boldsymbol{\zeta}>0.$$

By Assumption 6, one can verify

$$\Pr(a|\boldsymbol{x})(R^{+}(\boldsymbol{x},a)\ell''(\kappa_2(\boldsymbol{x})) - R^{-}(\boldsymbol{x},a)\ell''(\kappa_3(\boldsymbol{x})))/\Pr(a^{(m)}|\boldsymbol{x})$$

is finite and positive. Thus, we have $\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta} > 0$ and $\boldsymbol{\Sigma}$ is positive definite. Then there exists some $\kappa_4 > 0$ [34] such that

$$e_V(\boldsymbol{f},\boldsymbol{f}^*) \geq \kappa_4 E[\|\boldsymbol{f}-\boldsymbol{f}^*\|^2].$$

From (SC.2), one can verify

$$\begin{aligned} &\sup_{\{\boldsymbol{f}\in\mathcal{F}:e_V(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}^*,Z)) \\ \leq &\sup_{\{\boldsymbol{f}\in\mathcal{F}:E[\|\boldsymbol{f}-\boldsymbol{f}^*\|^2]\leq\kappa_4^{-1}\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}^*,Z)) \leq \kappa_4^{-1}M_1^2\epsilon. \end{aligned}$$

Thus, we have

$$\sup_{\{\boldsymbol{f}\in\mathcal{F}:e_{V^{\widetilde{T}}}(\boldsymbol{f},\boldsymbol{f}^*)\leq\epsilon\}} \mathrm{Var}(V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}^*,Z)) \leq c_2\epsilon^{\beta} \leq \kappa_4^{-1}M_1^2\epsilon.$$

Therefore, Assumption 4 is satisfied with $\beta = 1$. This completes the proof. □

To verify Assumption 5, the following lemma gives the explicit expression of $\mathcal{H}_B(\epsilon,\mathcal{F}^V(t))$ for any $\epsilon > 0$.

Lemma S1. *Under Assumption 6, it holds that*

$$\mathcal{H}_B(\epsilon,\mathcal{F}^V(t)) \leq c_8\mathcal{H}_B(\epsilon/(c_7c_8),\widetilde{\mathcal{F}}(1)),$$

where $c_7 = M_A^{-1}M_R[4\tilde{\gamma}L^{(m)}(J_0t)^{1/2} + |\ell(0)|]$, $c_8 = \max_{A\in\mathcal{A}}\sum_{m=2}^{\mathcal{L}(A)}\left|Sib(A^{(m)})\right|$ *and* $\widetilde{\mathcal{F}}(1) = \{f : f \in \mathbb{R}, |f| \leq 1\}$.

Proof of Lemma S1. Note that

$$\mathcal{F}^V(t) = \{V^{\widetilde{T}}(\boldsymbol{f},Z) - V(\boldsymbol{f}_0,Z) : \boldsymbol{f}\in\mathcal{F}, J(\boldsymbol{f}) \leq J_0t\},$$

where $J_0 = \max\{J(\boldsymbol{f}_0),1\}$ and

$$V^{\widetilde{T}}(\boldsymbol{f},Z) = \widetilde{T} \wedge \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top\boldsymbol{f}(\boldsymbol{X})).$$

Note that $J(\boldsymbol{f}) \leq J_0t$ implies that $\|\boldsymbol{f}\| \leq (J_0t)^{1/2}$ for linear learning and for nonlinear learning with the Gaussian kernel.

According to our embedding algorithm, we have

$$\left|\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top\boldsymbol{f}(\boldsymbol{X}))\right|$$

$$
\begin{aligned}
&= \frac{1}{\Pr(A^{(m)}|\boldsymbol{X})}\left|R^{+}\ell((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X}))-R^{-}\ell((\boldsymbol{\xi}_m(\tilde{A})-\boldsymbol{\xi}_m(A))^\top \boldsymbol{f}(\boldsymbol{X}))\right| \\
&= \frac{1}{\Pr(A^{(m)}|\boldsymbol{X})}\Big|R^{+}[\ell((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X}))-\ell(0)]- \\
&\quad R^{-}[\ell((\boldsymbol{\xi}_m(\tilde{A})-\boldsymbol{\xi}_m(A))^\top \boldsymbol{f}(\boldsymbol{X}))-\ell(0)]+R^{+}\ell(0)-R^{-}\ell(0)\Big| \\
&\le \frac{1}{\Pr(A^{(m)}|\boldsymbol{X})}[R^{+}\tilde{\gamma}|(\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})|- \\
&\quad R^{-}\tilde{\gamma}|(\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})|+|R\ell(0)|] \\
&\le M_A^{-1}[4M_R\tilde{\gamma}L^{(m)}(J_0t)^{1/2}+M_R|\ell(0)|] \\
&= M_A^{-1}M_R[4\tilde{\gamma}L^{(m)}(J_0t)^{1/2}+|\ell(0)|].
\end{aligned}
$$

Define $c_7 = M_A^{-1}M_R[4\tilde{\gamma}L^{(m)}(J_0t)^{1/2} + |\ell(0)|]$. Then it holds that

$$
\begin{aligned}
&\left\{\frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})) : J(\boldsymbol{f}) \le J_0t\right\} \\
&\subseteq \{f : f \in \mathbb{R}, |f| \le c_7\} := \widetilde{\mathcal{F}}(c_7).
\end{aligned}
$$

For any set $\mathcal{F}_1$ of functions, define $\widetilde{T} \wedge \mathcal{F}_1 = \{\widetilde{T} \wedge f : f \in \mathcal{F}_1\}$. For any two sets of functions $\mathcal{F}_1$ and $\mathcal{F}_2$, denote $\mathcal{F}_1 \oplus \mathcal{F}_2 = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. Then one can see that

$$
\begin{aligned}
&\left\{V^{\widetilde{T}}(\boldsymbol{f}, Z) : J(\boldsymbol{f}) \le J_0t\right\} \\
&= \left\{\widetilde{T} \wedge \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \frac{|R|}{\Pr(A^{(m)}|\boldsymbol{X})}\ell_R((\boldsymbol{\xi}_m(A)-\boldsymbol{\xi}_m(\tilde{A}))^\top \boldsymbol{f}(\boldsymbol{X})) : J(\boldsymbol{f}) \le J_0t\right\} \\
&\subseteq \bigcup_{A\in\mathcal{A}} \bigoplus_{m=2}^{\mathcal{L}(A)} \bigoplus_{\tilde{A}\in[\mathscr{E}_m(A)]} \left\{\widetilde{T} \wedge \widetilde{\mathcal{F}}(c_7)\right\} := \mathcal{B}.
\end{aligned} \tag{SC.3}
$$

First, we construct an ϵ-bracketing set $\Pi = \{(f_i^l, f_i^u)\}$ for $\widetilde{\mathcal{F}}(1)$, that is, $E[|f_i^l - f_i^u|^2]^{1/2} \le \epsilon$. For any constant $\kappa > 0$, define $\kappa\Pi = \{(\kappa f_i^l, \kappa f_i^u)\}$. One can see that $\kappa\Pi$ is a $\kappa\epsilon$-bracketing set of $\widetilde{\mathcal{F}}(\kappa)$. Define $\widetilde{T}\wedge\Pi = \{(\widetilde{T}\wedge f_i^l, \widetilde{T}\wedge f_i^u)\}$. For any $\Pi_1 = \{(f_i^{1,l}, f_i^{1,u})\}$ and $\Pi_2 = \{(f_i^{2,l}, f_i^{2,u})\}$, we define $\Pi_1 \oplus \Pi_2 = \{(f_i^{1,l} + f_j^{2,l}, f_i^{1,u} + f_j^{2,u})\}$. Then one can verify that

$$
\Xi = \bigcup_{A\in\mathcal{A}} \bigoplus_{m=2}^{\mathcal{L}(A)} \bigoplus_{\tilde{A}\in[\mathscr{E}_m(A)]} \widetilde{T} \wedge (c_7\Pi)
$$

is a $c_7c_8\epsilon$-bracketing set of $\mathcal{B}$ with $c_8 = \max_{A\in\mathcal{A}} \sum_{m=2}^{\mathcal{L}(A)} \left|\mathrm{Sib}(A^{(m)})\right|$. In fact, any function g in $\mathcal{B}$ can be written as

$$g = \widetilde{T} \wedge \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \tilde{f}_{m,\tilde{A}},$$

for some $A \in \mathcal{A}$ and $\tilde{f}_{m,\tilde{A}} \in \widetilde{\mathcal{F}}(c_7)$. Let

$$g^u = \widetilde{T} \wedge \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \tilde{f}^u_{m,\tilde{A}}, \quad g^l = \widetilde{T} \wedge \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} \tilde{f}^l_{m,\tilde{A}},$$

where $(\tilde{f}^l_{m,\tilde{A}}, \tilde{f}^u_{m,\tilde{A}}) \in c_7\Pi$ such that $\tilde{f}^l_{m,\tilde{A}} \leq \tilde{f}_{m,\tilde{A}} \leq \tilde{f}^u_{m,\tilde{A}}$. One can verify that $g^l \leq g \leq g^u$ and that

$$|g^u - g^l| \leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} |\tilde{f}^l_{m,\tilde{A}} - \tilde{f}^u_{m,\tilde{A}}|.$$

Thus, we have

$$E\left[|g^u - g^l|^2\right] \leq \sum_{m=2}^{\mathcal{L}(A)} \sum_{\tilde{A}\in[\mathscr{E}_m(A)]} E\left[|\tilde{f}^l_{m,\tilde{A}} - \tilde{f}^u_{m,\tilde{A}}|^2\right] \leq c_7 c_8 \epsilon,$$

where $c_8 = \max_{A\in\mathcal{A}} \sum_{m=2}^{\mathcal{L}(A)} \left|\mathrm{Sib}(A^{(m)})\right|$. Therefore, Ξ is a $c_7c_8\epsilon$-bracketing set of $\mathcal{B}$. The cardinality $|\Xi| \leq |\Pi|^{c_8}$, that is

$$\mathcal{H}_B(c_7c_8\epsilon, \mathcal{B}) \leq c_8 \mathcal{H}_B(\epsilon, \widetilde{\mathcal{F}}(1)).$$

Thus, we have

$$\mathcal{H}_B(\epsilon, \mathcal{F}^V(t)) \leq \mathcal{H}_B(\epsilon, \mathcal{B}) \leq c_8 \mathcal{H}_B(\epsilon/(c_7c_8), \widetilde{\mathcal{F}}(1)).$$

This completes the proof. □

To compute α, we consider a specific hierarchical structure shown as FIG S1, which is a binary tree of 3 layers. The corresponding embedded points are shown as follows,

$$\left(\begin{array}{cccccc} \boldsymbol{\xi}_{1,1} & \boldsymbol{\xi}_{1,2} & \boldsymbol{\xi}_{1,1,1} & \boldsymbol{\xi}_{1,1,2} & \boldsymbol{\xi}_{1,2,1} & \boldsymbol{\xi}_{1,2,2} \\ -1 & 1 & -1 & -1 & 1 & 1 \\ 0 & 0 & -\sqrt{5}/5 & \sqrt{5}/5 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\sqrt{5}/5 & \sqrt{5}/5 \end{array}\right).$$

Let $\boldsymbol{X} = (X^{(1)}, X^{(2)}, X^{(3)})^\top$, where $X^{(1)}, X^{(2)}, X^{(3)}$ are predictors. Assume the marginal distribution of $\boldsymbol{X}$ is non-zero only on the embedded points $\boldsymbol{\xi}_{1,1,1}$, $\boldsymbol{\xi}_{1,1,2}$, $\boldsymbol{\xi}_{1,2,1}, \boldsymbol{\xi}_{1,2,2}$ such that

$$P(\boldsymbol{X} = \boldsymbol{\xi}_{1,1,1}) = P(\boldsymbol{X} = \boldsymbol{\xi}_{1,1,2}) = P(\boldsymbol{X} = \boldsymbol{\xi}_{1,2,1}) = P(\boldsymbol{X} = \boldsymbol{\xi}_{1,2,2}) = 1/4.$$

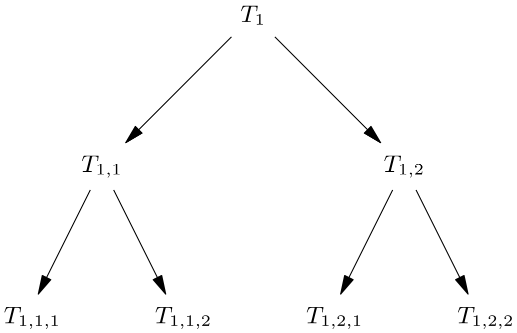

FIG S1. *The hierarchical structure for the illustrative example.*

Consider the random trials, where

$$\begin{aligned}&\Pr(A=\{T_1,T_{1,1},T_{1,1,1}\}|\boldsymbol{X})=\Pr(A=\{T_1,T_{1,1},T_{1,1,2}\}|\boldsymbol{X})\\=&\Pr(A=\{T_1,T_{1,2},T_{1,2,1}\}|\boldsymbol{X})=\Pr(A=\{T_1,T_{1,2},T_{1,2,2}\}|\boldsymbol{X})=1/4.\end{aligned}$$

The conditional distribution of R is given by

$$\begin{aligned}&P(R\le r|A=\{T_1,T_{1,1},T_{1,1,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,1})=r/2,0\le r\le 2\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,1,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,2})=P(R\le r|A=\{T_1,T_{1,1},T_{1,1,1}\},\\&\boldsymbol{X}=\boldsymbol{\xi}_{1,2,1})=P(R\le r|A=\{T_1,T_{1,1},T_{1,1,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,2})=r,0\le r\le 1\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,1,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,2})=r/2,0\le r\le 2\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,1,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,1})=P(R\le r|A=\{T_1,T_{1,1},T_{1,1,2}\},\\&\boldsymbol{X}=\boldsymbol{\xi}_{1,2,1})=P(R\le r|A=\{T_1,T_{1,1},T_{1,1,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,2})=r,0\le r\le 1\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,2,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,1})=r/2,0\le r\le 2\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,2,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,1})=P(R\le r|A=\{T_1,T_{1,1},T_{1,2,1}\},\\&\boldsymbol{X}=\boldsymbol{\xi}_{1,1,2})=P(R\le r|A=\{T_1,T_{1,1},T_{1,2,1}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,2})=r,0\le r\le 1\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,2,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,2})=r/2,0\le r\le 2\\&P(R\le r|A=\{T_1,T_{1,1},T_{1,2,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,1,1})=P(R\le r|A=\{T_1,T_{1,1},T_{1,2,2}\},\\&\boldsymbol{X}=\boldsymbol{\xi}_{1,1,2})=P(R\le r|A=\{T_1,T_{1,1},T_{1,2,2}\},\boldsymbol{X}=\boldsymbol{\xi}_{1,2,1})=r,0\le r\le 1.\end{aligned}$$

We consider linear learning for this example. One can verify the optimal minimizer is $\boldsymbol{f}^*=(f_1^*,f_2^*,f_3^*)^\top$ with $f_j^*=X^{(j)}$ for $j=1,2,3$. Let $\boldsymbol{f}=((1+\Delta_1)X^{(1)}+\Delta_2,(1+\Delta_3)X^{(2)}+\Delta_4,(1+\Delta_5)X^{(3)}+\Delta_6)^\top$. For $|\Delta_i|<(\sqrt{5}-1)/4, i=1,\ldots,6$, we have $e(\boldsymbol{f},\boldsymbol{f}^*)=0$. Then it holds that

$$|e(\boldsymbol{f},\boldsymbol{f}^*)|=|\mathcal{R}(\boldsymbol{f})-\mathcal{R}(\boldsymbol{f}^*)|\le M_R M_A^{-1}(\sqrt{5}+1)\sup_{i=1,\ldots,6}(|\Delta_i|).$$

Note that $E[\|\boldsymbol{f}-\boldsymbol{f}^*\|^2]\ge(2/5)\sum_{i=1}^6\Delta_i^2\ge((3-\sqrt{5})M_R^{-2}M_A^2/20)|e(\boldsymbol{f},\boldsymbol{f}^*)|^2$. By the proof of Proposition S2, $e_V(\boldsymbol{f},\boldsymbol{f}^*)\ge(2\kappa_1)^{-1}E[\|\boldsymbol{f}-\boldsymbol{f}^*\|^2]$ for the linear loss and $e_V(\boldsymbol{f}, \boldsymbol{f}^*) \geq \kappa_4 E[\|\boldsymbol{f} - \boldsymbol{f}^*\|^2]$ for the loss functions that are twice continuously differentiable with $\ell''(u) > 0$. Thus, under the conditions of Proposition S2, there exists some constant c_1 such that $|e(\boldsymbol{f}, \boldsymbol{f}^*)|^2 \leq (c_1)^2 e_V(\boldsymbol{f}, \boldsymbol{f}^*)$. Then

$$\sup_{\{\boldsymbol{f} \in \mathcal{F} : e_{V^{\widetilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \leq \epsilon\}} |e(\boldsymbol{f}, \boldsymbol{f}^*)| \leq c_1 \epsilon^{1/2}.$$

Therefore, Assumption 3 is satisfied with $\alpha = 1/2$.

Combining with $\beta = 1$ from Proposition S2, we then verify Assumption 5. From Wang, Shen and Pan [31], $\mathcal{H}_B(\epsilon, \widetilde{\mathcal{F}}(1)) \leq O(\log(1/\epsilon))$. Thus, by Lemma S1, $\mathcal{H}_B(\epsilon, \mathcal{F}^V(t)) \leq O(c_8 \log(c_7 c_8/\epsilon))$. By the definitions of $\phi(\varepsilon_n, t)$ and $\tilde{L}$ in Assumption 5, it follows that $\sup_{t \geq 1} \phi(\varepsilon_n, t) \leq O((c_8 \log(c_7 c_8/\varepsilon_n))^{1/2}/\varepsilon_n)$, and consequently that $\varepsilon_n = (c_8 n^{-1} \log n)^{1/2}$. Moreover, for this example, one can compute that $c_8 = k - 1$, where k is the number of layers. By Theorem 2 and Corollary 1, we have $|e(\hat{\boldsymbol{f}}_\lambda, \boldsymbol{f}^*)| = O_p(\varepsilon_n) = O_p((kn^{-1} \log n)^{1/2})$.

C.6. Proof of Theorem 2

Proof of Theorem 2. Under Assumptions 3–5, the proof is immediate from Theorem 1 in Shen and Wang [26].

C.7. Proof of Theorem 3.

Proof of Theorem 3. To bound $P(|e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)| \geq c_1 \delta_n^{2\alpha})$, we first establish a connection between $e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)$ and $e_{V^{\widetilde{T}}}(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)$. Under Assumptions 6 and 7, we have

$$n^{-1} \sum_{i=1}^n V^{\hat{\pi}}(\boldsymbol{f}, z_i) = n^{-1} \sum_{i=1}^n V(\boldsymbol{f}, z_i) + O_p(s_n).$$

By the definition of $\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}_0$ and Assumption 3, it holds that

$$\begin{aligned}
&\left\{|e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)| \geq c_1 \delta_n^{2\alpha}\right\} \subset \left\{e_{V^{\widetilde{T}}}(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*) \geq \delta_n^2\right\} \\
&\subset \left\{\sup_{\{\boldsymbol{f} \in \mathcal{F} : e_{V^{\widetilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \geq \delta_n^2\}} n^{-1} \sum_{i=1}^n [V^{\hat{\pi}}(\boldsymbol{f}_0, z_i) - V^{\hat{\pi}}(\boldsymbol{f}, z_i) + \lambda(J(\boldsymbol{f}_0) - J(\boldsymbol{f}))] \geq 0\right\} \\
&\subset \left\{\sup_{\{\boldsymbol{f} \in \mathcal{F} : e_{V^{\widetilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \geq \delta_n^2\}} n^{-1} \sum_{i=1}^n [V(\boldsymbol{f}_0, z_i) - V(\boldsymbol{f}, z_i) + \lambda(J(\boldsymbol{f}_0) - J(\boldsymbol{f}))] + O_p(s_n) \geq 0\right\} \\
&\subset \left\{\sup_{\{\boldsymbol{f} \in \mathcal{F} : e_{V^{\widetilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \geq \delta_n^2\}} n^{-1} \sum_{i=1}^n [V(\boldsymbol{f}_0, z_i) - V^{\widetilde{T}}(\boldsymbol{f}, z_i) + \lambda(J(\boldsymbol{f}_0) - J(\boldsymbol{f}))] + O_p(s_n) \geq 0\right\}.
\end{aligned}$$

Let $D_n = n^{-1} \sum_{i=1}^n [V(\boldsymbol{f}_0, z_i) - V^{\widetilde{T}}(\boldsymbol{f}, z_i) + \lambda(J(\boldsymbol{f}_0) - J(\boldsymbol{f}))]$ and

$$I \equiv P^* \left(\sup_{\{\boldsymbol{f} \in \mathcal{F} : e_{V^{\widetilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \geq \delta_n^2\}} D_n \geq -\delta_n^2/4 \right),$$

where P^* is the outer probability. Let $U_{ij} = \{\boldsymbol{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^{\tilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \leq 2^i\delta_n^2, 2^{j-1}J_0 \leq J(\boldsymbol{f}) < 2^j J_0\}$ and $U_{i0} = \{\boldsymbol{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^{\tilde{T}}}(\boldsymbol{f}, \boldsymbol{f}^*) \leq 2^i\delta_n^2, J(\boldsymbol{f}) < J_0\}$ for $i, j = 1, 2, \ldots$. We then consider a sequence of empirical processes on $\{U_{ij}, i = 1, 2, \ldots, j = 0, 1, \ldots\}$. Let $I = I_1 + I_2$, where

$$I_1 = \sum_{i,j\geq 1} P^*\left(\sup_{\boldsymbol{f}\in U_{ij}} D_n \geq -\delta_n^2/4\right), \quad I_2 = \sum_{i=1}^{\infty} P^*\left(\sup_{\boldsymbol{f}\in U_{i0}} D_n \geq -\delta_n^2/4\right).$$

Let $M(i,j) = 2^{i-1}\delta_n^2 + \lambda 2^{j-1}J_0$. Based on the fact that $e_V(\boldsymbol{f}_0, \boldsymbol{f}^*) < \delta_n^2/2$, $\lambda J_0 \leq \delta_n^2/2$, $|V(\boldsymbol{f}_0, Z) - V^{\widetilde{T}}(\boldsymbol{f}, Z)| \leq 2\widetilde{T}$ and Assumptions 4–5, by Theorem 3 of Shen and Wong [27] with $M = 3n^{1/2}M(i,j)/4, v = 4c_2M(i,j)^\beta, \epsilon = 9/10$, we have that for $i, j = 1, 2, \ldots,$

$$\begin{aligned}
&P^*\left(\sup_{\boldsymbol{f}\in U_{ij}} D_n \geq -\delta_n^2/4\right) \leq P^*\left(\sup_{\boldsymbol{f}\in U_{ij}} D_n \geq -M(i,j)/4\right) \\
\leq& P^*\left(\sup_{\boldsymbol{f}\in U_{ij}} D_n - E\{(V(\boldsymbol{f}_0, Z) - V^{\widetilde{T}}(\boldsymbol{f}, Z)) + \lambda(J(\boldsymbol{f}_0) - J(\boldsymbol{f}))\} \geq 3M(i,j)/4\right) \\
\leq& 3\exp\left\{-\frac{(1-\epsilon)M^2}{2(4v + M\widetilde{T}n^{-1/2}/3)}\right\} = 3\exp\left\{-\frac{9nM(i,j)^2/160}{2(16c_2M(i,j)^\beta + M(i,j)\widetilde{T}/4)}\right\}.
\end{aligned}$$

Thus, it holds that

$$\begin{aligned}
I_1 \leq& \sum_{i,j\geq 1} 3\exp\left\{-\frac{9nM(i,j)^2/160}{2(16c_2M(i,j)^\beta + M(i,j)\widetilde{T}/4)}\right\} \\
\leq& \sum_{i,j\geq 1} 3\exp\left\{-c_6 n M(i,j)^{2-\min(\beta,1)}\right\} \\
\leq& \sum_{i,j\geq 1} 3\exp\left\{-c_6 n (2^{i-1}\delta_n^2 + (2^{j-1}-1)\lambda J_0)^{2-\min(\beta,1)}\right\} \\
\leq& 3\exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)})/\{1 - \exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)})\}^2,
\end{aligned}$$

where c_6 is a positive generic constant. Similarly, I_2 can be bounded. Then we have

$$I \leq I_1 + I_2 \leq 6\exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)})/\{1 - \exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)})\}^2,$$

and consequently,

$$I \leq I^{1/2} \leq 3.5\exp(-c_6 n(\lambda J_0)^{2-\min(\beta,1)}).$$

Let $B_n = O_p(s_n)$, where $\delta_n^{-2}s_n = o(1)$. When $n(\lambda J_0)^{2-\min\{\beta,1\}}$ is bounded away from 0 as $n \to \infty$, for any $\tilde{\epsilon} > 0$, there exists N_e such that for any $n > N_e$,

$$P(|e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)| \geq c_1\delta_n^{2\alpha}) \leq I + P\left(|B_n| > \delta_n^2/4\right) < \tilde{\epsilon},$$

which implies $|e(\hat{\boldsymbol{f}}_\lambda^{\hat{\pi}}, \boldsymbol{f}^*)| = O_p(\delta_n^{2\alpha})$. This completes the proof. □

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and reviewers, whose helpful comments and suggestions led to a much improved presentation.

References

[1] Aziz, M. S., Kahle, M., Meier, J. J. and Nauck, M. A. (2017). A meta-analysis comparing clinical effects of short-or long-acting GLP-1 receptor agonists versus insulin treatment from head-to-head studies in type 2 diabetic patients. *Diabetes, Obesity and Metabolism* **19** 216–227.

[2] Bartlett, P. L., Jordan, M. I. and Mcauliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101** 138-156. MR2268032

[3] Cai, L. and Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM2004)* 78–87. ACM Press.

[4] Cesa-Bianchi, N., Gentile, C. and Zaniboni, L. (2006). Hierarchical classification: combining Bayes with SVM. In *Proceedings of the 23rd international conference on Machine learning (ICML2006)* 177–184. ACM Press.

[5] Chen, J., Fu, H., He, X., Kosorok, M. R. and Liu, Y. (2018). Estimating individualized treatment rules for ordinal treatments. *Biometrics* **74** 924–933. MR3860713

[6] Davies, M. N., Secker, A., Freitas, A. A., Mendao, M., Timmis, J. and Flower, D. R. (2007). On the hierarchical classification of G protein-coupled receptors. *Bioinformatics* **23** 3113–3118.

[7] Fan, R., Chang, K., Hsieh, C., Wang, X. and Lin, C. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* **9** 1871–1874.

[8] Fan, J., Zhang, J., Mei, K., Peng, J. and Gao, L. (2015). Cost-sensitive learning of hierarchical tree classifiers for large-scale image classification and novel category detection. *Pattern Recognition* **48** 1673-1687.

[9] Fan, Y., Lu, X., Liu, Y. and Zhao, J. (2020). Angle-based hierarchical classification using exact label embedding. *Journal of the American Statistical Association.* In press.

[10] Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., Van Staa, T. and Smeeth, L. (2015). Data resource profile: clinical practice research datalink (CPRD). *International Journal of Epidemiology* **44** 827–836.

[11] Holman, R. R., Thorne, K. I., Farmer, A. J., Davies, M. J., Keenan, J. F., Paul, S. and Levy, J. C. (2007). Addition of biphasic, prandial, or basal insulin to oral therapy in type 2 diabetes. *New England Journal of Medicine* **357** 1716–1730.

[12] Kasi, P. M., Ansell, S. M. and Gertz, M. A. (2015). Waldenström macroglobulinemia. *Clinical Advances in Hematology and Oncology* **13** 56–66.

[13] Lange, K. and Wu, T. T. (2008). An MM algorithm for multicategory vertex discriminant analysis. *Journal of Computational and Graphical Statistics* **17** 527-544. MR2528236

[14] Liu, Y., Zhang, H. H. and Wu, Y. (2011). Hard or soft classification? Large-margin unified machines. *Journal of the American Statistical Association* **106** 166-177. MR2816711

[15] Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y. and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine* **37** 3776–3788. MR3869154

[16] Morel, C. F. and Clarke, J. T. (2009). The use of agalsidase alfa enzyme replacement therapy in the treatment of Fabry disease. *Expert Opinion on Biological Therapy* **9** 631–639.

[17] Mosenzon, O., Pollack, R. and Raz, I. (2016). Treatment of type 2 diabetes: from "guidelines" to "position statements" and back: recommendations of the Israel National Diabetes Council. *Diabetes Care* **39** S146–S153.

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

[19] Murphy, S. A., Der Laan, M. J. V. and Robins, J. M. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96** 1410–1423. MR1946586

[20] Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S. et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304** 1497–1500.

[21] Pelletier, C. (2003). *Lange smart charts: pharmacology.* McGraw Hill Education.

[22] Qi, Z., Liu, D., Fu, H. and Liu, Y. (2020). Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association* **115** 678–691. MR4107672

[23] Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39** 1180-1210. MR2816351

[24] Robins, J., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* **27** 4678–4721. MR2528576

[25] Shao, Y. H., Chen, W. J., Wang, Z., Li, C. N. and Deng, N. Y. (2015). Weighted linear loss twin support vector machine for large-scale classification. *Knowledge-Based Systems* **73** 276-288.

[26] Shen, X. and Wang, L. (2007). Generalization error for multi-class margin classification. *Electronic Journal of Statistics* **1** 307–330. MR2336036

[27] Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* **22** 580–615. MR1292531

[28] Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* **22** 31-72. MR2764552

[29] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* **6** 1453-1484. MR2249862

[30] Wang, J., Shen, X. and Pan, W. (2009). On large margin hierarchical classification with multiple paths. *Journal of the American Statistical Association* **104** 1213-1223. MR2562009

[31] Wang, H., Shen, X. and Pan, W. (2011). Large margin hierarchical classification with mutually exclusive class membership. *Journal of Machine Learning Research* **12** 2721–2748. MR2845677

[32] Wu, Y. and Liu, Y. (2013). Adaptively Weighted Large Margin Classifiers. *Journal of Computational and Graphical Statistics* **22** 416-432. MR3173722

[33] Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* **5** 1225–1251. MR2248016

[34] Zhang, C. and Liu, Y. (2014). Multicategory angle-based large-margin classification. *Biometrika* **101** 625-640. MR3254905

[35] Zhang, C., Liu, Y., Wang, J. and Zhu, H. (2016). Reinforced angle-based multicategory support vector machines. *Journal of Computational and Graphical Statistics* **25** 806–825. MR3533639

[36] Zhang, C., Chen, J., Fu, H., He, X., Zhao, Y.-Q. and Liu, Y. (2020). Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica Sinica* **30** 1857-1879. MR4260747

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

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