Sequential Outcome-Weighted Multicategory Learning for Estimating Optimal Individualized Treatment Rules

Xuan Zhou*[1], Yuanjia Wang†[2] and Donglin Zeng‡[1]

[1]Department of Biotatistics, University of North Carolina at Chapel Hill
[2]Department of Biotatistics, Columbia University

Abstract

Personalized medicine has received increasing interest among clinicians and statisticians. One particular aspect of personalized medicine is to consider an individualized treatment strategy based on an individual's characteristics that leads to the greatest benefit. Recently, powerful machine learning methods have been proposed to estimate optimal individualized treatment rule (ITR) but are mostly restricted to the situation with only two treatments. When many treatment options are considered, which is often the case in practice, existing methods have to transform multicategory treatment selection into multiple binary treatment selections, for example, via one vs one or one vs all comparison. However, combining conclusions from multiple binary treatment selection is not straightforward and it may lead to inconsistent decision rules. In this article, we propose a novel and efficient method to generalize outcome weighted learning (O-learning) to multi-treatment settings. Specifically, we solve a multicategory treatment selection problem via sequential weighted support vector machines. Theoretically, we show that the resulting ITR is Fisher consistent. We demonstrate the performance of the proposed method with extensive simulations. An application to a three-arm randomized trial of treating major depressive disorder (MDD) shows that an individualiZed treatment strategy tailored to individual patients' expectancy of treatment efficacy and their baseline depression severity reduces depressive symptoms more than non-personaliZed treatment strategies.

Keywords: PersonaliZed medicine, Randomized clinical trial, Multicategory classification, Fisher consistency, Weighted support vector machine, Major depression disorder

*xuanz@live.unc.edu
†yw2016@cumc.columbia.edu
‡dzeng@email.unc.edu

1 Introduction

For many chronic diseases such as major depression and type 2 diabetes, treatment heterogeneity has been well documented where a treatment that is effective in the overall population may be highly ineffective in a subgroup of patients with specific characteristics (Trivedi Madhukar et al., 2008), or no longer beneficial after patients develop resistance (Lipska and Krumholz, 2014). On the other hand, in some cases a newly developed intervention may not be more efficacious than existing treatments in the overall population, but may reveal a large effect in subgroups of patients (Carini et al., 2014). Henceforth, there has been a growing interest in understanding treatment heterogeneity and discovering individualiZed treatment rules tailored to patient-specific characteristics so as to achieve personaliZed medicine and maximize efficacy (Kosorok and Moodie, 2015). More specifically, tailored treatment strategy aims to recommend optimal treatment decision for an individual patient based on a combination of his or her characteristics such as genomic features, medical treatment history, preference, and treatment expectancy.

Recently, there has been a surge of statistical methods on estimating optimal treatment regimes involving a single decision point or multiple decision points using data collected from clinical trials or observational studies (Murphy, 2003; Robins, 2004; Moodie et al., 2007; Qian and Murphy, 2011; Zhao et al., 2011, 2012; Zhang et al., 2012, 2013). The most widely used method is regression based Q-learning (Watkins, 1989; Murphy, 2005), which relies on some postulated models to incorporate treatment-by-covariate interactions. Alternatively, Zhao et al. (2012); Zhang et al. (2012, 2013) proposed machine learning algorithms, for instance, outcome weighted learning (O-learning), to choose the treatment rules by directly optimizing the expected clinical outcome, called the value function, and draw connection with a classification problem. Most of these methods aim at estimating optimal treatment rules for each patient between only two treatment options. However, in many clinical applications it is common that there are more than two treatments being compared. In our motivating stud of Research Evaluating the Value of Augmenting Medication with Psychotherapy (REVAMP) trial (Kocsis et al., 2009), non responders or partial responders to a first-line antidepressant were randomized to three second-line treatment strategies.

When it comes to multiple arm trials, Q-learning approach, which relies heavily on the correctness of postulated models, is more prone to model misspecification. For machine learning methods in Zhao et al. (2012), Zhang et al. (2012, 2013), multicategory comparisons may be obtained via one versus

one or one versus all comparisons. However, it is well documented in multicategory learning literature that the resulting classification rules from these methods is inconsistent (Dietterich and Bakiri, 1995; Kreßel, 1999; Allwein et al., 2001; Lee et al., 2004; Liu and Shen, 2006), due to possible conflicted decisions from pairwise or one versus all comparisons. To the best of our knowledge, there has been no work on using multicategory learning to consistently estimate an optimal ITR for multi-arm trials.

In this paper, we propose a new approach to identify optimal individualiZed treatment rules (ITRs) from multiple treatment options. Specifically, we transform the value maximization problem into a sequence of binary weighted classifications, which we name as sequential outcome-weighed multicateogry (SOM) learning. At each step, we use a weighted binary support vector machine (SVM) for determining the optimal treatment for patients into one treatment category versus remaining treatment categories, where weights are proportional to outcome values and reflect the fact that one single treatment category is compared to one or more treatment categories. We first estimate optimal rule for a designated treatment class by excluding the possibility of deciding any other treatment as optimal via sequential SVMs; we then exclude already determined treatments and repeat the same learning approach for the remaining treatment options. Theoretically, we show that the derived treatment rule is Fisher consistent. We demonstrate through extensive simulations that SOM learning has superior performance in comparison to Q-learning. Finally, an application of SOM learning to REVAMP shows that an ITR tailored to individual characteristics such as patients' expectancy of treatment efficacy and baseline depression severity reduces depressive symptoms more than a non-personaliZed treatment strategy.

The rest of the paper is structured as follows. Section 2 introduces the main idea and the mathematical framework for multicategory individualiZed treatment rules and formulate the problem for SOM learning. The detailed algorithm is then provided in the section. In Section 3, we provide theoretical justification for SOM learning. Sections 4 and 5 present extensive simulations and application to REVAMP to examine the performance of SOM learning. Finally, concluding remarks are given in Section 6. The proof of theoretical results are presented in the Appendix.

2 Methodology

2.1 Optimal individualiZed treatment rule with multicategory treatments

Assume data are collected from a randomized trial with n patients and k different treatment options. For each patient i, we observe a d-dimensional vector of feature variables, denoted by $\mathbf{X}_i \in \mathcal{X}$, a treatment assignment $A_i \in \mathcal{A} = \{1, 2, ..., k\}$, $i = 1, ..., n$, and the clinical outcome after treatment denoted by R_i (also referred as the "reward"), with larger values of R_i being more desirable. A multicategory ITR, denoted by $\mathcal{D}$, is a mapping from the space of feature variables, $\mathcal{X}$, to the domain of treatments, $\mathcal{A}$. An optimal ITR is a treatment assignment rule that maximizes the mean reward $E[R(\mathcal{D}(\mathbf{X}))|\mathbf{X}]$, where $R(a)$ is the potential outcome if treatment a is given. According to Qian and Murphy (2011), under a randomization setting and assuming consistency of the potential outcomes, the optimal ITR maximizes the following value function:

$$E\left[\frac{I(A = \mathcal{D}(\boldsymbol{x}))}{\pi_A(\mathbf{X})}R\right], \tag{1}$$

where $\pi_a(\boldsymbol{x}) = P(A = a|\mathbf{X} = \boldsymbol{x})$ is the randomization probability for treatment a, $a = 1, ..., k$, so $\sum_{a=1}^{k} \pi_a(\boldsymbol{x}) = 1$. The goal is to learn the optimal treatment rule using empirical observations $(R_i, A_i, \mathbf{X}_i), i = 1, ..., n$.

Theoretically, it can be easily shown that the optimal ITR is given as

$$\mathcal{D}^*(\boldsymbol{x}) = \text{argmax}_a E[R|A = a, \mathbf{X} = \boldsymbol{x}].$$

Therefore, one approach to estimate the optimal ITR is using a regression model to estimate the conditional means in the right-hand side. However, this approach heavily relies on the correctness of the postulated model, and model misspecification can lead to substantially non-optimal ITR even for a binary treatment situation (Zhao et al., 2012). Alternatively, (**?**) directly maximized the empirical version of the value function but replaced $I(A = \mathcal{D}(\boldsymbol{x}))$ by $1 - \max(0, 1 - Af(\boldsymbol{x}))$, where $f(\boldsymbol{x})$ is the decision function such that $\mathcal{D}(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x})$. The latter corresponds to a weighted support vector machine where the weight for each observation is proportional to R_i. Because of this, they called their method outcome weighted learning (abbreviated as O-learning). They demonstrated that O-learning outperformed the regression model based method. However, the proposed method can only be applied

to a binary treatment, not for more than 2 treatment options. Therefore, we aim to develop a robust method based on machine learning, which builds on O-learning for binary decision rule to learn optimal multicategory treatment decision rules.

2.2 Main idea

The main idea of our method, named SOM learning, is to perform a sequence of binary treatment rule learning, where each step in the sequence decides whether the optimal treatment for a patient should be one candidate treatment category or the others. To illustrate this idea, we order the candidate treatment categories based on the descending order of their prevalence so without loss of generality, we assume that the order of the labels of treatments are $k, k-1, ..., 1$.

We first aim to learn an optimal treatment rule that will decide whether one subject should be optimally treated with treatment k. Equivalently, we partition the domain of $\mathbf{X}$ into $\mathcal{X}_k$ and $\mathcal{X}_k^c$ such that for subject with feature values $\mathbf{X}$ in $\mathcal{X}_k$, the optimal treatment is k; for subject with $\mathbf{X}$ in $\mathcal{X}_k^c$, the optimal treatment should not be k. To this end, we consider an ordered sequence of $\{1, 2, ..., k-1\}$, denoted by $\{j_1, ..., j_{k-1}\}$, and let $j_k = k$. A sequential ITR learning is then conducted in the following.

In the first step, starting with j_1 versus $\{j_2, ..., j_k\}$, we determine whether a subject should be treated optimally with treatment j_1 or not. Since this is only a binary decision problem, we can use existing methods for learning a binary treatment decision rule, for example, O-learning, with additional modifications as explained in later section. With this binary rule, for a future patient with $\mathbf{X}$, if he or she is assigned to treatment j_1, then clearly, $\mathbf{X} \in \mathcal{X}_k^c$. Otherwise, we cannot determine whether $\mathbf{X}$ should be in $\mathcal{X}_k$ or $\mathcal{X}_k^c$ since his/her optimal treatment can be one of $j_2, ..., j_k$.

In the second step of this sequential learning, we only consider patients whose optimal treatments are not determined as j_1 in the previous step. We then aim to learn a binary treatment rule to decide whether this subject should be optimally treated with j_2 or the remaining treatments, $\{j_3, ..., j_k\}$. Again, this is a problem of learning a binary treatment rule so we can perform estimation similar to the first step. With the second decision rule, we can check whether the patient should be treated with j_2 or not. If yes, we conclude $\mathbf{X} \in \mathcal{X}_k^c$; otherwise, we are still uncertain whether $\mathbf{X} \in \mathcal{X}_k$.

Continue this process sequentially in the third step till the kth step when there is only treatment category k in consideration. Consequently, for this given sequence, $\{j_1, ..., j_{k-1}\}$, the optimal treatment for this patient is k, i.e., $\mathbf{X} \in \mathcal{X}_k$, if and only if at each step, the binary decision learning concludes

that the patient should not be treated by $j_1, j_2, ..., j_{k-1}$ in turn. The choice of the ordered sequence $\{j_1, ..., j_{k-1}\}$ is arbitrary, so we propose to consider all possible permutations of $\{1, ..., k-1\}$. Then a patient with $\mathbf{X}$ should be treated with treatment k once he/she is determined to have optimal treatment as k in at least one permuted sequence.

The above procedure only provides a treatment rule that decides whether a subject should be treated with k or not. Thus, for a subject who is determined not to be treated with k, we need to determine which treatment from the remaining $\{1, 2, ..., k-1\}$ options is optimal. This can be carried out using the following procedure. We only consider patients whose optimal treatment is not k based the previous procedure and whose actual treatments received are not k. For these patients, the treatment options can only be one of $\{1, 2, ..., k-1\}$, so the goal reduces to finding the optimal treatment decision within $(k-1)$ categories. Therefore, we can repeat the previous procedure but consider treatment $(k-1)$ as the target in treatment optimization. At the end, we should obtain a treatment rule that determines whether a subject should be treated with $(k-1)$.

Finally, the same procedure can be carried out sequentially to decide which patients should be treated optimally using $(k-2), ..., 1$ in turn. Clearly, an advantage of SOM learning is that at every step of the sequential learning, we only need to learn a binary decision rule. Thus many learning algorithms for binary decision are applicable. In particular, in our subsequent algorithm and implementation, we adopt the method from O-learning (Zhao et al., 2012) to use a weighted support vector machine (SVM) for this purpose. However, one significant difference is that due to multicategory nature, weights in SOM learning should not only be proportional to the outcome R as in O-learning, but should also reflect imbalanced comparison between one treatment category and the combination of multiple treatment categories. The latter ensures that the derived optimal treatment rule is Fisher consistent, as will be shown later.

2.3 Method and algorithm

Mathematically, we can express SOM learning algorithm as follows. Start from the target decision for treatment k. Consider the jth permutation $\{j_1, ..., j_{k-1}\}$, and let $j_k = k$.

1. At step 1, recall from the previous section that our goal is to learn a binary rule to decide whether a future patient should be treated by option j_1. It is equivalent to estimate the optimal decision function

$f^*_{j1}(\mathbf{X})$ such that the corresponding value for this decision, which is given by

$$E\left[RI(Z_{j_1} f_{j1}(\mathbf{X}) > 0)/\pi_A(\mathbf{X})\right],$$

is maximized with $Z_{j_1} = 2I(A_i = j_1) - 1$. According to Liu et al. (2014), even if R may be negative, this maximization is equivalent to

$$\min E\left[|R|I(Z_{j_1}\text{sign}(R) f_{j1}(\mathbf{X}) \leq 0)/\pi_A(\mathbf{X})\right].$$

Thus, using the empirical data, we minimize the following empirical risk for estimation:

$$n^{-1}\sum_{i=1}^{n} \frac{|R_i|I(Z_{ij_1}\text{sign}(R_i) f_{j1}(\mathbf{X}_i) \leq 0)}{\pi_{A_i}(\mathbf{X}_i)},$$

where $Z_{ij_l} = 2I\{A_i = j_l\} - 1$. Since solving the above problem is NP-hard, we propose to use a weighted support vector machine (SVM) which essentially replaces the above 0-1 loss with a continuous and convex hinge-loss function. Furthermore, since this learning is comparing one treatment category versus $(k-1)$ categories, it is necessary to weight observations with treatment j_1 by $(k-1)/k$ and the others by $1/k$ in order to balance the comparison.

Specifically, define $\pi_{j_l}(\boldsymbol{x}) = P(A_i = j_l | \mathbf{X}_i = \boldsymbol{x})$, where $l = 1, ..., k$. We estimate the optimal decision rule as $\text{sign}(\widehat{f}_{j1}(\boldsymbol{x}))$, where $\widehat{f}_{j1}(\boldsymbol{x})$ minimizes the following empirical risk of a weighted hinge loss:

$$\begin{aligned} V_{nj_1}(f) \;=\; & n^{-1}\sum_{i=1}^{n}\left\{\frac{|R_i|}{\pi_{j_1}(\mathbf{X}_i)} I(Z_{ij_1}\text{sign}(R_i) = 1)[1 - f(\mathbf{X}_i)]_+ \left(\frac{k-1}{k} I\{R_i > 0\} + \frac{1}{k} I\{R_i \leq 0\}\right)\right. \\ & \left. + \frac{|R_i|}{\pi^*_{j_1}(\mathbf{X}_i)} I(Z_{ij_1}\text{sign}(R_i) = -1)[1 + f(\mathbf{X}_i)]_+ \left(\frac{1}{k} I\{R_i > 0\} + \frac{k-1}{k} I\{R_i \leq 0\}\right)\right\} \\ & + \lambda_{nj_1}\|f\|^2, \end{aligned}$$

where $\pi^*_{j_1}(\mathbf{X}_i) = \sum_{l=2}^{k} I\{A_i = j_l\}\pi_{j_l}(\mathbf{X}_i)$, $x_+ = \max(x, 0)$ is the hinge loss, $\|\cdot\|$ denotes a semi-norm for f and λ_{nj_1} is a tuning parameter. Particularly, if we consider a linear decision rule, i.e., $f(\boldsymbol{x}) = \beta^T \boldsymbol{x} + \beta_0$, $\|f\|$ is chosen as the Euclidean norm of β; if a nonlinear decision rule is desired, f will be chosen from a reproduced kernel Hilbert space and $\|f\|$ is the corresponding norm in that space.

2. At step 2, recall from the previous section that this step aims to find the optimal decision to be either treatment j_2 or one of $\{j_3, ..., j_k\}$ among those patients whose optimal treatments are not determined as j_1 from Step 1. Thus, restrict the training data to those samples whose labels are not j_1 and whose optimal treatments are not j_1 as from the previous step. We then estimate a decision rule $\text{sign}(\widehat{f}_{j_2}(\boldsymbol{x}))$ using a weighted SVM by minimizing

$$\begin{aligned} V_{nj_2}(f) &= n^{-1}\sum_{i=1}^{n} I\left(A_i \neq j_1, \widehat{f}_{j_1}(\mathbf{X}_i) < 0\right) \\ &\quad \times \left\{ \frac{|R_i|}{\pi_{j_2}(\mathbf{X}_i)} I(Z_{ij_2}\text{sign}(R_i) = 1)[1 - f(\mathbf{X}_i)]_+ \left(\frac{k-2}{k-1} I\{R_i > 0\} + \frac{1}{k-1} I\{R_i \leq 0\}\right) \right. \\ &\quad \left. + \frac{|R_i|}{\pi^*_{j_2}(\mathbf{X}_i)} I(Z_{ij_2}\text{sign}(R_i) = -1)[1 + f(\mathbf{X}_i)]_+ \left(\frac{1}{k-1} I\{R_i > 0\} + \frac{k-2}{k-1} I\{R_i \leq 0\}\right) \right\} \\ &\quad + \lambda_{nj_2}\|f\|^2, \end{aligned}$$

where $\pi^*_{j_2}(\mathbf{X}_i) = \sum_{l=3}^{k} I\{A_i = j_l\}\pi_{j_l}(\mathbf{X}_i)$, $Z_{ij_l}, \pi_{j_l}(\mathbf{X}_i)$ are defined as same as in step 1, and λ_{nj_2} is a tuning parameter. Note that in additional to weights based on the outcome values, we also weigh the observations from treatment j_2 by $(k-2)/(k-1)$ and the others by $1/(k-1)$ in order to account for the fact that the decision rule is based on comparing one category versus $(k-2)$ categories.

3. In turn, at step $h = 3, ..., k-1$, we obtain the rule $\text{sign}(\widehat{f}_{jh}(\boldsymbol{x}))$ by minimizing

$$\begin{aligned} V_{nj_h}(f) &= n^{-1}\sum_{i=1}^{n} I\left(A_i \neq j_1, ..., A_i \neq j_{h-1}, \widehat{f}_{j_1}(\mathbf{X}_i) < 0, ..., \widehat{f}_{j_{h-1}}(\mathbf{X}_i) < 0\right) \\ &\quad \times \left\{ \frac{|R_i|}{\pi_{j_h}(\mathbf{X}_i)} I(Z_{ij_h}\text{sign}(R_i) = 1)[1 - f(\mathbf{X}_i)]_+ \right. \\ &\quad \times \left(\frac{k-h}{k-h+1} I\{R_i > 0\} + \frac{1}{k-h+1} I\{R_i \leq 0\}\right) \\ &\quad + \frac{|R_i|}{\pi^*_{j_h}(\mathbf{X}_i)} I(Z_{ij_h}\text{sign}(R_i) = -1)[1 + f(\mathbf{X}_i)]_+ \\ &\quad \left. \times \left(\frac{1}{k-h+1} I\{R_i > 0\} + \frac{k-h}{k-h+1} I\{R_i \leq 0\}\right) \right\} \\ &\quad + \lambda_{nj_h}\|f\|^2, \end{aligned}$$

where $\pi^*_{j_h}(\mathbf{X}_i) = \sum_{l=h+1}^{k} I\{A_i = j_l\}\pi_{j_l}(\mathbf{X}_i)$, $Z_{ij_l}, \pi_{j_l}(\mathbf{X}_i)$ are defined as same as above steps. Again, we use weight $(k-1)/(k-h+1)$ for treatment j_h versus $1/(k-h+1)$ for the others to balance comparison. At the end of this sequence, we conclude that the optimal treatment for a patient with

$\boldsymbol{x}$ will be treatment k, the pre-determined target treatment category, if

$$\widehat{f}_{j1}(\boldsymbol{x}) < 0, \quad \widehat{f}_{j2}(\boldsymbol{x}) < 0, \quad ... \quad \widehat{f}_{j,k-1}(\boldsymbol{x}) < 0.$$

For notational simplification, we denote $\widehat{\mathcal{D}}^k_{(j_1,...,j_{k-1})}(\boldsymbol{x}) = 1$ if the above conditions hold and let $\widehat{\mathcal{D}}^k_{(j_1,...,j_{k-1})}(\boldsymbol{x}) = -1$ otherwise.

The choice of this sequential decision rule is based on the permutation $(j_1, ..., j_{k-1})$, and thus may not exhaust the correct classification with label k due to this specific choice. We repeat the above sequential learning for any possible $(k-1)!$ permutations to obtain $\widehat{\mathcal{D}}^k_{(j_1,...,j_{k-1})}(\boldsymbol{x})$. Consequently, our final decision rule to assign a patient with treatment k if and only if $\widehat{\mathcal{D}}^k_{(j_1,...,j_{k-1})}(\boldsymbol{x}) = 1$ for at least one permutation $(j_1, ..., j_{k-1})$. That is, if we define

$$\widehat{\mathcal{D}}^k(\boldsymbol{x}) = \max_{(j_1,...,j_{k-1}) \text{ is permutation of} \{1,..,k-1\}} \widehat{\mathcal{D}}^k_{(j_1,...,j_{k-1})}(\boldsymbol{x}),$$

then the optimal treatment for patient with $\boldsymbol{x}$ is treatment k if and only if $\widehat{\mathcal{D}}^k(\boldsymbol{x}) = 1$.

4. We aim to construct a decision rule to decide whether a patient should be optimally treated with treatment $(k-1)$. We adopt a backward elimination procedure. We delete the patients whose treatment labels are k or whose optimal treatments are k from in the previous step. In other words, we restrict the training dataset to samples with $A_i \neq k$ and $\widehat{\mathcal{D}}^k(\boldsymbol{x}_i) = -1$. Because the data consist of only $(k-1)$ class labels, we use the same SOM learning procedure as before but now set label $(k-1)$ as the target treatment, i.e., the last category in consideration in the above sequential learning algorithm. By this procedure, we obtain a decision rule at each step for each permutation of $\{1, 2, .., k-2\}$, denoted by $\widehat{\mathcal{D}}^{(k-1)}_{(j_1,...,j_{k-2})}(\boldsymbol{x})$ for permutation $(j_1, ..., j_{k-2})$. Let

$$\widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = \max_{(j_1,...,j_{k-2}) \text{ is permutation of} \{1,..,k-2\}} \widehat{\mathcal{D}}^{(k-1)}_{(j_1,...,j_{k-2})}(\boldsymbol{x}).$$

Consequently, the optimal treatment for a patient with $\boldsymbol{x}$ is $(k-1)$ if and only if $\widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = 1$ and $\widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1$.

We continue this backward elimination and sequential learning in turn for treatment $(k-2), ..., 1$

so as to obtain $\widehat{\mathcal{D}}^{(k-2)}(\boldsymbol{x}), ..., \widehat{\mathcal{D}}^{1}(\boldsymbol{x})$. Our final estimated optimal ITR is

$$\widehat{\mathcal{D}}(\boldsymbol{x}) = \begin{cases} k & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = 1 \\ k-1 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = 1 \\ \vdots & \quad \vdots \\ 2 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(2)}(\boldsymbol{x}) = 1 \\ 1 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(2)}(\boldsymbol{x}) = -1 \end{cases}$$

Our algorithm for k-category SOM learning can be summariZed as follows:

Backward loop with target class $s \in \{k, ..., 1\}$:

Inner loop: for each permutation of the remaining treatment assignments except the previously classified ones and target treatment label s , perform a sequence of weighted O-learning to learn $\widehat{\mathcal{D}}^{(j_1,...,j_{s-1})}(\boldsymbol{x})$ for each permutation $(j_1, ..., j_s)$ of $\{1, 2, .., s\}$.

Collect all rules to obtain $\widehat{\mathcal{D}}^{s}(\boldsymbol{x}) = \max_{(j_1,...,j_{s-1}) \text{ is a permutation of } \{1,..,s\}} \widehat{\mathcal{D}}^{s}_{(j_1,...,j_{s-1})}(\boldsymbol{x})$.

After eliminating all samples with actual treatment labels are previously considered treatment or whose optimal treatments are within any of the previous labels, go to the backward loop step.

We note that SOM learning requires a total of

$$\sum_{l=1}^{k} (l-1) \times (l-1)! = k! - 1$$

weighted binary SVM classifications. However, because of the sequential data elimination, the size of the input dataset keeps decreasing in a proportional fashion. Therefore, SOM learning can be computationally efficient due to the fast implementation of SVM and reduced data sizes. In our numeric examples, SVM at each step is implemented in MATLAB with quadratic programming.

3 Theoretical Justification

In this section, we establish Fisher consistency of the optimal ITR estimated using SOM learning. We further obtain a risk bound for the estimated ITR and show how the bound can be improved for certain situations.

3.1 Fisher consistency

We provide the theoretical property of Fisher consistency for the proposed SOM learning. Specifically, when the sample size is infinity, we show that the derived ITR is the same as the true optimal ITR given as

$$\text{argmax}_{l=1}^{k} E[R|\mathbf{X} = \boldsymbol{x}, A = l].$$

Let $f_{jl}^{*}(\boldsymbol{x})$ be the counterpart of $\widehat{f}_{jl}(\boldsymbol{x})$ in the SOM learning procedure when $n = \infty$ and the tuning parameters vanishes. Let $\mathcal{D}_{(j_1,\ldots,j_s)}^{*l}(\boldsymbol{x})$ and $\mathcal{D}^{*l}(\boldsymbol{x})$ be the corresponding limits of $\widehat{\mathcal{D}}_{(j_1,\ldots,j_s)}^{l}(\boldsymbol{x})$ and $\widehat{\mathcal{D}}^{l}(\boldsymbol{x})$, respectively, when $n = \infty$. Then the limit of the ITR from SOM learning is given by

$$\mathcal{D}^{*}(\boldsymbol{x}) = \begin{cases} k & \mathcal{D}^{*(k)}(\boldsymbol{x}) = 1 \\ k-1 & \mathcal{D}^{*(k)}(\boldsymbol{x}) = -1, \mathcal{D}^{*(k-1)}(\boldsymbol{x}) = 1 \\ \vdots & \quad\vdots \\ 2 & \mathcal{D}^{*(k)}(\boldsymbol{x}) = -1, \ldots, \mathcal{D}^{*(3)}(\boldsymbol{x}) = -1, \mathcal{D}^{*(2)}(\boldsymbol{x}) = 1 \\ 1 & \mathcal{D}^{*(k)}(\boldsymbol{x}) = -1, \ldots, \mathcal{D}^{*(3)}(\boldsymbol{x}) = -1, \mathcal{D}^{*(2)}(\boldsymbol{x}) = -1. \end{cases} \tag{2}$$

The following result holds.

Theorem 1. *SOM learning rule $\mathcal{D}^{*}(X)$ is Fisher consistent. That is, $\mathcal{D}^{*}(\boldsymbol{x}) = l$ if and only if $E[R|\mathbf{X} = \boldsymbol{x}, A = l] = \max_{h=1}^{k} E[R|\mathbf{X} = \boldsymbol{x}, A = h]$ for $l = 1, ..., k$.*

Theorem 1 provides a theoretical justification that the proposed SOM learning yields the true optimal ITR asymptotically. The proof of Theorem 1 is given in the appendix. The key result is to show that at each step of SMO learning, we compare the conditional mean $E[R|\mathbf{X}, A = j_1]$ with the average value of $E[R|\mathbf{X}, A = j_2]$, where j_1 is the treatment category in consideration at this step while j_2 is any treatment category among the remaining options.

3.2 Risk bounds

For any ITR $\mathcal{D}(\boldsymbol{x})$ associated with decision function $\mathcal{D}(\boldsymbol{x})$, define

$$\mathcal{R}(\mathcal{D}) = E\left[\frac{R}{\pi_A(\boldsymbol{x})} I(A \neq \mathcal{D}(\mathbf{X})\right]$$

where $j = 1, ..., k$, $\pi_A(\boldsymbol{x}) = \sum_{j=1}^{k} I(A = j)P(A = j|\boldsymbol{x})$; and let $\mathcal{R}^* = \mathcal{R}(\mathcal{D}^*)$. Clearly, $\mathcal{R}(\mathcal{D})$ and $\mathcal{R}^*$ correspond to $E[R]$ subtracting the value for $\mathcal{D}$ and $\mathcal{D}^*$, respectively. In the section, we will derive the convergence rate of the estimated value function from the optimal value, which is equivalent to $\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}^*$, under some regularity conditions and assuming that the functional spaces for f_{jl} in our SOM learning are from a reproducing kernel Hilbert space (RKHS) with Gaussian kernel and bandwidth $1/\sigma_n$.

For any l and subset, $\mathcal{S}$, in $\{1, 2, ..., k\}$ where $l \notin \mathcal{S}$, we define

$$\eta_{l,\mathcal{S}}(\boldsymbol{x}) = \frac{E[R|\mathbf{X} = \boldsymbol{x}, A = l]}{|\mathcal{S}|^{-1} \sum_{h \in \mathcal{S}} E[R|\mathbf{X} = \boldsymbol{x}, A = h]},$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. That is, $\eta_{l,\mathcal{S}}(\boldsymbol{x})$ is the ratio between the mean outcome in treatment arm l and the average mean outcome in treatment options from $\mathcal{S}$. We assume that the following conditions hold:

(C.1) (Geometric noise conditions) There exist $q, \beta > 0$, and a constant c such that for any l and set $\mathcal{S}$ with $l \notin \mathcal{S}$, it holds that

$$P\left\{\left|\eta_{l,\mathcal{S}}(\mathbf{X}) - 1\right| < t\right\} \leq (ct)^q,$$

and moreover,

$$E\left[\exp\left(-\frac{\Delta(\mathbf{X})^2}{t}\right)\left|\eta_{l,\mathcal{S}}(\mathbf{X}) - 1\right|\right] \leq ct^\beta,$$

where $\Delta(\mathbf{X})$ denotes the distance from $\mathbf{X}$ to the boundary defined as $\{\boldsymbol{x} : \eta_{l,\mathcal{S}}(\boldsymbol{x}) = 1\}$.

(C.2) The distribution of $\mathbf{X}$ satisfies tail component condition $P(|\mathbf{X}| \geq r) \leq cr^{-\tau}$ for some $\tau \in (0, \infty]$ and $E[|R||A = a, \mathbf{X} = \boldsymbol{x}]$ is uniformly bounded away from zero and infinity.

(C.3) There exists λ_n such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$. Moreover, all tuning parameters λ_{nj}'s in SOM satisfy $M^{-1}\lambda_n \leq \lambda_{nj} \leq M\lambda_n$ for a positive constant M. We further assume $\sigma_n \to \infty$.

Remark 1. In condition (C.1), the constants q and β are called noise exponent and marginal noise exponent, respectively. They are used to characteriZe the data distribution near the decision boundary

at each step of SOM where we compare treatment j_l versus any subset of $\{j_{l+1}, ..., j_k\}$. In particular, when the boundary is fully separable, that is, $|\eta_{l,\mathcal{S}} - 1| > \delta_0$ for a constant δ_0, these conditions hold for $q = \beta = \infty$. In condition (C.2), τ describes the decay of the distribution of $\mathbf{X}$. Obviously, when $\mathbf{X}$ is bounded, $\tau = \infty$. Condition (C.3) assumes the choice of tuning parameter and bandwidth in RKHS. We choose this simplification for convenience, although we can allow the tuning parameter and bandwidth to be different for each treatment decision in the proposed method. Under conditions (C.1)-(C.3), the following theorem holds.

Theorem 2. Under conditions (C.1)-(C.3), for any $\epsilon_0 > 0$, $d/(d+\tau) < p \leq 2$, there exists a constant C such that for any $\epsilon > 1$ and $\sigma_n = \lambda_n^{-q/(2\beta(1+q))}$, with probability at least $1 - e^{-\epsilon}$,

$$\mathcal{R}(\widehat{\mathcal{D}}) \leq \mathcal{R}^* + C \left\{ \lambda_n^{-\frac{2}{2+p} + \frac{(2-p)(1+\epsilon_0)}{(2+p)(1+q)}} n^{-\frac{2}{2+p}} + \frac{\epsilon}{n\lambda_n} + \lambda_n^{\frac{q}{1+q}} \right\}^{\frac{q}{1+q}}.$$

Remark 2. Suppose that $\mathbf{X}$ is bounded such that $\tau = \infty$ in condition (C.2). By choosing the optimal λ_n for the last two term in the right-hand side, i.e., $\lambda_n = n^{-(1+q)/(1+2q)}$, we find that the convergence rate is a polynomial order of n, where the order is given by $q/(1+2q)$. If furthermore, the separating boundaries are all completely separable such that $q = \infty$, then the convergence rate is close to the square-root rate.

4 Simulation Studies

We conducted extensive simulation studies from four different settings to examine the small-sample performance of SOM learning. In the first three settings, 20 feature variables were simulated from multivariate normal distribution, where the first 10 variables $X_1, X_2, ..., X_{10}$ had a pairwise correlation of 0.8, the remaining 10 variables were uncorrelated, and the marginal distribution for each variable was $N(0, 1)$. We generated 3-category random treatment assignments with equal probability, i.e. $P(A = 1|\mathbf{X}) = P(A = 2|\mathbf{X}) = P(A = 3|\mathbf{X}) = 1/3$. The reward functions were generated as follows:

Setting 1. $R = X_4 + (X_1 + X_2)I\{A = 2\} + (-X_1 + X_3)I\{A = 3\} + 0.5 \times N(0, 1)$

Setting 2. $R = X_4 + (X_2^2 - X_1^2)I\{A = 2\} + X_3^3 I\{A = 3\} + 0.5 \times N(0, 1)$

Setting 3. $R = (X_1 - 0.2) \times (I\{A = 1\} - I\{X_1 > 0.3\})^2 + (X_2 + 0.3) \times (I\{A = 2\}$
$-I\{X_2 > -0.5\})^2 + (X_3 + 0.5) \times (I\{A = 3\} - I\{X_3 > 0\})^2 + 0.5 \times N(0, 1)$.

In the last setting (*Setting 4*), we imitated a situation where the entire population consisted of

a finite number of latent subgroups for which the optimal treatment rule was the same within each subgroup. Specifically, we considered 10 latent groups and the true optimal treatment category of each groups was in turn $A^* = 3, 3, 1, 2, 2, 1, 2, 3, 3, 1$ in turn. To generate data mimicking a randomized trial, for each subject, a 3-category treatment A, was randomly assigned with equal probability. The reward outcome was generated as $R = 3 \times I\{A = A^*\} - I\{A \neq A^*\} + 0.5 \times N(0, 1)$. Furthermore, instead of observing the group labels, we generated feature variables that were informative of the latent group membership: we simulated 30 feature variables from a multivariate normal distribution, where the first 10 variables $X_1, X_2, ..., X_{10}$ had a pairwise correlation of 0.8, the remaining 20 variables were uncorrelated, and the variance for each variable was 1. Moreover, $X_1, X_2, ..., X_{10}$ had mean values of μ_l for the latent group l, which were generated from $N(0, 5)$, while the means of $X_{11}, ..., X_{30}$ were all 0. Therefore, only $X_1, X_2, ..., X_{10}$ were informative of the group labels due to different μ_l. The empirical observation for each subject consisted of the treatment assignment A, the feature variables $X_1, ..., X_{30}$, and the outcome R.

For each simulated data, we applied SOM learning to estimate the optimal ITR. At each step, we fitted a weighted SVM with a linear kernel by solving the corresponding dual problem via quadratic programming. The tuning parameter was chosen using cross-validation. Furthermore, we compared SOM learning with regression-based Q-learning, one-vs-all (OVA) and one-vs-one (OVO) based on the value function (reward) of the estimated optimal treatment rules. Q-learning was obtained by fitting a linear model, regressing R on $\mathbf{X}$, A and their interactions, in which A was replaced by dummy variables created for each category of A. For OVA and OVO, how were they exactly done? For each setting, we compared the four methods for different sample sizes: n =300, 600, and 900.

Figure 1 to 4 present the results of the optimal treatment mis-allocation rates and the estimated value functions from 100 replicates and difference sample sizes, which were computed in an independently generated test data of size 3 million. Furthermore, Table 1 to 4 summariZe the average of the marginal mis-allocation rates of each category.

In the first setting, we observe that Q-learning gains higher values and lower mis-allocation rates of the optimal ITR compared to SOM under all sample sizes because the regression model used in Q-learning is correctly specified. The estimated values of SOM learning become closer to those of Q-learning as the sample size increases. In the latter three non-linear settings, the regression model in Q-learning is misspecified, so it performs poorly under all sample sizes. Instead, SOM learning

outperforms all comparators including OVA and OVO in all the simulation settings. For SOM learning, we also used Gaussian kernel in our method and found negligible difference from using linear kernel. However, since computation using the former is much more intensive, we recommend to use linear kernel in practice.

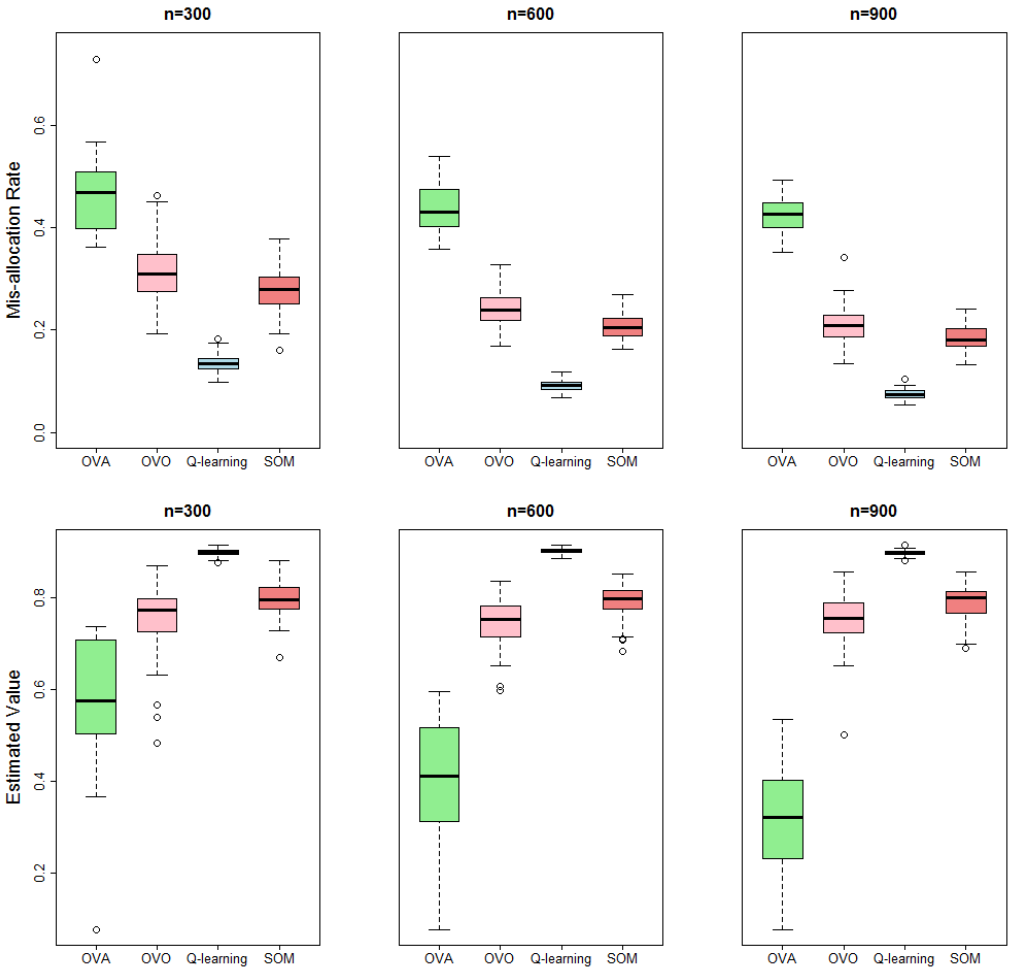

Figure 1: Box plots of the optimal treatment mis-allocation rates and estimated value functions of SOM, Q-learning, OVA and OVO for setting 1 with sample size of 300, 600 and 900. The optimal value is 0.9245.

Table 1: Category Mis-allocation Rates (%) of Setting 1

Category	n=300				n=600				n=900			
	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO
1	19.6	10.9	41.3	23.3	14.6	7.4	40.1	17.8	12.7	6.0	39.6	15.3
2	12.6	5.4	19.3	14.5	10.5	3.6	16.8	11.6	9.8	2.9	16.0	10.5
3	23.2	10.7	31.7	25.3	16.5	7.2	30.6	18.9	14.4	5.9	29.4	16.0

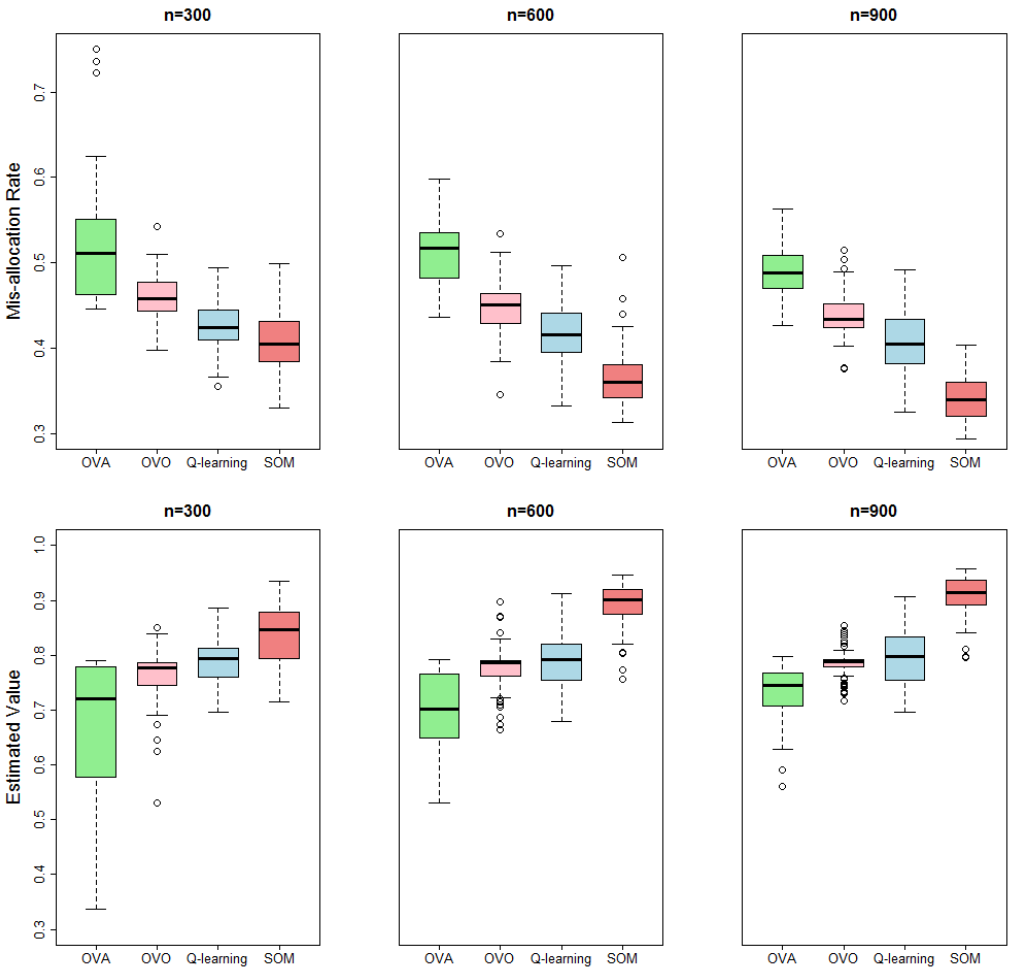

Figure 2: See Table 1. The optimal value is 1.0585.

Table 2: Category Mis-allocation Rates (%) of Setting 2

Category	n=300				n=600				n=900			
	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO
1	23.8	27.7	41.4	31.2	20.9	27.1	42.1	31.0	19.0	26.7	40.4	30.3
2	34.4	38.8	39.3	39.8	30.7	38.8	38.9	39.4	29.0	38.5	38.9	39.0
3	23.1	19.0	22.4	21.1	21.3	17.0	20.7	19.5	20.3	16.3	18.7	18.3

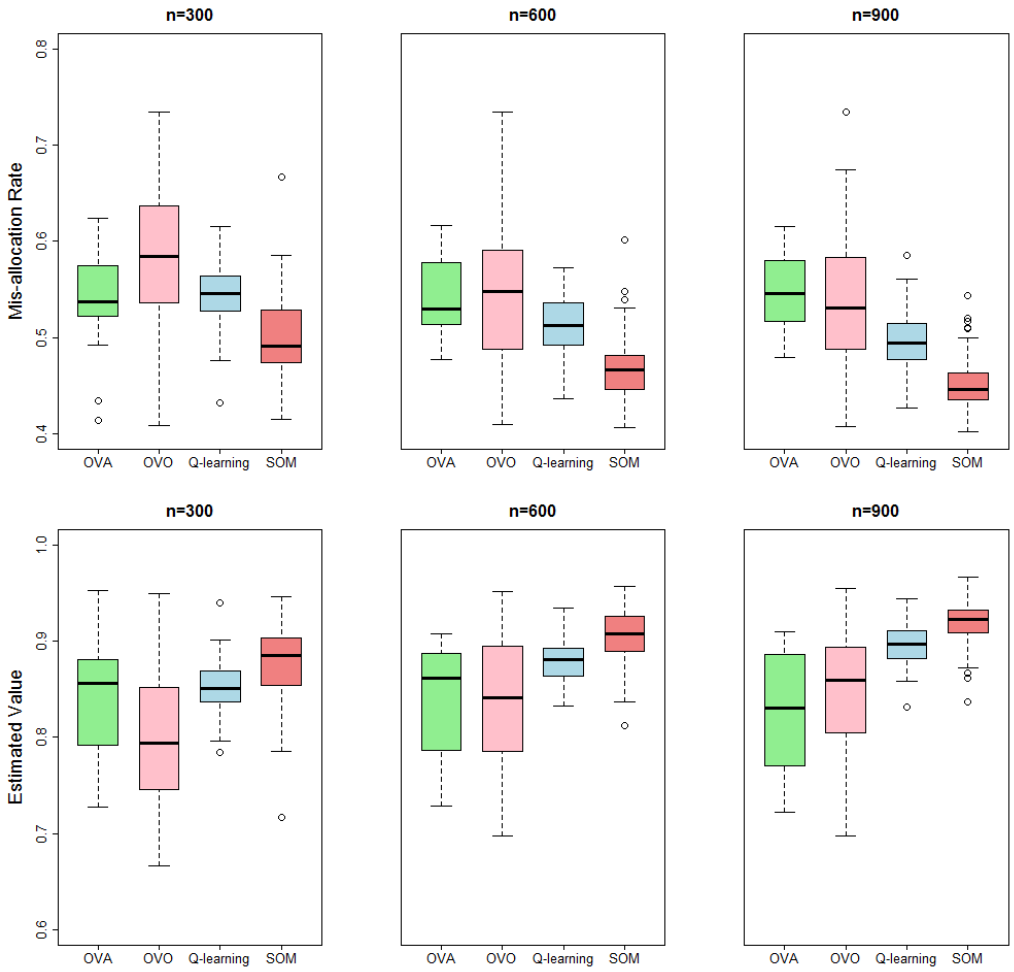

Figure 3: See Table 1. The optimal value is 1.1438.

Table 3: Category Mis-allocation Rates (%) of Setting 3

Category	n=300				n=600				n=900			
	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO
1	28.1	31.0	35.1	35.8	25.4	27.7	34.6	31.7	23.7	25.7	36.6	29.9
2	33.0	39.1	36.7	40.0	31.5	38.2	36.9	38.6	30.3	37.5	36.2	37.6
3	38.5	38.8	37.2	40.7	36.6	36.6	37.0	38.7	36.0	36.1	36.9	39.5

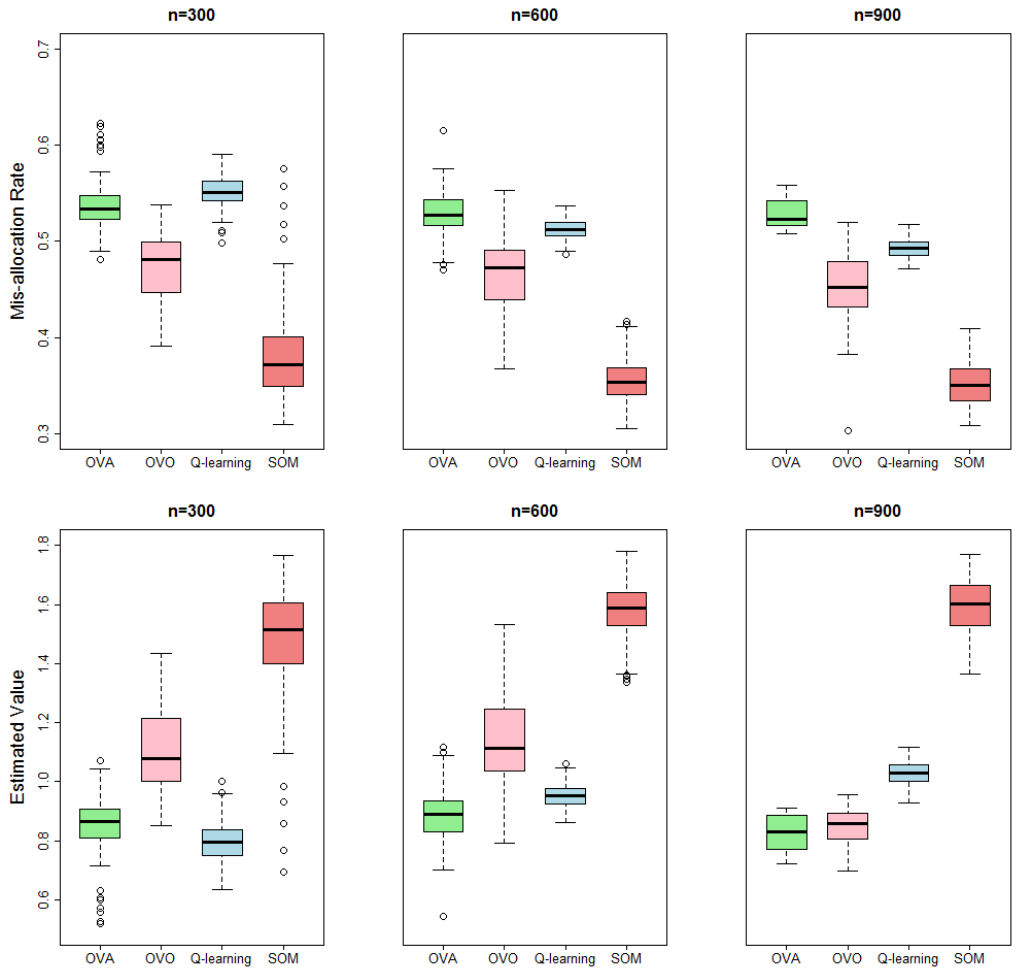

Figure 4: See Table 1. The optimal value is 2.9999.

Table 4: Category Mis-allocation Rates (%) of Setting 4

Category	n=300				n=600				n=900			
	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO	SOM	Qlearn	OVA	OVO
1	14.7	35.7	43.3	30.6	13.1	33.0	43.6	30.2	13.2	31.2	44.3	28.6
2	27.5	37.5	30.9	32.0	25.7	34.8	29.4	31.2	25.1	33.4	29.0	30.8
3	33.9	37.1	33.4	32.0	32.6	34.8	32.8	31.8	31.9	34.0	32.3	31.4

5 Application to REVAMP Study

We applied the proposed method to real data collected from the Research Evaluating the Value of Augmenting Medication with Psychotherapy (REVAMP) trial (Kocsis et al., 2009). The study

aimed to evaluate the efficacy of adjunctive psychotherapy in the treatment of patients with chronic depression who have failed to fully respond to initial treatment with an antidepressant medication. Among the 808 participants in phase I, 491 were classified as nonresponders (NRs) or partial responders (PRs) and entered phase II. The 491 participants were then randomized to receive (1) continued pharmacotherapy and augmentation with brief supportive psychotherapy (MEDS+BSP), (2) continued pharmacotherapy and augmentation with cognitive behavioral analysis system of psychotherapy (MEDS+CBASP), or (3) continued optimized pharmacotherapy (MEDS) alone, and followed for 12 weeks. The primary outcome is the Hamilton Scale for Depression (HAM-D) scores at the end of 12-week follow-up. There were 17 feature variables including participants' demographics, patient's treatment efficacy expectation, social adjustment scale, mood and anxiety symptom, and depression experience, as well as phase I depressive symptom measures such as rate of change in HAM-D score over phase I, HAM-D score at the end of phase I, and Quick Inventory of Depression Symptoms (QIDS) scores during phase I. After eliminating participants with missing data, the final dataset contained 348 participants, among which 147, 135, and 66 were assigned in MEDS+BSP, MEDS+CBASP, and MEDS only group, respectively. The mean HAM-D at the end of Phase II study in each treatment arm is summariZed in Table 6. MEDS+CBASP had the lowest post-treatment HAM-D score, but there was that no statistically significant differences in changes on HAM-D scores during phase II detected among the 3 treatment groups (Kocsis et al., 2009).

Table 5: Mean and standard deviation of the value function (HAM-D scores) from 2-fold cross-validation procedure with 500 repetitions.

Method	SOM learning	Q-learning	OVA	OVO
Value in test sample*	9.95 (2.085)	12.64 (2.009)	11.97 (1.150)	11.15 (1.458)
One-fits-all	MEDS+BSP	MEDS+CBASP	MEDS	
Value in test sample	12.90	10.62	12.53	

*: Value function is the average HAM-D score at end of phase II for patients following an estimated optimal treatment (a smaller HAM-D score indicates a better outcome).

Our analysis goal is to use 17 feature variables to estimate the optimal individualiZed treatment strategy among three different options. The feature variables include participants' value function (average HAM-D scores) under the ITR can be as low as possible. All feature variables were standardiZed

before the analyses. We applied SOM learning and compared with Q-learning that uses $(1, \mathbf{X}, A, \mathbf{X}A)$ in the regression model, where $\mathbf{X}$ represents feature variables and A is the randomized treatment assignments, as well as OVA and OVO. The expected HAM-D for an ITR was calculated from 2-fold cross validation of the data with 500 replicates: at each replicate, we randomly split the data into one training sample and testing sample; we then applied SOM learning to learn the optimal ITR using the training data and computed the expected value in the testing sample under this estimate rule. The averages of the cross-validated value functions from the two methods are presented in Table 5 and their distributions over cross-validations are plotted in Figure 5. The last 3 columns in Table 5 are the values from non-personaliZed rules where the same treatment is recommended for all patients. With a value function of 9.95, the SOM learning achieved a lower HAM-D compared to Q-learning and any of the non-personaliZed rule.

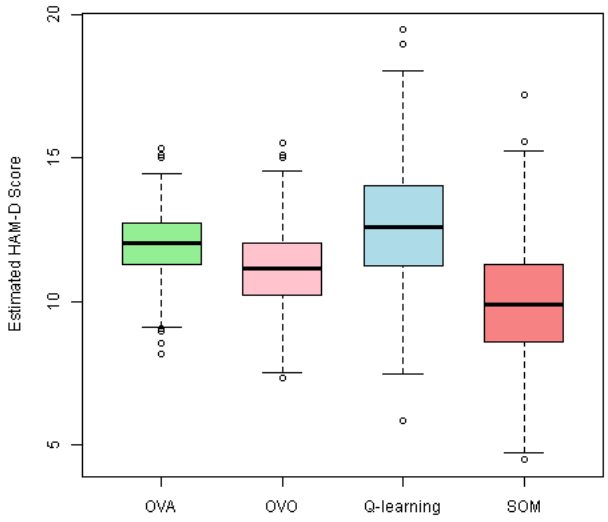

Figure 5: Box plot of the value function for the optimal ITR estimated by various methods from 2-fold cross-validation with 500 repetitions using REVAMP data: HAM-D score after phase II treatment (a smaller score indicates a better outcome).

There were 122, 114, and 112 patients predicted to have MEDS+BSP, MEDS+CBASP, and MEDS alone as optimal treatment, respectively. Table 6 presents the coefficients of the $3! - 1 = 5$ models derived from SOM learning rule in REVAMP study. Model 1 and model 2 corresponds to the 2 permutations of inner loop, determining whether a subject should be assigned to MEDS only group or not. After eliminating the possibility of being assigned to MEDS only group, model 3 classifies

a subject into MEDS+BSP or MEDS+CBASP treatment group. Let $\widehat{\boldsymbol{\beta}}_{11}, \widehat{\boldsymbol{\beta}}_{12}, \widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}, \widehat{\boldsymbol{\beta}}_{3}$ be the estimated coefficients of model 1(1), 1(2), 2(1), 2(2) and 3, respectively. A patient will be assigned with: MEDS if he has $\{\mathbf{X}^T\widehat{\boldsymbol{\beta}}_{11} < 0, \mathbf{X}^T\widehat{\boldsymbol{\beta}}_{12} < 0\}$, or $\{\mathbf{X}^T\widehat{\boldsymbol{\beta}}_{21} < 0, \mathbf{X}^T\widehat{\boldsymbol{\beta}}_{22} < 0\}$; MEDS+CBASP if he has not been assigned to MEDS and $\mathbf{X}^T\widehat{\boldsymbol{\beta}}_{3} < 0$; MEDS+BSP if he has not been assigned to MEDS and $\mathbf{X}^T\widehat{\boldsymbol{\beta}}_{3} > 0$. The column "Norm" reports the overall effect of feature variables on the optimal treatment decision rule as the L_2 norm of all coefficients for predicting each model.

Table 6: Coefficients estimated from SOM learning in REVAMP study (ranked by the overall effect of a feature variable).

Feature Variable	Model 1(1)	Model 1(2)	Model 2(1)	Model 2(2)	Model 3	Norm*
HAM-D phase I change	0.1604	-0.0558	1.5224	3.4055	-0.0120	3.7342
QISD phase I change	-0.1809	0.1523	0.4728	0.8132	-0.0585	1.8896
Gender (Male)	-0.4162	-0.5120	-0.1823	-1.0248	-0.0101	1.2325
Drug abuse	0.7450	0.0934	0.1487	0.9402	0.0221	1.2125
Tx efficacy expectation	0.5920	0.2344	0.2541	0.8006	0.0115	1.0541
Social adjustment	-0.3592	0.4067	0.6779	-0.0529	0.0014	0.8699
CBASP expectation	-0.3941	0.1358	0.1948	-0.5616	-0.0649	0.7289
Current alcohol use	0.1987	0.0608	-0.4850	-0.2004	0.0515	0.5664
Anxious Arousal	0.0170	0.5054	0.1416	0.1307	-0.0066	0.5412
Phase I response	-0.1683	-0.2817	0.1355	0.2846	-0.0282	0.4559
BSP expectation	0.2014	-0.1000	-0.1100	0.3462	0.0455	0.4296
General Distress Anxious	-0.0512	-0.3460	-0.0804	-0.0555	0.0111	0.3633
Freq of side effects	0.0459	0.1544	0.0879	0.0694	-0.0436	0.2010
QISD end of phase I	0.1524	-0.0524	-0.0897	-0.0311	0.0587	0.1961
HAMD end of phase I	-0.0634	0.0035	-0.0710	-0.1520	-0.0206	0.1806
Dysfunctional Attitudes	-0.0077	0.0119	0.0007	-0.0111	-0.0190	0.0262
Age	0.0022	-0.0013	-0.0009	0.0008	0.0006	0.0029

*: "Norm" measures the overall effect of a variable on the optimal treatment assignment rule as the L_2 norm of all coefficients for predicting each model.

The overall most predictive variable as determined by the norm of the coefficients in estimating

the optimal ITR is the phase I HAM-D rate of change, followed by phase I QIDS rate of change. Both variables are most predictive of patients with MEDS alone as the optimal choice compared to two other combined pharmacotherapy and psychotherapy. Gender, history of drug use, and patients expectancy of treatment efficacy are also informative with an overall effect size greater than 1. Gender is also most predictive of MEDS alone versus two combined therapies alternatives with females preferring the latter. Other predictive variables include social adjustment scale and CBASP expectation, and current alcohol use. No feature variable has a substantially large effect in model 3, implies that potentially many variables are in play to distinguish MEDS+BSP versus MEDS+CBASP. In a recent analysis of another randomized trial on major depressive disorder comparing NefaZodone, CBASP or the combination of the two treatments, obsessive compulsive and past history of alcohol dependence (Gunter et al., 2011), race, and education level (Klein et al., 2011) were identified as predictive by Q-learning. Our analyses identified several additional feature variables as informative.

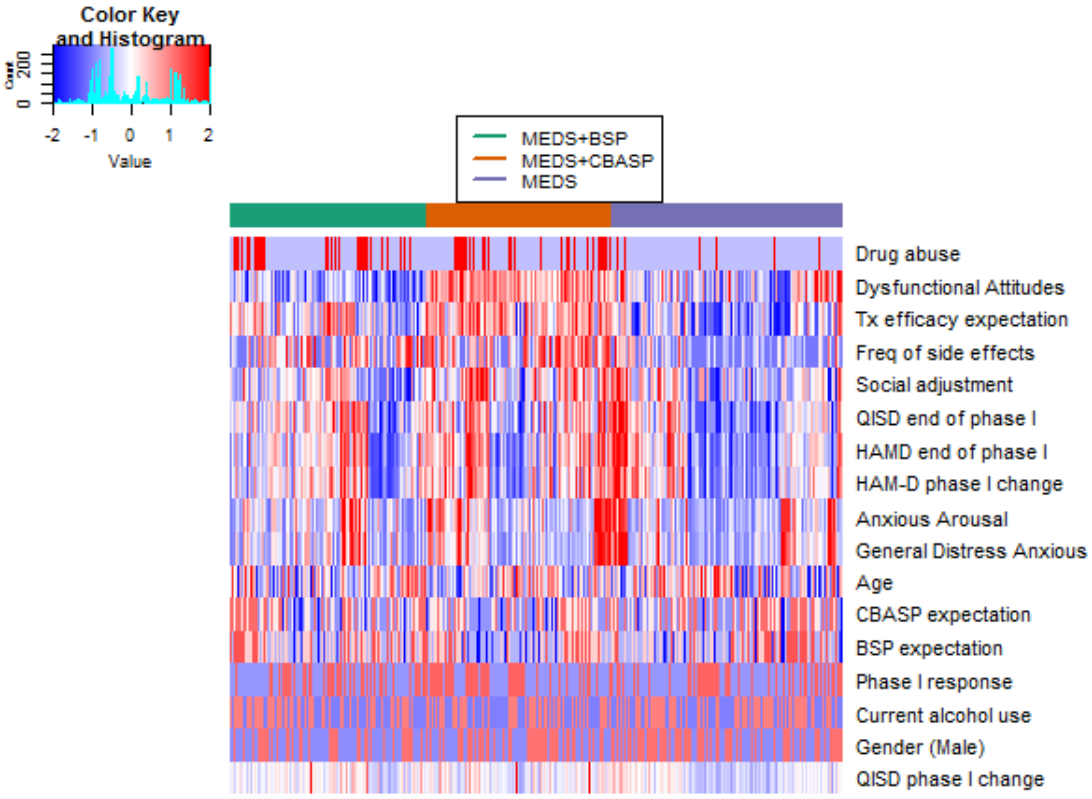

Figure 6: Heatmap of 17 standardiZed feature variables on all patients. Row corresponds to feature variable and column corresponds to patients stratified by predicted optimal treatment.

To further visualiZe the relationship between feature variables and the optimal treatment for each individual, in figure 6, we present the heatmap of 17 standardiZed feature variables by predicted optimal treatment on all subjects. We can see that history of drug abuse has a different pattern between patients with MEDS alone as optimal choice and the other two groups and thus may be informative of distinguishing MEDS alone versus others; dysfunctional attitudes, and patient's treatment efficacy expectation, frequency of sides effects, HAM-D rate of change during phase I, QIDS and HAM-D end of phase I score are informative for distinguishing all three treatments. It is clear that no single variable has a dominating effect on estimating the optimal ITR, and all feature variables in combination may be more effective.

6 Conclusions

We have proposed a sequential outcome weighted learning, SOM learning, to estimate the optimal ITRs with multicategory treatment studies, where each step solves a weighted binary classification problem via support vector machines (SVMs). By carefully choosing weights in each SVM step and combining the treatment decision functions from all steps, we showed that the derived rule from the proposed learning algorithm is Fisher consistent. In the numeric studies, SOM learning yielded more desirable expected value functions as compared to the method based on a standard regression model.

The proposed method can be extended in several directions. First, for some chronic diseases with multi-stage therapy, dynamic treatment regimens (DTRs) can be more powerful in obtaining favorable outcomes than a simple combination of single-stage treatment rules. Various approaches have been developed to estimate optimal DTR, such as Murphy (2003); Robins (2004); Moodie et al. (2007); Zhao et al. (2011); Zhang et al. (2012, 2013); Liu et al. (2014). While our method has focused on single-stage studies only, the proposed procedure can be easily generaliZed to handle multicategory DTR for multiple stage trials. Second, although the proposed method was only applied to a finite number of categories, it can be naturally extended to find optimal personaliZed dose, where treatment is in a continuous scale, after discretiZing the dose variable into a number of categories. However, one challenge is to determine the number of the categories and the way of discretiZation. One possibility is to include these uncertainty as parameters to learn in SOM learning. Further research is worth pursuing.

A major computational cost for SOM learning is to go through all possible permutations of the

treatment categories. Since the sequential learning for each permutation can be carried out independent of one another, one potential improvement in implementation is to incorporate distributed computing to use this parallel computation structure.

Finally, although we suggested to treat the most prevalent treatment as the first target optimal treatment in the SOM learning procedure, this may result in few cases for later treatments in consideration and cause large misallocation rates for patients whose optimal treatments are less prevalent. In practice, when different treatments have different importance, for instance, due to the need to balance efficacy and risk, the order of the targeted treatments should be taken into account of the practical importance.

Acknowledgements

This research is support in part by U.S. NIH grants NS082062 and NS073671. Data and used in the preparation of this manuscript were obtained and analyZed from the controlled access datasets distributed from the NIMH-supported National Database for Clinical Trials (NDCT). NDCT is a collaborative informatics system created by the National Institute of Mental Health to provide a national resource to support and accelerate discovery related to clinical trial research in mental health. Dataset identifier(s): Research Evaluating the Value of Augmenting Medication with Psychotherapy (REVAMP) #2153. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIMH or of the Submitters submitting original data to NDCT.

Appendix

A.1 Proof of Theorem 1

We start from class label k following the order in SOM. First, we show $\mathcal{D}^*(\boldsymbol{x}) = k$ if and only if $E[R|\mathbf{X} = \boldsymbol{x}, A = k] = \max_{l=1}^{k} E[R|\mathbf{X} = \boldsymbol{x}, A = l]$. For any $\boldsymbol{x}$ with $\mathcal{D}^*(\boldsymbol{x}) = k$, by the definition of $\mathcal{D}^*$, there exists a permutation $(j_1, ..., j_{k-1})$ of $\{1, ..., k-1\}$ such that $\mathcal{D}_l^{*(k)}(\boldsymbol{x}) = -1$ for $l = j_1, ..., j_{k-1}$. That is,

$$f_{j_1}^*(\boldsymbol{x}) < 0, f_{j_2}^*(\boldsymbol{x}) < 0, \ldots, f_{j_{k-1}}^*(\boldsymbol{x}) < 0, \tag{3}$$

where f^*_{jl} is the counterpart of $\widehat{f}_{j_1}$ when $n = \infty$.

On the other hand, from the estimation of $\widehat{f}_{j_1}$, it is clear that $f^*_{j_1}$ is the minimizer of the expectation of a weighted hinge loss corresponding to $V_{n,j1}$, which is given by

$$
\begin{aligned}
&E\left[\frac{k-1}{k}\frac{R^+}{\pi_{j_1}(\boldsymbol{x})}I(A=j_1)[1-f(\mathbf{X})]_+\middle|\mathbf{X}=\boldsymbol{x}\right]+E\left[\frac{1}{k}\sum_{l=2}^{k}\frac{R^-}{\pi_{j_l}(\boldsymbol{x})}I(A=j_l)[1-f(\mathbf{X})]_+\middle|\mathbf{X}=\boldsymbol{x}\right]\\
&+E\left[\frac{1}{k}\sum_{l=2}^{k}\frac{R^+}{\pi_{j_l}(\boldsymbol{x})}I(A=j_l)[1+f(\mathbf{X})]_+\middle|\mathbf{X}=\boldsymbol{x}\right]+E\left[\frac{k-1}{k}\frac{R^-}{\pi_{j_1}(\boldsymbol{x})}I(A=j_1)[1+f(\mathbf{X})]_+\middle|\mathbf{X}=\boldsymbol{x}\right]\\
=&E\left[\frac{k-1}{k}R^+\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right][1-f(\mathbf{X})]_+ +\sum_{l=2}^{k}E\left[\frac{R^-}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right][1-f(\mathbf{X})]_+\\
&+\sum_{l=2}^{k}E\left[\frac{R^+}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right][1+f(\mathbf{X})]_+ +E\left[\frac{k-1}{k}R^-\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right][1+f(\mathbf{X})]_+
\end{aligned}
$$

where $R^+ = RI\{R>0\}$, $R^- = -RI\{R\le 0\}$, and $R = R^+ - R^-$.

We first consider the case when $f(\boldsymbol{x}) \in (-\infty, -1]$, the equation above can be reduced to

$$
\left(E\left[\frac{k-1}{k}R^+\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right]+\sum_{l=2}^{k}E\left[\frac{R^-}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right]\right)(-f(\mathbf{X}))+constant \tag{4}
$$

It's clear that we cannot find a minimizer for (2). Similarly, the minimizer cannot be in the interval $f(\boldsymbol{x}) \in [1,\infty)$. Therefore, we only consider $f(\mathbf{X}) \in (-1,1)$. Then the expectation of a weighted hinge loss corresponding to $V_{n,j1}$ above is:

$$
\begin{aligned}
&E\left[\frac{k-1}{k}R^+\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right][1-f(\mathbf{X})]_+ +\sum_{l=2}^{k}E\left[\frac{R^-}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right][1-f(\mathbf{X})]_+\\
&+\sum_{l=2}^{k}E\left[\frac{R^+}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right][1+f(\mathbf{X})]_+ +E\left[\frac{k-1}{k}R^-\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right][1+f(\mathbf{X})]_+\\
=&\left(\sum_{l=2}^{k}E\left[\frac{R}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right]-E\left[\frac{k-1}{k}R\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right]\right)f(\mathbf{X})+constant
\end{aligned}
$$

That is, $f^*_{j_1}(\mathbf{X}) < 0$ is equivalent to

$$
E\left[\frac{k-1}{k}R\middle|\mathbf{X}=\boldsymbol{x},A=j_1\right] < \sum_{l=2}^{k}E\left[\frac{R}{k}\middle|\mathbf{X}=\boldsymbol{x},A=j_l\right],
$$

which is equivalent to

$$E\left[R|\,\mathbf{X}=\boldsymbol{x},A=j_1\right]<\frac{1}{k-1}\sum_{l=2}^{k}E\left[R|\,\mathbf{X}=\boldsymbol{x},A=j_l\right].$$

Next, we restrict to data with $A\neq j_1$ and $f^*_{j_1}(\mathbf{X})<0$, it is clear that $f^*_{j_2}$ minimizes

$$\begin{aligned}
&E\left[\frac{k-2}{k-1}\frac{R^+}{\pi_{j_2}(\boldsymbol{x})}I(A=j_2)[1-f(\mathbf{X})]\middle|\,\mathbf{X}=\boldsymbol{x},A\neq j_1,f^*_{j_1}(\mathbf{X})<0\right]\\
&+E\left[\frac{1}{k-1}\sum_{l=3}^{k}\frac{R^-}{\pi_{j_l}(\boldsymbol{x})}I(A=j_l)[1-f(\mathbf{X})]\middle|\,\mathbf{X}=\boldsymbol{x},A\neq j_1,f^*_{j_1}(\mathbf{X})<0\right]\\
&+E\left[\frac{1}{k-1}\sum_{l=3}^{k}\frac{R^+}{\pi_{j_l}(\boldsymbol{x})}I(A=j_l)[1+f(\mathbf{X})]\middle|\,\mathbf{X}=\boldsymbol{x},A\neq j_1,f^*_{j_1}(\mathbf{X})<0\right]\\
&+E\left[\frac{k-2}{k-1}\frac{R^-}{\pi_{j_2}(\boldsymbol{x})}I(A=j_2)[1+f(\mathbf{X})]\middle|\,\mathbf{X}=\boldsymbol{x},A\neq j_1,f^*_{j_1}(\mathbf{X})<0\right]\\
=&E\left[\frac{k-2}{k-1}R^+\middle|\,\mathbf{X}=\boldsymbol{x},A=j_2,f^*_{j_1}(\mathbf{X})<0\right][1-f(\mathbf{X})]\\
&+\sum_{l=3}^{k}E\left[\frac{R^-}{k-1}\middle|\,\mathbf{X}=\boldsymbol{x},A=j_l,f^*_{j_1}(\mathbf{X})<0\right][1-f(\mathbf{X})]\\
&+\sum_{l=3}^{k}E\left[\frac{R^+}{k-1}\middle|\,\mathbf{X}=\boldsymbol{x},A=j_l,f^*_{j_1}(\mathbf{X})<0\right][1+f(\mathbf{X})]\\
&+E\left[\frac{k-2}{k-1}R^-\middle|\,\mathbf{X}=\boldsymbol{x},A=j_2,f^*_{j_1}(\mathbf{X})<0\right][1+f(\mathbf{X})]\\
=&\left(\sum_{l=3}^{k}E\left[\frac{R}{k-1}\middle|\,\mathbf{X}=\boldsymbol{x},A=j_l,f^*_{j_1}(\mathbf{X})<0\right]-E\left[\frac{k-2}{k-1}R\middle|\,\mathbf{X}=\boldsymbol{x},A=j_2,f^*_{j_1}(\mathbf{X})<0\right]\right)f(\mathbf{X})\\
&+constant
\end{aligned}$$

Thus, we conclude that

$$\text{sign}(f^*_{j_2}(\mathbf{X}))=\text{sign}(E\left[(k-2)R|\,\mathbf{X}=\boldsymbol{x},A=j_2\right]-\sum_{l=3}^{k}E\left[R|\,\mathbf{X}=\boldsymbol{x},A=j_l\right])I\{f^*_{j_1}(\mathbf{X})<0\}.$$

That is, $f^*_{j2}(\boldsymbol{x})<0$ if and only if

$$E\left[R|\,\mathbf{X}=\boldsymbol{x},A=j_2\right]<\frac{1}{k-2}\sum_{l=3}^{k}E\left[R|\,\mathbf{X}=\boldsymbol{x},A=j_l\right]$$

Continue the same arguments so we establish the relationship between $f^*_{j_l}$ and $E[R|\mathbf{X}=\boldsymbol{x},A=j_l]$

as

$$\text{sign}(f^*_{j_l}(\boldsymbol{x})) = \text{sign}\left(E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_l\right] - \frac{1}{k-l}\sum_{h=l+1}^{k} E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_h\right]\right)$$

In other words, we obtain that for this subject with $f^*_{j_1}(\boldsymbol{x}) < 0, ..., f^*_{j_{k-1}}(\boldsymbol{x}) < 0$, it holds

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_1\right] < \frac{1}{k-1}\sum_{l=2}^{k} E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_l\right], \tag{5}$$

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_2\right] < \frac{1}{k-2}\sum_{l=3}^{k} E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_l\right], \tag{6}$$

$$\vdots \tag{7}$$

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_{k-2}\right] < \frac{1}{2}(E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_{k-1}\right] \tag{8}$$

$$+ E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right]), \tag{9}$$

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_{k-1}\right] < E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right]. \tag{10}$$

Starting from the last inequality in the above, in turn, we have

$$\begin{aligned}
&E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_{k-1}\right] < E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right] \\
&E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_{k-2}\right] < \frac{1}{2}(E[R|\mathbf{X}=\boldsymbol{x}, A=j_{k-1}] + E[R|\mathbf{X}=\boldsymbol{x}, A=k]) \\
&\qquad\qquad < E[R|\mathbf{X}=\boldsymbol{x}, A=k], \\
&\vdots \\
&E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_1\right] < \frac{1}{k-1}\sum_{l=2}^{k} E[R|\mathbf{X}=\boldsymbol{x}, A=j_1] < E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right].
\end{aligned}$$

Therefore,

$$E[R|\mathbf{X}=\boldsymbol{x}, A=k] = \max_{l=1}^{k} E[R|\mathbf{X}=\boldsymbol{x}, A=l].$$

For the other direction, we suppose that

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right] = \max_{l=1}^{k} E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=l\right].$$

We order the expectations to obtain

$$E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_1\right] \leq E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_2\right] \leq ... \leq E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=k\right]$$

Thus all the inequalities in (5)-(10) hold, from equivalence between $f^*_{j_l}$ and $E\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_l\right]$'s, it is straightforward to see that

$$f^*_{j_1}(\boldsymbol{x})<0, ..., f^*_{j_{k-1}}(\boldsymbol{x})<0.$$

In other words, $\mathcal{D}^*(\boldsymbol{x})=k$. Hence, we have proved that SOM learning correctly assigns subjects whose conditional mean outcomes are maximal in treatment k into the optimal treatment k.

To prove the consistency of the remaining classes, obtains the rule for class $(k-1)$ conditional on $A\neq k$ and $\mathcal{D}^*(X)\neq k$. Using the same proof as above, we conclude

$$\mathcal{D}^*(\boldsymbol{x})=(k-1) \text{ if and only if } (k-1)=\text{argmax}_{l=1}^{k-1}\widetilde{E}[R|\mathbf{X}=\boldsymbol{x}, A=l],$$

where $\widetilde{E}\left[R|\,\mathbf{X}=\boldsymbol{x}, A=j_l\right]$ is the conditional expectation of R given $\mathbf{X}=\boldsymbol{x}$, $A\neq k$ and $\mathcal{D}^*(X)\neq k$. Moreover, $\mathcal{D}^*(\boldsymbol{x})\neq k$ implies that $E[R|\mathbf{X}=\boldsymbol{x}, A=k]$ cannot be the maximum. Therefore,

$$(k-1)=\text{argmax}_{l=1}^{k-1}E[R|\mathbf{X}=\boldsymbol{x}, A=l]=\text{argmax}_{l=1}^{k}E[R|\mathbf{X}=\boldsymbol{x}, A=l].$$

That is,

$$\mathcal{D}^*(\boldsymbol{x})=(k-1) \text{ if and only if } (k-1)=\text{argmax}_{l=1}^{k}E[R|\mathbf{X}=\boldsymbol{x}, A=l].$$

We continue this proof for the remaining classes and finally obtain Fisher consistency.

A.2 Proof of Theorem 2

We first note

$$
\begin{aligned}
& \mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}(\mathcal{D}^*) \\
= & \sum_{l=1}^{k} \left\{ E\left[\frac{R}{\pi_l(\mathbf{X})} I(A=l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l)\right] - E\left[\frac{R}{\pi_l(\mathbf{X})} I(A=l, \mathcal{D}^*(\mathbf{X}) \neq l)\right] \right\} \\
= & \sum_{l=1}^{k} \left\{ E\left[\frac{R}{\pi_l(\mathbf{X})} I(A=l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \right. \\
& \left. -E\left[\frac{R}{\pi_l(\mathbf{X})} I(A=l, \mathcal{D}^*(\mathbf{X}) \neq l, \widehat{\mathcal{D}}(\mathbf{X}) = l)\right] \right\} \\
= & \sum_{l=1}^{k} \left\{ E\left[\frac{R}{\pi_l(\mathbf{X})} I(A=l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \right. \\
& \left. -E\left[\frac{R}{\pi_A(\mathbf{X})} I(A\neq l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \right\}. \\
\leq & \sum_{l=1}^{k} \left\{ E\left[\frac{R^+}{\pi_A(\mathbf{X})} I(A=l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \right. \\
& \left. +E\left[\frac{R^-}{\pi_A(\mathbf{X})} I(A\neq l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \right\}.
\end{aligned}
$$

We let Δ_l to denote each term on the right-hand side of the above equation. That is,

$$
\begin{aligned}
\Delta_l &= E\left[\frac{R^+}{\pi_A(\mathbf{X})} I(A=l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] + E\left[\frac{R^-}{\pi_A(\mathbf{X})} I(A\neq l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right] \\
&= E\left[\frac{|R|}{\pi_A(\mathbf{X})} I(Z_l \text{sign}(R) = 1, \widehat{\mathcal{D}}(\mathbf{X}) \neq l, \mathcal{D}^*(\mathbf{X}) = l)\right],
\end{aligned}
$$

where we recall $Z_l = 2I(A=l) - 1$.

We first examine Δ_k. For any $\boldsymbol{x}$ in the domain of $\mathbf{X}$, we let $j_1, j_2, ..., j_{k-1}$ be the permutation of $\{1, ..., k-1\}$ such that

$$
E[R|A=j_1, \mathbf{X}=\boldsymbol{x}] < ... < E[R|A=j_{k-1}, \mathbf{X}=\boldsymbol{x}].
$$

Then according to SOM learning, $\mathcal{D}^*(\boldsymbol{x}) = k$ implies that $f^*_{j_l(\boldsymbol{x})}(\boldsymbol{x}) < 0$ for any $l = 1, .., k-1$, while $\widehat{\mathcal{D}}(\mathbf{X}) \neq k$ implies that for this particular permutation, there exists some $l = 1, ..., k-1$ such that $\widehat{f}_{j_l}(\boldsymbol{x}) > 0$ so $\widehat{f}_{j_l}(\boldsymbol{x}) f^*_{j_l}(\boldsymbol{x}) < 0$. Recall that $f^*_{j_l}(\boldsymbol{x}) = \eta_{j_l,S}$ with $S = \{j_{l+1}, ..., k\}$ and it is the limit of

$\widehat{f}_{j_l}$ from Theorem 1. Therefore, we obtain

$$\begin{aligned}
\Delta_k &\leq E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{\sum_{(j_1,\ldots,j_{k-1})} I(Z_k\text{sign}(R)=1, \text{there exists } l\leq k-1 \text{ such that } \widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X})<0)\right\}\right] \\
&\leq \sum_{(j_1,\ldots,j_{k-1})} E\left[\frac{|R|}{\pi_A(\mathbf{X})} I\left(Z_{j_1}\text{sign}(R)=-1,\ldots,Z_{j_{l-1}}\text{sign}(R)=-1,\widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X})<0\right)\right] \\
&\leq \sum_{(j_1,\ldots,j_{k-1})} E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{I(A=j_l)(k-l+1)+I(A\neq j_l)\right\}\right. \\
&\qquad \left.\times I\left(Z_{j_1}\text{sign}(R)=-1,\ldots,Z_{j_{l-1}}\text{sign}(R)=-1,\widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X})<0\right)\right].
\end{aligned}$$

Hence, it suffices to bound each term on the right-hand side of the above inequality.

When $l=1$, under conditions (C.1)-(C.4), we use the same proof of Theorem 3.2 in Zhao et al. (2012), which extends the result in Stienwart and Christmann (2008) to a weighted support vector machine. Particularly, in their proof, we let the weight for subject i be

$$|R_i|/\pi_{A_i}(\mathbf{X}_i)\left\{(k-1)I(A_i=j_1)+I(A_i\neq j_1)\right\}$$

and the class label be $Z_{j_1}\text{sign}(R_i)$. Furthermore, from the proof of Theorem 1, $f^*_{j_1}(\boldsymbol{x})$ has the same sign as $\eta_{j_1,\{j_2,\ldots,j_k\}}(\boldsymbol{x})$. Thus, from condition (C.1), we conclude that there exists at least probability $1-3e^{-\epsilon}$ and a constant C_1 such that it holds

$$E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-1)I(A=j_1)+I(A\neq j_1)\right\}I(Z_{j_1}\text{sign}(R)\widehat{f}_{j_1}(\mathbf{X})<0)\right]$$

$$-E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-1)I(A=j_1)+I(A\neq j_1)\right\}I(Z_{j_1}\text{sign}(R)f^*_{j_1}(\mathbf{X})<0)\right]\leq C_1Q_n(\epsilon),$$

where

$$Q_n(\epsilon)=\left\{\lambda_n^{\frac{\tau}{2+\tau}}\sigma_n^{-\frac{d\tau}{d+\tau}}+\sigma_n^{\beta}+\epsilon\left(n\lambda_n^p\sigma_n^{\frac{1-p}{1+\epsilon_0 d}}\right)^{-\frac{q+1}{q+2-p}}\right\}$$

with any constant $\epsilon_0>0$ and $d/(d+\tau)<p<2$. Then according to the proof of Lemma 5 in Barlette et al. (2006) and conditions (C.1) and (C.2), this gives

$$P(\widehat{f}_{j_1}(\mathbf{X})f^*_{j_1}(\mathbf{X})<0)\leq[C_1'Q_n(\epsilon)]^{\alpha},$$

where $\alpha = q/(1+q)$ and C_1' is a constant.

When $l = 2$, the step at j_2 in SOM is to minimize

$$n^{-1}\sum_{i=1}^{n} I(Z_{ij_1} = -1, Z_{ij_1}\text{sign}(R_i)\widehat{f}_{j_1}(\mathbf{X}_i) < 0) w_i (1 - Z_{ij_2}\text{sign}(R_i) f(\mathbf{X}_i))_+ + \lambda_{n,j_2}\|f\|^2,$$

where $w_i = |R_i|/\pi_{A_i}(\mathbf{X}_i)\left\{(k-2)I(A_i = j_2) + I(A_i \neq j_2)\right\}$. Thus, we can proceed the same proof of Theorem 3.2 in Zhao et al. (2012) except that only subjects in the random set

$$\left\{i : Z_{ij_1} = -1, Z_{ij_1}\text{sign}(R_i)\widehat{f}_{j_1}(\mathbf{X}_i) < 0\right\}$$

are used in the derivation. We obtain that

$$\begin{aligned}
& E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-2)I(A = j_2) + I(A \neq j_2)\right\} I(Z_{j_1} = -1, Z_{j_2}\text{sign}(R)\widehat{f}_{j_2}(\mathbf{X}) < 0)\right] \\
& \quad -E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-2)I(A = j_2) + I(A \neq j_2)\right\} I(Z_{j_1} = -1, Z_{j_2}\text{sign}(R) f^*_{j_2}(\mathbf{X}) < 0)\right] \\
\leq \; & C_2\left\{Q_n(\epsilon) + |P(Z_{j_1}\text{sign}(R)\widehat{f}_{j_1}(\mathbf{X}) > 0) - P(Z_{j_1}\text{sign}(R) f^*_{j_1}(\mathbf{X}) > 0)|\right\} \\
\leq \; & C_2\left\{Q_n(\epsilon) + Q_n(\epsilon)^\alpha\right\}
\end{aligned}$$

with a probability at least $1 - 3e^{-\epsilon}$ for a constant C_2. Note that the second term on the right-hand side is due to the estimated random set in this step. Again, the proof of Lemma 5 in Barlette et al. (2006) gives

$$P(Z_{j_1} = -1, \widehat{f}_{j_2}(\mathbf{X}) f^*_{j_2}(\mathbf{X}) < 0) \leq [C_2' Q_n(\epsilon)]^\alpha.$$

We continue the same arguments for $l = 3, ..., k-1$ to obtain

$$\begin{aligned}
& E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-l+1)I(A = j_l) + I(A \neq j_l)\right\} I\left\{Z_{j_l}\text{sign}(R)\widehat{f}_{j_l}(\mathbf{X}) < 0, Z_{j_{l-1}} = -1, ..., Z_{j_1} = -1\right\}\right] \\
& -E\left[\frac{|R|}{\pi_A(\mathbf{X})}\left\{(k-l+1)I(A = j_l) + I(A \neq j_l)\right\} I\left\{Z_{j_l} f^*_{j_l}(\mathbf{X}) < 0, Z_{j_{l-1}} = -1, ..., Z_{j_1} = -1\right\}\right] \\
& \leq C_l\left\{Q_n(\epsilon) + Q_n(\epsilon)^\alpha\right\}
\end{aligned}$$

with a probability at least $1 - 3le^{-\epsilon}$ for some constant C_l, and

$$P(Z_{j_1} = -1, ..., Z_{j_{l-1}} = -1, \widehat{f}_{j_l}(\mathbf{X}) f^*_{j_l}(\mathbf{X}) < 0) \leq [C'_l Q_n(\epsilon)]^\alpha$$

for a constant C'_l. Hence, with a probability $1 - [3k(k-1)/2]e^{-\epsilon}$, $\Delta_k \leq CQ_n(\epsilon)^\alpha$ for a constant C.

Similarly, we can examine the difference for Δ_{k-1}. We follow exactly the same arguments as before by considering all possible permutations from $\{1, ..., k-2\}$ and $l = 1, ..., k-2$. The only difference in the argument is that the random set is restricted to subjects with $A \neq k$ and $\widehat{\mathcal{D}}^{(k)}(\mathbf{X}) = -1$. However, the probability of the latter differs from the probability $A \neq k$ and $\mathcal{D}^{*(k)}(\mathbf{X}) = -1$ by $CQ_n(\epsilon)^\alpha$ from the previous conclusion. Therefore, we obtain that with probability at least $1 - [3k(k-1)/2 + 3(k-1)(k-2)/2]e^{-\epsilon}$, $\Delta_{k-1} \leq CQ_n(\epsilon)^\alpha$ for another constant C. Continue the same arguments for $\Delta_l, l = k-2, ..., 1$ so we finally conclude

$$\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}^* \leq CQ_n(\epsilon)^\alpha$$

with probability at least $1 - C'e^{-\epsilon}$ where C' is a constant depending on k. Thus Theorem 2 holds.

References

Allwein, E. L., Schapire, R. E., and Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, 1:113–141, 2001.

Carini, C., Menon, S. M., and Chang, M. *Clinical and Statistical Considerations in Personalized Medicine*. CRC Press, 2014.

Dietterich, T. G. and Bakiri, G. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, pages 263–286, 1995.

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

Klein, D. N., Leon, A. C., Li, C., D'Zurilla, T. J., Black, S. R., Vivian, D., Dowling, F., Arnow, B. A., Manber, R., Markowitz, J. C., et al. Social problem solving and depressive symptoms over time: A randomized clinical trial of cognitive-behavioral analysis system of psychotherapy, brief supportive

psychotherapy, and pharmacotherapy. *Journal of Consulting and Clinical Psychology*, 79(3):342, 2011.

Kocsis, J. H., Gelenberg, A. J., Rothbaum, B. O., Klein, D. N., Trivedi, M. H., Manber, R., Keller, M. B., Leon, A. C., Wisniewski, S. R., Arnow, B. A., et al. Cognitive behavioral analysis system of psychotherapy and brief supportive psychotherapy for augmentation of antidepressant nonresponse in chronic depression: the revamp trial. *Archives of General Psychiatry*, 66(11):1178–1188, 2009.

Kosorok, M. R. and Moodie, E. E. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, volume 21. SIAM, 2015.

Kreßel, U. H.-G. Pairwise classification and support vector machines. In *Advances in Kernel Methods*, pages 255–268. MIT Press, 1999.

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

Lipska, K. J. and Krumholz, H. M. Comparing diabetes medications: where do we set the bar? *JAMA Internal Medicine*, 174(3):317–318, 2014.

Liu, Y. and Shen, X. Multicategory ψ-learning. *Journal of the American Statistical Association*, 101 (474):500–509, 2006.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. Robust hybrid learning for estimating personalized dynamic treatment regimens. *arXiv preprint arXiv:1611.02314*, 2014.

Moodie, E. E., Richardson, T. S., and Stephens, D. A. Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455, 2007.

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

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

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

Robins, J. M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

Trivedi Madhukar, H. et al. Treatment strategies to improve and sustain remission in major depressive disorder. *Dialogues in Clinical Neuroscience*, 10(4):377–384, 2008.

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

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

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3), 2013.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

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