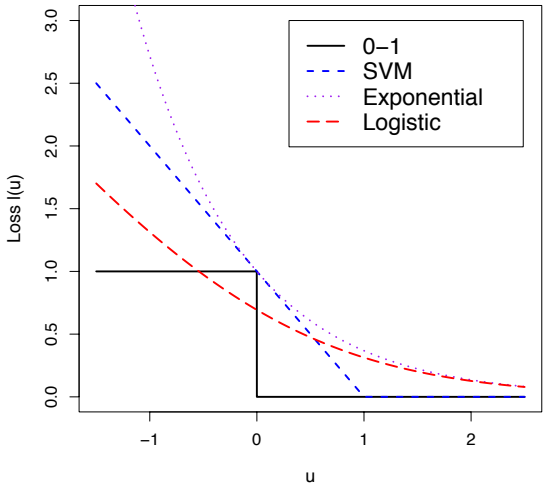
Supplementary Materials for "Multicategory Outcome Weighted Margin-based Learning for Estimating Individualized Treatment Rules"

Chong Zhang[1], Jingxiang Chen[1], Haoda Fu[2], Xuanyao He[2], Ying-Qi Zhao[3], and Yufeng Liu[1]

[1] *University of North Carolina at Chapel Hill,* [2] *Eli Lilly and Company,*

and [3] *Fred Hutchinson Cancer Research Center*

A1. Plot of Common Large Margin Loss Functions



A2. Additional Discussions on Soft and Hard Classifiers

Following the discussion at the end of Section 2 of the main paper, we would like continue to illustrate the discussions about the performance comparison between soft and hard classifiers. To explore the difference between soft and hard MOML classifiers, we use a toy example in Figure S1 for demonstration. In particular, we let

$k = 3$, and plot the log of the reward ratio $\log\{R(\boldsymbol{x}, 1)/R(\boldsymbol{x}, 2)\}$ (denoted by r_{12} in Figure S1) against $\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle$ and $\langle \boldsymbol{f}^*, \boldsymbol{W}_2 \rangle$ ($\boldsymbol{f}^*$ indicates the underlying optimal classifier, and $\boldsymbol{W}_j$ indicates the jth vertex of the simplex). One can see that for $c = 0$ (the soft classifier), MOML can provide estimation of $R(\boldsymbol{x}, 1)/R(\boldsymbol{x}, 2)$ for $\{\boldsymbol{f}^* : \langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle > 0 \text{ or } \langle \boldsymbol{f}^*, \boldsymbol{W}_2 \rangle > 0\}$. As c increases, the flat region enlarges and the function gets closer to a step function. Consequently, the ratio estimation becomes more difficult. In the limit when $c \to \infty$, the hard classifier in MOML provides little information about the rewards ratio. This is similar to the binary case as discussed in Section 2.2 of the main paper.

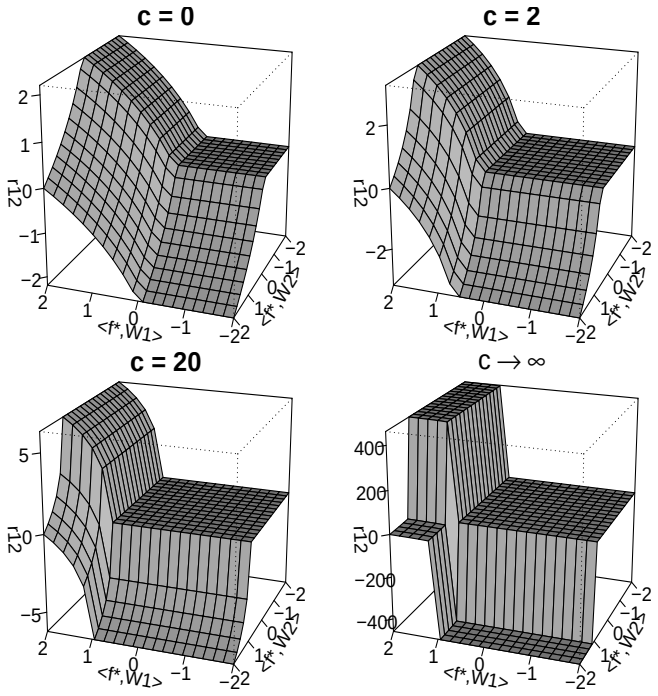


Figure S1: Plot of $\log\{R(\boldsymbol{x}, 1)/R(\boldsymbol{x}, 2)\}$ (r_{12} on the y axis) against $\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle$ and $\langle \boldsymbol{f}^*, \boldsymbol{W}_2 \rangle$ for some LUM loss functions. Here $c = 0$ corresponds to the soft LUM loss, and $c \to \infty$ corresponds to the SVM hinge loss, which is a hard classifier. We fix $a = 1$ as in the binary case (see Figure 1).

In Section 2.1 of the main paper, we claimed that soft classifiers can show better performance than hard classifiers when the underlying optimal treatment probability

ratios are relatively smooth functions of the covariates. Otherwise, hard classifiers may outperform soft classifiers. To show such differences, we use Example 1 and Example 6 from the numerical section of the main paper as two representatives of the smooth and non-smooth ratios. We repeat the settings of the two examples as below:

Example 1 We consider three points $(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3)$ of equal distances from the p-dimensional space to represent the cluster centroids of the true optimal treatments. For each $\boldsymbol{c}_j$ where $j = 1, 2, 3$, we generate its covariate X_i from a multivariate normal distribution $N(\boldsymbol{c}_j, I_p)$ where I_p is a p-dimensional identity matrix. The actually assigned A_i follows a discrete uniform distribution $U\{1, 2, 3\}$. The reward R_i follows a Gaussian distribution $N(\mu(X_i, A_i, d_i), 1)$, where the $\mu(X_i, A_i, d_i) = X_i^T\boldsymbol{\beta} + 5 \cdot I(A_i = d_i)$, $\boldsymbol{\beta}^T = (\mathbf{1}_{p/2}^T, -\mathbf{1}_{p/2}^T)$ and d_i is the optimal treatment for X_i determined by the cluster centroids. The training dataset is of size 300.

Example 6 In this example, the optimal treatment d_i for each X_i is determined with probability 95% by the signs of two underlying non-linear functions $f_1(X) = X_1^2 + X_2^2 + \exp\{0.5X_3\}$ and $f_2(X) = X_4^2 - X_5^3 - X_6$ while a random noise is added to d_i with probability 5% to create a positive Bayes error. In particular, we have d_i defined as

$$d_i = d(X_i) = \begin{cases} 1 + [\text{sign}(f_1(X_i) - m_1)]_+ + 2 \times [\text{sign}(f_2(X_i) - m_2)]_+ & \text{w/t prob. } 0.95 \\ U_i & \text{w/t prob. } 0.05 \end{cases},$$

where m_1 and m_2 are the medians of f_1 and f_2 respectively, and U_i follows a discrete $U\{1, 2, 3, 4\}$ which is independent of (A_i, X_i). The covariates X_i follows a continuous uniform distribution $U(0, 1)$, $A_i \sim U\{1, \cdots, 4\}$, and $R_i \sim N(\mu(X_i, A_i, d_i), 1)$, where

$\mu(X_i, A_i, d_i) = X_i^T\boldsymbol{\beta} + 5 \cdot I(A_i = d_i) - 1$ and $\boldsymbol{\beta}^T = (\mathbf{1}_{p/2}^T, -\mathbf{1}_{p/2}^T)$. The training dataset is of size 500.

We draw the predicted optimal treatment for test datasets under different values of c. Note that the classifier of MOML becomes soft when $c \to 0$ and hard when $c \to \infty$. Thus, we pick $c \in \{0, 10, 1000\}$, and note that $c = 1000$ leads to an approximately hard classifier. In addition, we only include the first two covariates in Example 1 to better visualize the observations in a 2-D graph. The underlying true optimal treatments and predicted optimal treatments in one realization are displayed in Figure S2 (Example 1) and Figure S3 (Example 6). The dashed lines indicate the underlying boundaries that are determined by the Bayes classifiers.

By Figure S2, when the underlying optimal treatment probability ratios are smooth functions of covariates, the predicted treatment results become worse as c goes up in the sense that the estimated boundaries move away from the Bayes classifier. In particular, the misclassification rate increases from 6.2% when $c = 0$ to 10.5% when $c = 1000$. In contrast, Figure S3 shows that when the underlying optimal treatment probability ratios are non-smooth functions, the predicted treatment accuracy becomes slightly better as c becomes larger, i.e. the misclassification rate drops from 23.5% at $c = 0$ to 22.9% at $c = 1000$. This is consistent to our previous conclusion that hard classifiers can perform better than soft ones when it is difficult to fit the probability ratios.

A3. Additional Simulation Results for the Main Paper

Tables S1 and S2 report the sample means and standard deviations of the misclassification rates and the empirical value functions produced by all the models for the

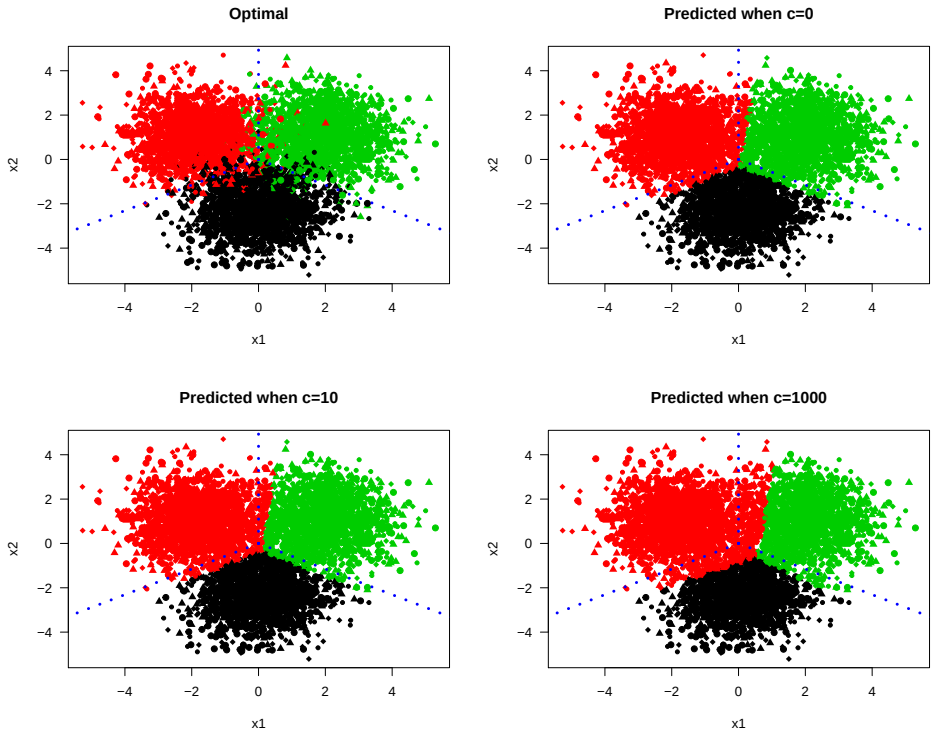


Figure S2: Prediction performance of soft classifiers ($c = 0$) and an approximately hard classifier ($c = 1000$) in one realization of Example 1 in two-dimension when the underlying optimal treatment probability ratios are smooth functions. The dashed lines indicate the true underlying decision boundary. The point symbols show the actually assigned treatments and point colors show the predicted ITRs for testing samples. The corresponding misclassification rates are 6.2%, 6.8% and 10.5%, for $c = 0$, $c = 10$ and $c = 1000$.

simulation examples in the main paper.

A4. Additional Statistical Learning Theory

In this section, we explore some additional theoretical properties of our proposed MOML. We begin by showing that the one-versus-rest SVM approach in ITR problems can be inconsistent in Section A4.1. Then in Section A4.2, we demonstrate that under certain conditions, MOML can enjoy selection consistency in linear learning. Asymptotic convergence rates for the excess risks of MOML are obtained in Section A4.3.

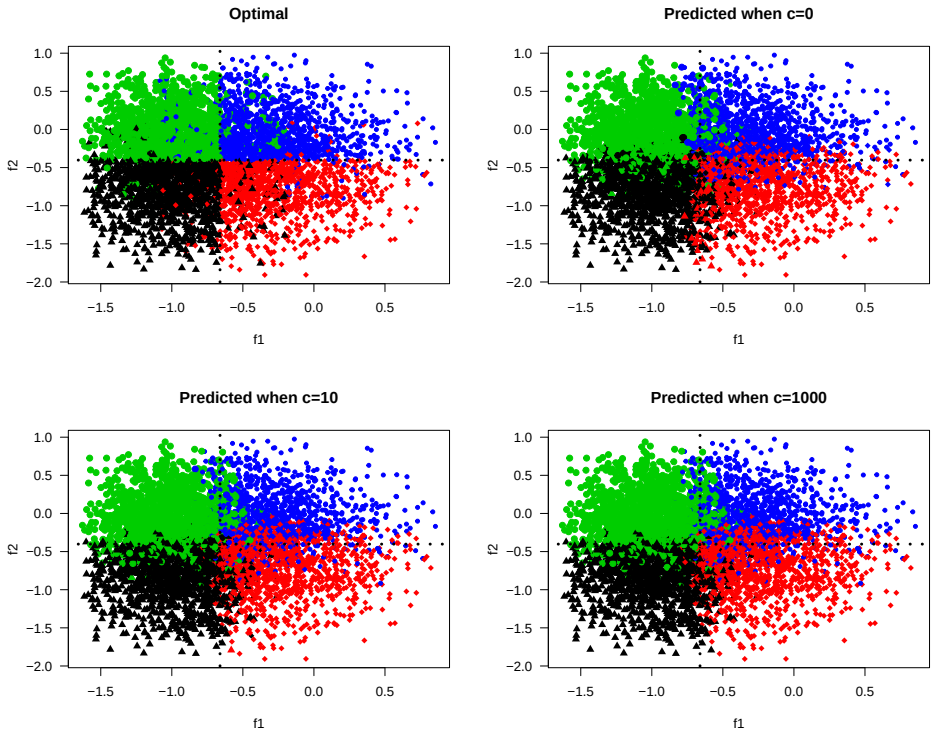


Figure S3: Prediction performance of soft classifiers ($c = 0$) and an approximately hard classifier ($c = 1000$) in one realization of Example 6 when the underlying optimal treatment probability ratios are non-smooth functions. The dashed lines indicate the true underlying decision boundary. The point symbols show the actually assigned treatments and point colors show the predicted ITRs for testing samples. The corresponding misclassification rates are 23.5%, 23.4% and 22.9%, for $c = 0$, $c = 10$ and $c = 1000$.

A4.1 Fisher Inconsistency of the One-versus-rest SVM Approach

In Zhao et al. (2012), it was shown that the binary OWL approach using the SVM hinge loss is Fisher consistent. However, its direct generalization to the multicategory framework can be more involved, and the Fisher consistency cannot be always guaranteed. The next proposition shows that if one generalizes a binary margin-based classifier to handle multiple treatments using the one-versus-rest approach, the classifier can be inconsistent. Consequently, the corresponding methods can be suboptimal for real applications. Therefore, it is desirable to consider multiple treatments in one

Dimension			$p = 10$					$p = 50$		
Method	OWL-1	OWL-2	MOML	MOML-l_1	*Bayes*	OWL-1	OWL-2	MOML	MOML-l_1	*Bayes*
Example 1	0.083	0.127	**0.081**	0.086	*0.068*	0.139	0.173	**0.136**	0.174	*0.067*
	(0.007)	(0.043)	(0.013)	(0.015)	*(0.004)*	(0.054)	(0.069)	(0.031)	(0.049)	*(0.004)*
Example 2	0.185	0.252	**0.150**	0.151	*0.098*	0.369	0.362	0.224	**0.220**	*0.097*
	(0.065)	(0.036)	(0.016)	(0.017)	*(0.005)*	(0.081)	(0.078)	(0.024)	(0.022)	*(0.004)*
Example 3	0.508	0.669	**0.318**	0.326	*0.074*	0.636	0.844	0.431	**0.419**	*0.074*
	(0.055)	(0.048)	(0.050)	(0.072)	*(0.004)*	(0.049)	(0.062)	(0.046)	(0.047)	*(0.003)*
Example 4	0.197	0.257	0.154	**0.149**	*0.098*	0.378	0.386	0.237	**0.216**	*0.097*
	(0.094)	(0.037)	(0.018)	(0.013)	*(0.005)*	(0.127)	(0.085)	(0.026)	(0.019)	*(0.004)*
Example 5	0.254	0.333	**0.148**	-	*0.076*	0.355	0.387	**0.291**	-	*0.077*
	(0.063)	(0.061)	(0.016)	-	*(0.005)*	(0.040)	(0.071)	(0.024)	-	*(0.005)*
Example 6	0.276	0.397	**0.220**	-	*0.038*	0.384	0.487	**0.323**	-	*0.038*
	(0.031)	(0.063)	(0.020)	-	*(0.003)*	(0.017)	(0.021)	(0.019)	-	*(0.003)*

Table S1: Misclassification results of simulation studies: means and standard deviations (in parenthesis) of the misclassification rates. OWL-1 and OWL-2 represent the two extensions of outcome weighted learning (one-versus-rest and one-versus-one), MOML and MOML-l_1 represent the outcome weighted margin-based learning with l_2 and l_1 penalties respectively, and Bayes represents the empirical Bayes error. In each scenario, the model producing the best criterion is in bold.

optimization problem.

Proposition S1. *Suppose for a given* $\boldsymbol{x}$*, we have* $R(\boldsymbol{x}, j) < \sum_{i \neq j} R(\boldsymbol{x}, i)$ *for all* $j \in \{1, \ldots, k\}$*. Then for finding optimal ITRs using the one-versus-rest approach with a binary Fisher consistent loss function* $\ell(\cdot)$*, the corresponding method is not Fisher consistent.*

A4.2 Selection Consistency

In the statistical learning literature, selection consistency of regression methods has been well established. See, for example, Zhao and Yu (2006), Zou (2006), Fan and Lv (2010), and the references therein. In contrast, selection consistency of classification methods has received much less attention. Recently, Zhang et al. (2014) studied the selection consistency of SVMs for standard classification problems, and Song et al. (2015) studied the selection consistency of ITRs for binary treatments. In the literature of ITRs for multicategory treatments, to our knowledge, no work has been

Dimension			$p = 10$					$p = 50$		
Method	OWL-1	OWL-2	MOML	MOML-l_1	*Optimal*	OWL-1	OWL-2	MOML	MOML-l_1	*Optimal*
Example 1	4.905	4.525	**4.932**	4.878	*5.319*	4.599	4.432	**4.672**	4.315	*5.304*
	(0.062)	(0.173)	(0.083)	(0.086)	*(0.159)*	(0.268)	(0.375)	(0.175)	(0.238)	*(0.208)*
Example 2	2.989	2.657	**3.165**	3.164	*3.919*	2.061	2.231	2.806	**2.809**	*3.916*
	(0.241)	(0.189)	(0.086)	(0.093)	*(0.016)*	(0.307)	(0.390)	(0.127)	(0.129)	*(0.025)*
Example 3	2.466	2.164	**2.701**	2.644	*2.900*	2.201	2.037	2.572	**2.597**	*2.897*
	(0.095)	(0.156)	(0.054)	(0.057)	*(0.008)*	(0.136)	(0.272)	(0.065)	(0.061)	*(0.015)*
Example 4	2.951	2.737	3.151	**3.251**	*4.001*	2.034	2.063	2.797	**2.897**	*3.976*
	(0.277)	(0.189)	(0.090)	(0.070)	*(0.004)*	(0.402)	(0.421)	(0.156)	(0.086)	*(0.006)*
Example 5	1.730	1.811	**2.273**	-	*2.997*	1.289	1.291	**1.848**	-	*3.025*
	(0.213)	(0.318)	(0.103)	-	*(0.111)*	(0.353)	(0.377)	(0.257)	-	*(0.245)*
Example 6	2.744	2.316	**3.168**	-	*3.989*	2.651	1.864	**2.750**	-	*4.019*
	(0.171)	(0.191)	(0.117)	-	*(0.038)*	(0.223)	(0.224)	(0.228)	-	*(0.105)*

Table S2: Value functions results of simulation studies: means and standard deviations (in parenthesis) of the estimated value functions. OWL-1 and OWL-2 represent the two extensions of outcome weighted learning (one-versus-rest and one-versus-one), MOML and MOML-l_1 represent the outcome weighted margin-based learning with l_2 and l_1 penalties respectively, and Optimal represents the optimal value function. In each scenario, the model producing the best criterion is in bold.

done on the selection consistency for existing learning methods. In this section, we focus on linear learning with the l_1 penalty, and explore the selection consistency of MOML. We show that if the number of observations n and the number of covariates p grow simultaneously, in a way such that $\log(p)^2/n \to 0$, MOML can enjoy asymptotic selection consistency under certain conditions.

To begin with, we need to introduce some further notations. Recall that $\boldsymbol{f} = (f_1, \ldots, f_{k-1})^T$ for an ITR problem with k treatments. In linear learning, we let $f_j(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_j + \beta_{j,0};\ j = 1, \ldots, k-1$, where $\boldsymbol{\beta}_j$ is the coefficient vector for the jth classification function, and $\beta_{j,0}$ is the corresponding intercept. Denote by $\beta_{j,q};\ q \geq 1$, the qth element of $\boldsymbol{\beta}_j$. For brevity, we let $\beta_{\cdot,q}$ represent an indicator for the $k-1$ parameters that correspond to the qth covariate, and define $\beta_{\cdot,q} = 0$ if $\beta_{j,q} = 0$ for all $j \geq 1$, and $\beta_{\cdot,q} = 1$ if $\beta_{j,q} \neq 0$ for some $j \geq 1$.

Let $\boldsymbol{f}_0$ be the underlying function that minimizes the expected loss of (2.9). In

other words, for linear learning, let

$$\boldsymbol{f}_0 = \underset{\boldsymbol{f}}{\operatorname{argmin}} E\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle \boldsymbol{W}_A, \boldsymbol{f}(\boldsymbol{X})\rangle\},$$

where the expectation is taken with respect to the underlying distribution P. For $\boldsymbol{f}_0$, let the corresponding parameters be $\boldsymbol{\beta}_j^*$ and $\beta_{j,0}^*$ for all j. Similar to $\beta_{\cdot,q}$, we define $\beta_{\cdot,q}^* = 0$ if $\beta_{j,q}^* = 0$ for all $j \geq 1$, and $\beta_{\cdot,q}^* = 1$ if $\beta_{j,q}^* \neq 0$ for some $j \geq 1$. When the learning signal is sparse, many $\beta_{j,q}^*$'s are zero. In other words, the coefficient vectors $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{k-1}^*)$ are parsimonious. Note that the qth covariate does not contain useful information for finding the optimal ITRs if and only if $\beta_{\cdot,q} = 0$. Hence, we define the set of important covariates to be $\boldsymbol{X}_1 = \{x_q : \ \beta_{\cdot,q}^* = 1\}$ and its complement to be $\boldsymbol{X}_0 = \{x_q : \ \beta_{\cdot,q}^* = 0\}$. In this paper, we focus on the case where the number of important covariates is a fixed number, i.e., $|\boldsymbol{X}_1| < \infty$. Note that $\hat{\beta}_{\cdot,q}$ and $\beta_{\cdot,q}^*$ are the estimated and underlying true indicators that whether the qth covariate is useful. By this definition, selection consistency of the OML method means that as the sample size n increases, the probability of $\hat{\beta}_{\cdot,q} = \beta_{\cdot,q}^*$ for all q tends to 1. As the dimension p may become unbounded, we assume that the underlying distribution P is defined on $([0,1]^\infty \times \{1, \ldots, k\}, \sigma^\infty([0,1]^\infty) \times 2^{\{1,\ldots,k\}})$ with $\sigma^\infty([0,1]^\infty)$ being the σ-field generated by the open balls introduced by the uniform metric $d(\boldsymbol{x}, \boldsymbol{x}') = \sup_j |x_j - x_j'|$, where x_j is the jth element of $\boldsymbol{x}$.

Next, we introduce some regularity conditions for selection consistency. The first condition requires that the marginal distribution of the covariates is bounded in $[0,1]^p$, where p is the number of covariates. One can verify that our theorem can be generalized to the case where the covariates are bounded (not necessarily in $[0,1]^p$).

Condition 1 (C1). Every element in $\boldsymbol{X}$ ranges in $[0,1]$, and the corresponding

distribution is absolutely continuous with respect to the Lebesgue measure.

We would like to point out that the second assumption in C1 can be removed if the loss function ℓ has the second order derivative everywhere. See the discussions after C3 for more details.

The next condition requires that the marginal distribution of the clinical rewards for all patients and treatments does not have a heavy tail. Recall that if a random variable U is sub-Gaussian with parameter s, then we have $\text{pr}(|U| > u) \leq 2\exp(-u^2/s)$ for large enough u.

Condition 2 (C2). The marginal distribution of $(R \mid \boldsymbol{X} = \boldsymbol{x}, A = a)$ is sub-Gaussian with a universal parameter $s < \infty$ for any $\boldsymbol{x}$ and a.

C2 is very general, and many commonly seen distributions are sub-Gaussian. For example, normal random variables are known to be sub-Gaussian, and random variables with bounded ranges or small kurtosis are also sub-Gaussian. C2 excludes the possibility that a patient's reward is significantly different from its expectation, which can lead to biased ITRs. We note that C2 is sufficient for selection consistency. Hence, if for some $\boldsymbol{x}$ and a, the marginal distribution of the reward is not sub-Gaussian, selection consistency may still be established. See the proof of Theorem S1 for more discussions.

The next condition requires that the loss function ℓ is differentiable and convex, and its second order derivative function is bounded.

Condition 3 (C3). The loss function $\ell(u)$ is differentiable, and has a second order derivative almost everywhere with respect to the Lebesgue measure, where $\ell''(u) < \infty$.

C3 is valid for many commonly used loss functions, such as the logistic deviance loss and the LUM loss family with $c < \infty$. The SVM hinge loss is not differentiable and thus our theorem does not apply. Note that our focus is to prove selection

consistency for general margin-based loss functions using mild conditions. To prove selection consistency for SVMs, one may need finer analysis. For example, Zhang et al. (2014) gave the conditions to establish selection consistency for binary SVMs.

Note that in C1, we require that the marginal distribution of the covariates is absolutely continuous with respect to the Lebesgue measure. This condition, together with C3, ensures that there is no probability mass at which the loss function is not second order differentiable. If one uses a loss function that is second order differentiable everywhere, such as the logistic loss, the corresponding requirement in C1 can be dropped and the selection consistency is still valid.

We are ready to present our main theorem in this section.

Theorem S1. *Suppose Conditions C1-C3 hold, and* $\log(p)/(n^{1/2}) \to 0$ *as* $n, p \to \infty$. *If we choose* $\lambda = O_P\{\log(p)^{1/2}/(n^{1/4})\}$, *we have that the corresponding solution* $\hat{\boldsymbol{f}}$ *to (2.9) satisfies that, with probability tending to 1,* $\hat{\beta}_{\cdot,q} = \beta^*_{\cdot,q}$ *for all* $q = 1, \ldots, p$.

From Theorem S1, one can verify that the MOML method can select covariates consistently for problems where p is of any polynomial form of n. This can help find important covariates on which different treatments have significant effects. In the next section, we show that the excess ℓ risk converges to zero at a fast rate, under certain conditions.

A4.3 Convergence Rate of Excess Risks

In this section, we study the convergence rate of the excess ℓ risk for MOML using linear learning. In particular, we first extend the notation of the excess ℓ risk from standard binary margin-based classification to the ITRs framework using multicategory classifiers, and show that the convergence rate of MOML can be fast under

Conditions C1-C3.

Let $e_\ell(\boldsymbol{f}, \boldsymbol{f}_0) = E\{|R|\ell_R(\langle \boldsymbol{W}_A, \boldsymbol{f}\rangle)/\pi_A(\boldsymbol{X})\} - E\{|R|\ell_R(\langle \boldsymbol{W}_A, \boldsymbol{f}_0\rangle)/\pi_A(\boldsymbol{X})\}$. We call $e_\ell(\boldsymbol{f}, \boldsymbol{f}_0)$ the excess ℓ risk. One can verify that this definition is a natural generalization of the excess ℓ risk used in Bartlett et al. (2006) for binary classifiers. For MOML we have the following result.

Theorem S2. *Suppose Conditions C1-C3 hold, and* $\log(p)/(n^{1/2}) \to 0$ *as* $n, p \to \infty$. *If we choose* $\lambda = O_P\{\log(p)^{1/2}/(n^{1/4})\}$, *then* $e_l(\hat{\boldsymbol{f}}, \boldsymbol{f}_0)$ *converges to 0 at the rate* $O_P\{\log(p)/(n^{1/2})\}$.

By Theorem S2, the excess ℓ risk of MOML converges to zero at a desirable rate, under mild conditions. Consequently, the fitted $\hat{\boldsymbol{f}}$ can enjoy a good prediction performance.

A5. Technical Proofs of theorems

Proof of Proposition 1: This proposition is a special case of Theorem 2. ■

Large-margin Unified Loss Function: The Large-margin Unified Machines (LUM) use loss functions

$$\ell(u) = \begin{cases} 1 - u, & u < c/(1+c), \\ [a/\{(1+c)u - c + a\}]^a/(1+c), & u \geq c/(1+c), \end{cases}$$

where $c \geq 0$ and $a > 0$ are parameters of the LUM family. Note that $a = 1$, $c = 1$ corresponds to the distance discriminant analysis (Marron et al., 2007), and $c \to \infty$ corresponds to the SVM hinge loss. ■

Estimation of Class Conditional Probabilities in Standard Margin-based Classification: For a classification problem with k classes, we let $\boldsymbol{X}$ be the covariate

vector, and let Y be the corresponding label. It is common to assume that $\boldsymbol{X}$ and Y follow an unknown joint distribution. Define the class conditional probabilities as $P_j(\boldsymbol{x}) = \text{pr}(Y = j \mid \boldsymbol{X} = \boldsymbol{x});\ j = 1, \ldots, k$, where the probability is taken with respect to the joint distribution of $(\boldsymbol{X}, Y)$. In other words, the vector of class conditional probabilities is the marginal probability vector of a multinomial distribution that depends on $\boldsymbol{x}$.

For margin-based classifier, we can define the theoretical minimizer $S(\boldsymbol{x})$ in an analogous manner as in Section 2.2 of the original paper. For some loss functions, one can prove that there exist functions $g_j(\cdot)$, such that $P_j(\boldsymbol{x}) = g_j\{\boldsymbol{f}^*(\boldsymbol{x})\}$. Hence, it is common to use $\hat{\boldsymbol{f}}$ to estimate $\boldsymbol{f}^*$, and the corresponding estimation for P_j is $\hat{P}_j(\boldsymbol{x}) = g_j\{\hat{\boldsymbol{f}}(\boldsymbol{x})\}$. See Zhang et al. (2013) and Zhang and Liu (2014), among others, for details on how to estimate P_j using $\hat{\boldsymbol{f}}$. ■

Proof of Theorem 1: The proof is contained in the proof of Theorem 3 and is omitted. ■

Proof of Theorem 2: Define $R^+_{+1}(\boldsymbol{x}) = \int (R \mid \boldsymbol{X} = \boldsymbol{x}, A = +1) I(R > 0) dP$ and $R^-_{+1}(\boldsymbol{x}) = \int (R \mid \boldsymbol{X} = \boldsymbol{x}, A = +1) I(R < 0) dP$, as in the main paper. One can verify that $R^+_{+1}(\boldsymbol{x}) + R^-_{+1}(\boldsymbol{x}) = R(\boldsymbol{x}, +1)$. In the proof we drop the dependence of R^+_{+1} and R^-_{+1} on $\boldsymbol{x}$ when there is no confusion. Define R^+_{-1} and R^-_{-1} in an analogous manner. We have that, minimizing the conditional loss $S(\boldsymbol{x})$ is equivalent to

$$\min_f (R^+_{+1} - R^-_{-1})\ell(f) + (-R^-_{+1} + R^+_{-1})\ell(-f). \tag{A5.1}$$

Next, assume that treatment $+1$ is better, in the sense that $R^+_{+1} + R^-_{+1} > R^+_{-1} + R^-_{-1}$. It suffices to show that the minimizer of (A5.1), f^*, is such that $f^* > 0$. Based on the assumption of Theorem 2, one can conclude that $f^* \geq 0$. Because if this is not true,

then let $f^{**} = -f^*$, and one can verify (through some calculation) that the objective function value of f^* is larger than that of f^{**}, which is a contradiction with respect to the definition of f^*. Therefore, we only need to prove that $f^* \neq 0$.

To this end, we only need to show that the objective function value when $f^* = \delta$ is smaller than when $f^* = 0$, where δ is a small positive number. To verify this, observe that

$$\begin{aligned} S(\boldsymbol{x}) \mid_{f^*=0} - S(\boldsymbol{x}) \mid_{f^*=\delta} &= (R^+_{+1} - R^-_{-1})\ell'(0)(-\delta) + (-R^-_{+1} + R^+_{-1})\ell'(0)\delta \\ &= \ell'(0)\delta(-R^-_{+1} + R^+_{-1} - R^+_{+1} + R^-_{-1}) > 0. \end{aligned}$$

Hence OML is Fisher consistent in the binary setting. ■

Proof of Theorem 3: Define R^+_{+1}, R^-_{+1}, R^+_{-1}, and R^-_{-1} as in the proof of Theorem 2. In (A5.1), take derivative with respect to f, and we have that

$$\frac{R^+_{+1} - R^-_{-1}}{-R^-_{+1} + R^+_{-1}} = \frac{\ell'(-f^*)}{\ell'(f^*)}. \tag{A5.2}$$

When $R > 0$, $R^-_{+1} = R^-_{-1} = 0$, and we have proved Theorem 1. When R can be negative, $R^+_{+1} + R^-_{+1} > 0$, and $R^+_{-1} + R^-_{-1} > 0$, one can verify, after some calculation, that

$$\frac{R^+_{+1} + R^-_{+1}}{R^+_{-1} + R^-_{-1}} - \frac{R^+_{+1} - R^-_{-1}}{-R^-_{+1} + R^+_{-1}} \begin{cases} < 0, & \text{if } R^+_{+1} + R^-_{+1} > R^+_{-1} + R^-_{-1} \\ > 0, & \text{if } R^+_{+1} + R^-_{+1} < R^+_{-1} + R^-_{-1}. \end{cases}$$

This completes the proof.

Note that for the relationship between f^* and $\{R(\boldsymbol{x}, +1), R(\boldsymbol{x}, +1)\}$, we only have (A5.2). Hence, one cannot directly estimate the rewards ratio without further

assumption, such as $R > 0$. ■

Proof of Theorem 4: First, we recall the definition of R_j^+ and R_j^- from the main paper. For multicategory problems, one can verify that minimizing the conditional expected loss $S(\boldsymbol{x})$ is equivalent to

$$\min_{\boldsymbol{f}} \sum_{j=1}^{k} \{R_j^+ \ell(\langle \boldsymbol{f}, \boldsymbol{W}_j \rangle) - R_j^- \ell(-\langle \boldsymbol{f}, \boldsymbol{W}_j \rangle)\}. \tag{A5.3}$$

Without loss of generality, assume that treatment 1 is the best, in the sense that $R_1^+ + R_1^- > R_j^+ + R_j^-$ for any $2 \leq j \leq k$. By Assumption 1, we have that $R_1^- > R_j^-$. Next, we prove that $\boldsymbol{f}^*$ is such that $\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle > \langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle$ for any $j \geq 2$.

We prove by contradiction. Suppose $\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle > \langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle$ is not true for one specific j. Then we can find $\boldsymbol{f}^{**}$, such that $\langle \boldsymbol{f}^*, \boldsymbol{W}_q \rangle = \langle \boldsymbol{f}^{**}, \boldsymbol{W}_q \rangle$ for $q \neq 1, j$, $\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle = \langle \boldsymbol{f}^{**}, \boldsymbol{W}_1 \rangle - \delta$, and $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle = \langle \boldsymbol{f}^{**}, \boldsymbol{W}_j \rangle + \delta$ (see Lemma 1 in Zhang and Liu, 2014), where $\delta > 0$ is a small positive number. Denote by $S(\boldsymbol{f}^*)$ the conditional loss with respect to $\boldsymbol{f}^*$, and define $S(\boldsymbol{f}^{**})$ in an analogous manner. One can verify that, after some calculation,

$$\begin{aligned} S(\boldsymbol{f}^*) - S(\boldsymbol{f}^{**}) = \delta \{ & [R_1^+ |\ell'(\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle)| + R_1^- |\ell'(-\langle \boldsymbol{f}^*, \boldsymbol{W}_1 \rangle)|] \\ & - [R_j^+ |\ell'(\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle)| + R_j^- |\ell'(-\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle)|] \} \end{aligned}$$

By assumptions in Theorem 4, we have that ℓ is convex, and consequently $|\ell'(u_1)| \leq |\ell'(u_2)|$ if $u_2 \leq u_1$. Hence, choose δ sufficiently small, and one can verify that, $S(\boldsymbol{f}^*) - S(\boldsymbol{f}^{**}) > 0$ because $R_1^+ + R_1^- > R_j^+ + R_j^-$, $0 \geq R_1^- > R_j^-$ and $\delta > 0$. Therefore, the conditional loss for $\boldsymbol{f}^{**}$ is smaller than that of $\boldsymbol{f}^*$, which contradicts with the definition of $\boldsymbol{f}^*$. This completes the proof of the first part.

The proof of inconsistency for MOML-SVM uses similar techniques. In particular, we show that even when $R > 0$, the MOML-SVM is not consistent. First, we can show that if $R(\boldsymbol{x}, i) \geq R(\boldsymbol{x}, j)$ for $i \neq j$, then $\langle \boldsymbol{f}^*, \boldsymbol{W}_i \rangle \geq \langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle$. Next, one can verify that if $R(\boldsymbol{x}, i) = \min_{i=1,\ldots,k} R(\boldsymbol{x}, i)$ and is unique, then $\langle \boldsymbol{f}^*, \boldsymbol{W}_i \rangle = -(k-1)$, and $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle = 1$ for $j \neq i$. This is because if $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle < 1$ for some j, one can modify $\boldsymbol{f}^*$ such that $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle$ increases to 1, while $\langle \boldsymbol{f}^*, \boldsymbol{W}_i \rangle$ decreases by the same amount. In this case, the expected loss decreases. On the other hand, if $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle > 1$ for any j, then one can decrease $\langle \boldsymbol{f}^*, \boldsymbol{W}_j \rangle$ to 1 and increase $\langle \boldsymbol{f}^*, \boldsymbol{W}_i \rangle$, such that the loss decreases. As a result, we have that the argmax of $\langle \boldsymbol{f}^*, \boldsymbol{W}_i \rangle;\ i = 1, \ldots, k$ is not unique. Hence, the MOML with the hinge loss is not Fisher consistent. ■

Proof of Theorem 5: Take partial derivative of $S(\boldsymbol{x})$ with respect to each element in $\boldsymbol{f}$ and set to zero, which can be written as

$$\frac{\partial S}{\partial f_j} = \sum_{i=1}^{k} R(\boldsymbol{x}, i) \ell'(\langle \boldsymbol{W}_i, \boldsymbol{f} \rangle) W_i^{(j)} = 0,$$

for $j \in \{1, \ldots, k-1\}$. Here $W_i^{(j)}$ is the jth element of $\boldsymbol{W}_i$. One can verify that this is equivalent to $\sum_{i=1}^{k} R(\boldsymbol{x}, i) \ell'(\langle \boldsymbol{W}_i, \boldsymbol{f} \rangle) \boldsymbol{W}_i = \boldsymbol{0}_{(k-1)}$. Notice that $\sum_{i=1}^{k} \boldsymbol{W}_i = \boldsymbol{0}_{(k-1)}$ and $\boldsymbol{W}_i;\ i = 1, \ldots, k-1$ are linearly independent, hence one can conclude that $R(\boldsymbol{x}, i) \ell'(\langle \boldsymbol{W}_i, \boldsymbol{f} \rangle) = R(\boldsymbol{x}, i) \ell'(\langle \boldsymbol{W}_j, \boldsymbol{f} \rangle)$ for $i \neq j$. This completes the proof. ■

Proof of Proposition S1: Using the same technique as in the proof of Proposition 3.1 in Zhao et al. (2012), one can verify that if $R(\boldsymbol{x}, j) < \sum_{i \neq j} R(\boldsymbol{x}, i)$, then the theoretical minimizer is such that the recommended treatment is not j. Hence, the one-versus-rest approach is not Fisher consistent for any binary loss function. ■

Proof of Theorem S1: Because the loss functions considered are convex, one can

verify that for any $\beta^*_{j,q} \neq 0$, we have that

$$\left[\frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle \boldsymbol{W}_A, \boldsymbol{f}_0(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}}\right] |_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0} \neq 0,$$

where $|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0}$ means the derivative is evaluated at the best $\boldsymbol{\beta}^*$ and $\beta^*_{\cdot,0}$, except that for $\beta_{j,q}$ it is at 0. Similarly, we have that for any $\beta^*_{j,q} = 0$,

$$\left[\frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle \boldsymbol{W}_A, \boldsymbol{f}_0(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}}\right] |_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0} = 0.$$

To prove selection consistency, we need to show that for the empirical loss function with the fitted $\hat{\boldsymbol{f}}$, the corresponding partial derivative is not far away from its expectation. Then, with the help of l_1 penalties, we can select the variables correctly.

Before we present our proof for the main theorem, we give some lemmas to make the flow clear. We first assume that the marginal distribution of the rewards has a bounded range uniformly. We will consider the more general case of sub-Gaussian distribution later.

Lemma S1. *Suppose Conditions C1-C3 are valid. With the λ specified in Theorem S1, we have that $\|\hat{\boldsymbol{\beta}}_j\|_1 \preceq O_P\{n^{1/4}\log(p)^{-1/2}\}$ and $|\hat{\beta}_{j,0}| \preceq O_P\{n^{1/4}\log(p)^{-1/2}\}$ for all j.*

Proof of Lemma S1: With $\boldsymbol{\beta}_j = \mathbf{0}$ and $\beta_{j,0} = 0$ for all j, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}(\boldsymbol{x}_i)\rangle\} \to E\{\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R(0)\},$$

which is a constant. On the other hand, $\hat{\boldsymbol{\beta}}_j$ and $\hat{\beta}_{j,0}$ are the solution to the objective

function in (2.9), hence

$$\lambda\sum_{j=1}^{k-1}\|\hat{\boldsymbol{\beta}}_j\|_1 \le \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i},\hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\} + \lambda\sum_{j=1}^{k-1}\|\hat{\boldsymbol{\beta}}_j\|_1 \le E\{\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R(0)\}$$

for n large enough. Consequently, we have $\|\hat{\boldsymbol{\beta}}_j\|_1 \preceq O_P\{n^{1/4}\log(p)^{-1/2}\}$. For $|\hat{\beta}_{j,0}|$, one can verify that $\max_j|\hat{\beta}_{j,0}| \preceq \max_j\|\hat{\boldsymbol{\beta}}_j\|_1$. Otherwise, the prediction would be all the same for any new patient, which is not desirable. This completes the proof. □

Next, we prove that at the solution $\hat{\boldsymbol{f}}$, the partial derivative of the loss with respect to any β values can only deviate from its expectation by a small amount. Therefore, with the help of l_1 penalization, we can achieve selection consistency. The following lemma controls the difference of the expected partial derivatives between $\hat{\boldsymbol{f}}$ and $\boldsymbol{f}_0$.

Lemma S2. *Suppose Conditions C1-C3 are valid. With the λ specified in Theorem S1, we have that for any $j = 1,\ldots,k-1$ and $q = 1,\ldots,p$,*

$$\left|\left[\frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle \boldsymbol{W}_A,\hat{\boldsymbol{f}}(\boldsymbol{X})\rangle\}]}{\partial\beta_{j,q}} - \frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle \boldsymbol{W}_A,\boldsymbol{f}_0(\boldsymbol{X})\rangle\}]}{\partial\beta_{j,q}}\right]|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0}\right|$$
$$= O_P\left\{\frac{\log(p)}{\sqrt{n}}\right\}^{1/2}.$$

Moreover, we have that the difference between the expected partial derivative and its empirical value is also small for $\hat{\boldsymbol{f}}$, as in the following lemma.

Lemma S3. *Suppose Conditions C1-C3 are valid. With the λ specified in Theo-*

rem S1, we have that for any $j = 1, \ldots, k-1$ *and* $q = 1, \ldots, p$,

$$\left| \left[\frac{\partial [\frac{1}{n} \sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)} \ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\}]}{\partial \beta_{j,q}} - \frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R\{\langle \boldsymbol{W}_A, \hat{\boldsymbol{f}}(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}} \right] |_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0} \right|$$
$$= O_P \left\{ \frac{\log(p)}{\sqrt{n}} \right\}^{1/2}. \tag{A5.4}$$

Consequently, one can verify that the solution $\hat{\boldsymbol{f}}$ enjoys selection consistency. Next, we give the proofs to the two important lemmas.

Proof of Lemma S2: The proof of this lemma consists of two parts. The first part is to prove that the excess ℓ risk (defined in Section A4.3) converges at a rate at least $O_P \left\{ \frac{\log(p)}{\sqrt{n}} \right\}^{1/2}$. The second part is to show that the convergence rate of $\|\hat{\boldsymbol{f}} - \boldsymbol{f}_0\|_2$ is dominated by that of the excess ℓ risk, which further leads to the bound on the convergence rate of the derivatives.

To prove the first part, note that the optimization problem (2.9) can be written in an equivalent form

$$\min_{\boldsymbol{f}} \frac{1}{n} \sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)} \ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}(\boldsymbol{x}_i)\rangle\}, \text{ subject to } \sum_{j=1}^{k-1} (\|\boldsymbol{\beta}_j\|_1 + |\beta_{j,0}|) < s(\lambda), \tag{A5.5}$$

where $s(\lambda)$ is a tuning parameter. By Lemma S1, we have that $s(\lambda) = O_P\{n^{1/4} \log(p)^{-1/2}\}$. Next, we note that by similar arguments as in the proof of Theorem 4 in Zhang and Liu (2014), one have that the excess ℓ risk $E\left\{ \frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R(\langle \boldsymbol{W}_A, \hat{\boldsymbol{f}}\rangle) \right\} - E\left\{ \frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R(\langle \boldsymbol{W}_A, \boldsymbol{f}_0\rangle) \right\}$ converges at the rate $O_P\{s(\lambda)\sqrt{\frac{\log(p)}{n}}\} \preceq O_P \left\{ \frac{\log(p)}{\sqrt{n}} \right\}^{1/2}$. Next, we show the relationship between the convergence rate of $\|\hat{\boldsymbol{f}} - \boldsymbol{f}_0\|_2$ and the excess ℓ risk in Lemma S4 below. Note that the proof of Lemma S4 is analogous to that of Theorems 5 and 6

in Zhang and Liu (2013) and is omitted here.

Lemma S4. *Suppose that Conditions C1-C3 are valid. Moreover, consider a loss function $L\{u(\boldsymbol{f}, y)\}$ that is second order differentiable with respect to u, where $u(\boldsymbol{f}, y)$ is a function of the response y and the learning function $\boldsymbol{f}$. Assume that u has second order derivative with respect to each element in $\boldsymbol{f}$, and the two second order derivatives are both bounded. Then we have that, if the function $\boldsymbol{f}^*$ minimizes $E(L)$,*

$$|E[L\{u(Y, \boldsymbol{f})\}] - E[L\{u(Y, \boldsymbol{f}^*)\}]| = O\{(\|\boldsymbol{f} - \boldsymbol{f}^*\|_2)^2\},$$

and if $\boldsymbol{f}^*$ is not the minimizer of $E(L)$,

$$|E[L\{u(Y, \boldsymbol{f})\}] - E[L\{u(Y, \boldsymbol{f}^*)\}]| = O\{(\|f - f^*\|_2)\}.$$

By Lemma S4, we can see that the convergence rate of $\|\hat{\boldsymbol{f}} - \boldsymbol{f}_0\|_2$ is dominated by that of the excess ℓ risk, which can further leads to that

$$\left| \left[\frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R\{\langle \boldsymbol{W}_A, \hat{\boldsymbol{f}}(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}} - \frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R\{\langle \boldsymbol{W}_A, \boldsymbol{f}_0(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}} \right] |_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0} \right|$$
$$= O_P \left\{ \frac{\log(p)}{\sqrt{n}} \right\}^{1/2}.$$

This completes the proof of Lemma S2. □

Proof of Lemma S3: This proof consists of two parts. The first part is to use the Rademacher complexity (Mohri et al., 2012) technique to show that with a high probability, the difference between the first and second terms in (A5.4) is bounded by the Rademacher complexity of the functional space considered in (A5.5). The second

part is to show that the Rademacher complexity of the functional space converges at the desired rate.

To begin with, we introduce the Rademacher complexity. Let $\sigma_i;\ i = 1, \ldots, n$ be *i.i.d.* random variables, each taking the value 1 with probability 1/2, and -1 with probability 1/2. Let the set of training observations $(\boldsymbol{x}_i, a_i, r_i);\ i = 1, \ldots, n$, which are observed from P, be denoted by S. Define the function class of (A5.5) to be

$$\mathcal{H} = \left\{ \hat{\boldsymbol{f}} : \hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)} \ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}(\boldsymbol{x}_i)\rangle\},\ \text{subject to } \sum_{j=1}^{k-1}(\|\boldsymbol{\beta}_j\|_1 + |\beta_{j,0}|) < s(\lambda) \right\}.$$

Fix S, and we define the empirical Rademacher complexity of the function class $\mathcal{H}$ as

$$\hat{R}_n\{\mathcal{H}\} = E_{\boldsymbol{\sigma}} \left\{ \sup_{\sum_{j=1}^{k-1}(\|\boldsymbol{\beta}_j\|_1 + |\beta_{j,0}|) < s(\lambda)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)} \ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}(\boldsymbol{x}_i)\rangle\} \right] \right\},$$

where $E_{\boldsymbol{\sigma}}$ represents the expectation with respect to $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$. Next, define the Rademacher complexity of $\mathcal{H}$ by

$$R_n\{\mathcal{H}\} = E_S \hat{R}_n\{\mathcal{H}\},$$

where E_S is the expectation with respect to the distribution of the sample S.

Next, we prove that, with the conditions C1-C3 valid and λ in Theorem S1, we

have that with probability at least $1-\delta$ ($\delta > 0$ is a small positive number),

$$\begin{aligned}
&\left| \left[\frac{\partial [\frac{1}{n}\sum_{i=1}^{n} \frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)} \ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\}]}{\partial \beta_{j,q}} - \frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})} \ell_R\{\langle \boldsymbol{W}_A, \hat{\boldsymbol{f}}(\boldsymbol{X})\rangle\}]}{\partial \beta_{j,q}} \right] \right| \\
&\leq C_1 R_n\{\mathcal{H}\} + T_n(\delta) \\
&\leq C_1 \hat{R}_n\{\mathcal{H}\} + 3T_n(\delta/2), \qquad \text{(A5.6)}
\end{aligned}$$

where $T_n(\delta) = C_2\{n^{-1}\log(p)\log(1/\delta)\}^{1/2}$, and C_1, C_2 are universal constants that are independent of n and p.

The proof to this claim is standard in the Rademacher complexity literature. To bound the first term of (A5.6) by $C_1 R_n\{\mathcal{H}\} + T_n(\delta)$, one can use the McDiarmid inequality (McDiarmid, 1989) and the symmetrization technique (Van der Vaart and Wellner, 2000). To bound $C_1 R_n\{\mathcal{H}\}$ by $C_1 \hat{R}_n\{\mathcal{H}\} + 2T_n(\delta/2)$, one can use the McDiarmid inequality again. See the proof of Lemma 3 in Zhang et al. (2015) for more details.

Note that there are one major difference between the proof of (A5.6) and that of Lemma 3 in Zhang et al. (2015). In particular, the maximum change in the first term of (A5.6) should one replace a $\boldsymbol{x}_i$ or a_i can be bounded by $C_3\log(p)^{1/2}/n^{5/4}$ (this is a direct result from Lemma S1), where C_3 is another universal constant, instead of $O_P(n^{-1})$ as in Zhang et al. (2015). Because $O_P\{\log(p)^{1/2}/n^{5/4}\} \preceq O_P(n^{-1})$, this does not change the conclusion much. The rest of the proof is analogous, and we omit the details here.

The next step is to bound the empirical Rademacher complexity of $\mathcal{H}$. To this end, we note that $\hat{R}_n\{\mathcal{H}\}$ can be upper bounded by the following Rademacher complexity,

up to a constant scalar

$$\hat{R}_n^*\{\mathcal{H}\} = E_{\boldsymbol{\sigma}}\left\{\sup_{\sum_{j=1}^{k-1}(\|\boldsymbol{\beta}_j\|_1+|\beta_{j,0}|)<s(\lambda)} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\{\sum_{j=1}^{k-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_j+\beta_{j,0})\}\right\}, \tag{A5.7}$$

because R is bounded, each element in $\boldsymbol{W}_j$ is bounded by 1, and we assume that ℓ is second order differentiable (see Lemma 4.2 in Mohri et al., 2012). Without loss of generality, we can rewrite (A5.7) as follows

$$\hat{R}_n^*\{\mathcal{H}\} = E_{\boldsymbol{\sigma}}\left\{\sup_{\|\boldsymbol{\gamma}\|_1<s(\lambda)} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{\gamma}^T\boldsymbol{x}_i^*\right\}, \tag{A5.8}$$

where $\boldsymbol{\gamma}$ can be regarded as a vector that contains all the elements in $\boldsymbol{\beta}_j$ and $\beta_{j,0}$ for $j = 1,\ldots,k-1$, and $\boldsymbol{x}_i^*$ is defined accordingly. Next, using Theorem 10.10 in Mohri et al. (2012), we have that $\hat{R}_n^*\{\mathcal{H}\} \preceq O_P\left\{\frac{\log(p)}{\sqrt{n}}\right\}^{1/2}$ (note that $s(\lambda) = O_P\{n^{1/4}\log(p)^{-1/2}\}$).

Next, choose $\delta = 2p^{-1}n^{-2}$, and one has that $T_n(\delta/2) \preceq O_P\left\{\frac{\log(p)}{\sqrt{n}}\right\}^{1/2}$. Consequently, with probability at least $2n^{-2}$, (A5.4) holds true for all the covariates. Using the Borel–Cantelli Lemma, we have proved Lemma S3. □

Combining Lemmas S2 and S3, we have that

$$\left|\left[\frac{\partial[\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle\boldsymbol{W}_{a_i},\hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\}]}{\partial\beta_{j,q}} - \frac{\partial E[\frac{|R|}{\pi_A(\boldsymbol{X})}\ell_R\{\langle\boldsymbol{W}_A,\boldsymbol{f}_0(\boldsymbol{X})\rangle\}]}{\partial\beta_{j,q}}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,\ \beta_{\cdot,0}=\beta^*_{\cdot,0},\ \beta_{j,q}=0}\right|$$
$$= O_P\left\{\frac{\log(p)}{\sqrt{n}}\right\}^{1/2}.$$

Hence, one can verify that at the solution $\hat{\boldsymbol{f}}$, selection consistency is equivalent to that $\lambda \to 0$ at a rate no faster than $O_P\left\{\frac{\log(p)}{\sqrt{n}}\right\}^{1/2}$. This completes the proof when the rewards are bounded random variables.

To finish the proof, we consider the general case where the distribution of the rewards is sub-Gaussian. As a matter of fact, one can show that with a high probability, the actual rewards would be bounded in a range. Then we can prove that the corresponding converge rate are the same, because the probability of sub-Gaussian random variables being significantly away from its expectation converges to zero very fast as the bound increases.

Without loss of generality, we assume that the reward for any patient and any treatment follows a common sub-Gaussian distribution, with the corresponding c.d.f. Φ_R. The generalization of this assumption to the heteroscedastic case is straightforward, because what really matters is the tail probability $\text{pr}(|R(\boldsymbol{X}, A)| > t)$ for large t. Next, define $t^* = \Phi_R^{-1}\{0.5+0.5(1-\delta/2)^{1/n}\}$, where δ is a small positive number. It can be verified that with probability at least $1-\delta/2$, all the rewards $r_i;\ i = 1, \ldots, n$ deviate from its expectation within $[-t^*, t^*]$. Since Φ_R is the c.d.f. of a sub-Gaussian distribution with a fixed parameter, t^* diverges at a rate slower than $O_P\{\log(n)\}$. One can check that the RHS of the displays in Lemmas S2 and S3 can be bounded similarly as in the corresponding proofs. This completes the proof.

Note that we assume sub-Gaussian distribution for the rewards to control the probability of observing a reward that is significantly away from its expectation. If one can verify that such a probability is small, for example, the joint distribution P can guarantee that such a tail probability is negligible (in particular, the Radon-Nikodym derivative of the covariates is small when the conditional reward for given $\boldsymbol{X}$ has a heavy tail), Condition C2 can be removed. ■

Proof of Theorem S2: The proof is similar to that of Theorem 4 in Zhang and Liu (2014). Here, we point out the major differences. In particular, consider the optimization problem (A5.5). Note that in the proof of Theorem S1, we argue that

$s(\lambda) \preceq O_P\{n^{1/4}\log(p)^{-1/2}\}$, hence the excess ℓ risk converges at the rate at least $O_P\{\log(p)^{1/2}n^{-1/4}\}$. In this proof, we are going to show that with similar results as those in Lemmas S2 and S3, one can obtain that $s(\lambda) = O_P(1)$, and this can lead to the fast convergence rate of the excess ℓ risk.

To begin with, note that we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\} + \lambda\sum_{j=1}^{k-1}\|\hat{\boldsymbol{\beta}}_j\|_1 \leq \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}_0(\boldsymbol{x}_i)\rangle\} + \lambda\sum_{j=1}^{k-1}\|\boldsymbol{\beta}_j^*\|_1,$$

therefore,

$$\lambda\sum_{j=1}^{k-1}\|\hat{\boldsymbol{\beta}}_j\|_1 \leq \lambda\sum_{j=1}^{k-1}\|\boldsymbol{\beta}_j^*\|_1 + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}_0(\boldsymbol{x}_i)\rangle\} - \sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\}\right]. \tag{A5.9}$$

Now decompose the second term on the RHS of (A5.9) into

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}_0(\boldsymbol{x}_i)\rangle\} - \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\} \\
&= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \boldsymbol{f}_0(\boldsymbol{x}_i)\rangle\} - E\frac{R}{\pi_A(\boldsymbol{X})}\ell_R\{W_A, \boldsymbol{f}_0(\boldsymbol{X})\}\right] && \text{(A5.10)} \\
&+ \left[E\frac{R}{\pi_A(\boldsymbol{X})}\ell_R\{W_A, \boldsymbol{f}_0(\boldsymbol{X})\} - E\frac{R}{\pi_A(\boldsymbol{X})}\ell_R\{W_A, \hat{\boldsymbol{f}}(\boldsymbol{X})\}\right] && \text{(A5.11)} \\
&+ \left[E\frac{R}{\pi_A(\boldsymbol{X})}\ell_R\{W_A, \hat{\boldsymbol{f}}(\boldsymbol{X}) - \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{\pi_{a_i}(\boldsymbol{x}_i)}\ell_{r_i}\{\langle \boldsymbol{W}_{a_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\rangle\}\right]. && \text{(A5.12)}
\end{aligned}$$

By similar arguments as in Lemmas S2 and S3, one can verify that the summation of (A5.11) and (A5.12) converges at a rate no slower than $O_P\{\log(p)^{1/2}/(n^{1/4})\}$. Furthermore, by Conditions C1-C3 and the law of large numbers, one has that (A5.10) is of order $o_P\{\log(p)^{1/2}/(n^{1/4})\}$. Thus, the second term on the RHS of (A5.9) is of

order $O_P\{\log(p)^{1/2}/(n^{1/4})\}$, and one can verify that $s(\lambda) = O_P(1)$.

Next, using similar techniques as in the proof of Theorem 4 in Zhang and Liu (2014), we have that the convergence rate of the excess ℓ risk in this paper is of the desired order. This completes the proof. ■

References

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.

Fan, J. and Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica* **20,** 101.

Marron, J. S., Todd, M., and Ahn, J. (2007). Distance Weighted Discrimination. *Journal of the American Statistical Association* **102,** 1267–1271.

McDiarmid, C. (1989). On the Method of Bounded Differences. In *In Surveys in Combinatorics*, pages 148–188. Cambridge University Press.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT press.

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E., and Yuan, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat* **4,** 59–68.

Van der Vaart, A. W. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes with Application to Statistics*. Springer, 1st edition.

Zhang, C. and Liu, Y. (2013). Multicategory Large-margin Unified Machines. *Journal of Machine Learning Research* **14,** 1349–1386.

Zhang, C. and Liu, Y. (2014). Multicategory Angle-based Large-margin Classification. *Biometrika* **101,** 625–640.

Zhang, C., Liu, Y., and Wu, Y. (2015). On Quantile Regression in Reproducing Kernel Hilbert Spaces with Data Sparsity Constraint. *Journal of Machine Learning Research* .

Zhang, C., Liu, Y., and Wu, Z. (2013). On the Effect and Remedies of Shrinkage on Classification Probability Estimation. *The American Statistician* **67,** 134–142.

Zhang, X., Wu, Y., Wang, L., and Li, R. (2014). Variable Selection for Support Vector Machines in Moderately High Dimensions. *Journal of the Royal Statistical Society: Series B* .

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* **7,** 2541–2563.

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.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

E-mail: zhangchong101@gmail.com

Department of Biostatistics, University of North Carolina at Chapel Hill

E-mail: jgxchen@email.unc.edu

Eli Lilly and Company

E-mail: fu_haoda@lilly.com, he_xuanyao@lilly.com

Public Health Sciences Division, Fred Hutchinson Cancer Research Center

E-mail: yqzhao@fredhutch.org

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

E-mail: yfliu@email.unc.edu