Support Vector Decision Making*

Yixiao Sun†
Department of Economics, UC San Diego

December 2024

Abstract

The paper develops a support vector machine (SVM) for binary decision-making within a utility framework. Given an information set, a decision-maker first predicts a binary outcome and then selects a binary action based on this prediction to maximize expected utility, where the utility function can depend on the action taken, observable covariates, and the binary outcome subsequently realized. The proposed maximum utility SVM differs from the traditional SVM in four key aspects. First, as a conceptual innovation, it incorporates the optimal cut-off function as a separate and special covariate. Second, there is a sign restriction on this special covariate. Third, it accounts for the dependence of the utility-induced loss on both the covariates and the binary outcome. Finally, it allows the margin to differ across different classes of outcomes. The paper proves that the proposed method is Bayes-consistent under the maximum utility criterion and establishes a finite-sample generalization bound. A simulation study shows that the proposed method outperforms existing methods under the data-generating processes considered in the literature.

Keywords: Decision-based binary prediction, Maximum margin decision, Maximum utility estimation, Support vector decision, Support vector machine.

JEL Classification: C14, C45, C52, C53

*I thank a coeditor, an associate editor, three anonymous referees, and seminar and conference participants at Indiana University, Princeton University, and the 2021 Korean Economic Review (KER) International Conference for their helpful and constructive comments and suggestions.

†Address correspondence to: Department of Economics, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508. Email: yisun@ucsd.edu

1 Introduction

This paper studies binary decision making in a utility maximization framework. After observing a training sample $\{(X_i, Y_i)\}_{i=1}^n$ and an out-of-sample covariate $X \in \mathbb{R}^{d_x}$, the decision-maker predicts the binary outcome variable $Y \in \{-1, 1\}$ based on covariate X, and then chooses a binary action $a \in \{-1, 1\}$ before the outcome Y is realized. The decision-maker's preference is captured by a utility function $U(a, Y, X)$, which quantifies the payoff associated with each action a, given the realized outcome Y and the observed covariate X. The objective of the decision-maker is to evaluate the likelihood of the outcome Y based on covariate X and then choose an action a that maximizes the expected payoff.[1]

To illustrate our binary decision-making framework, we present a few examples below.

Loan Default Prediction and Approval Decision: A bank uses an applicant's financial information, such as credit score, income, and loan amount (covariate X), to predict whether the applicant will default on the loan (outcome $Y \in \{-1, 1\} = \{\text{default, no default}\}$). Based on this prediction, the bank then decides whether to approve (action $a = 1$) or deny (action $a = -1$) the loan.

Job Performance Prediction and Hiring Decision: A company predicts whether a job candidate will perform well in a job (outcome $Y \in \{-1, 1\} = \{\text{poor performance, good performance}\}$) based on factors such as qualifications, experience, and interview results (covariate X). The company then decides whether to hire (action $a = 1$) or reject (action $a = -1$) the candidate.

Pandemic Prediction and Emergency Preparation: Public health officials use surveillance data, including infection rates, mobility patterns, vaccination rates, and other relevant factors (covariate X), to predict whether a pandemic will occur (outcome $Y \in \{-1, 1\} = \{\text{no pandemic, pandemic}\}$). Based on the prediction, the authorities decide whether to prepare emergency response teams for pandemic-related scenarios (action $a \in \{-1, 1\} = \{\text{do not prepare, prepare}\}$).

Sentiment Prediction and Moderation Decision: A social media platform uses the presence (or absence) of certain keywords (covariate X) to predict whether a post will receive negative or positive sentiment (outcome $Y \in \{-1, 1\} = \{\text{negative sentiment, positive sentiment}\}$). Based on this prediction, the platform decides whether to take moderation actions, such as temporarily muting the user for a certain period (action $a \in \{-1, 1\} = \{\text{moderate, do not moderate}\}$).

These examples highlight three key features in our binary decision framework.

First, the decision-maker considers the future and has to predict a binary outcome. The prediction is based on the historical information $\{(X_i, Y_i)\}_{i=1}^n$, as well as an out-of-sample observed covariate X. This prediction helps the decision-maker assess the likelihood of the outcome before making a choice.

Second, the decision-maker has to choose an action before the outcome Y is realized, as no further information about Y, beyond the covariate X and sample information, is available at the

[1] In our setting, the payoff may also depend on another vector of observed covariates, say Z, which does not help predict Y but can enter the payoff function. In this case, we can include Z as part of X, while continuing to use $U(a, Y, X)$ as the utility function, recognizing that not all elements of X are predictive of Y.

time of the decision. Importantly, the action does not influence the eventual outcome realized but is based on the predicted probability of that outcome.

Third, the utility function allows for asymmetric and covariate-specific consequences: the consequences of different actions may have varying degrees of severity depending on the outcome realized and the covariate observed. For instance, in the loan example above, approving a loan that defaults (i.e., $Y = -1$) may result in more significant financial loss than denying a loan that would have been repaid (i.e., $Y = 1$). Thus, the cost of making a "false positive" decision (approving a loan that defaults) may be much higher than the cost of making a "false negative" decision (denying a loan that would have been repaid). These costs may also depend on the loan characteristics and the applicant's credit profile. This asymmetry in the payoff structure is what distinguishes our decision framework from a standard binary decision problem, where the decision-maker cares only about the direction of the decision (approve or deny) and treats all decision errors equally.

Our framework is the same as Granger and Machina (2006), Elliott and Lieli (2013), and Su (2021). Elliott and Lieli (2013) propose and study a maximum utility (MU) action rule that maximizes expected utility when the model for the conditional distribution of Y given $X = x$ may be misspecified. Su (2021) employs a penalized MU approach for model selection within the MU framework. However, the MU approach does not account for the margin of a point defined as its distance to the decision boundary.

As an example, consider the case of complete separation when the number of elements d_x in X equals 2, as shown in Figure 1a. The figure plots $X_i = (X_{1i}, X_{2i}) \in \mathbb{R}^2$ together with the associated outcome Y_i indicated by the shape and color. It also shows the two decision boundaries corresponding to action rules a_1 and a_2. According to the MU approach, a_1 and a_2 are equivalent as they deliver the same in-sample empirical utility. However, the action rule a_1 is expected to have a smaller out-of-sample generalization error. The reason is that a_1 is closer than a_2 to achieving the maximum separation between the two classes of points. The problem of ignoring the margin is further illustrated in Figure 1b where a_1 is indistinguishable from a_2 according to the MU criterion, but it is expected to have a smaller generalization error.

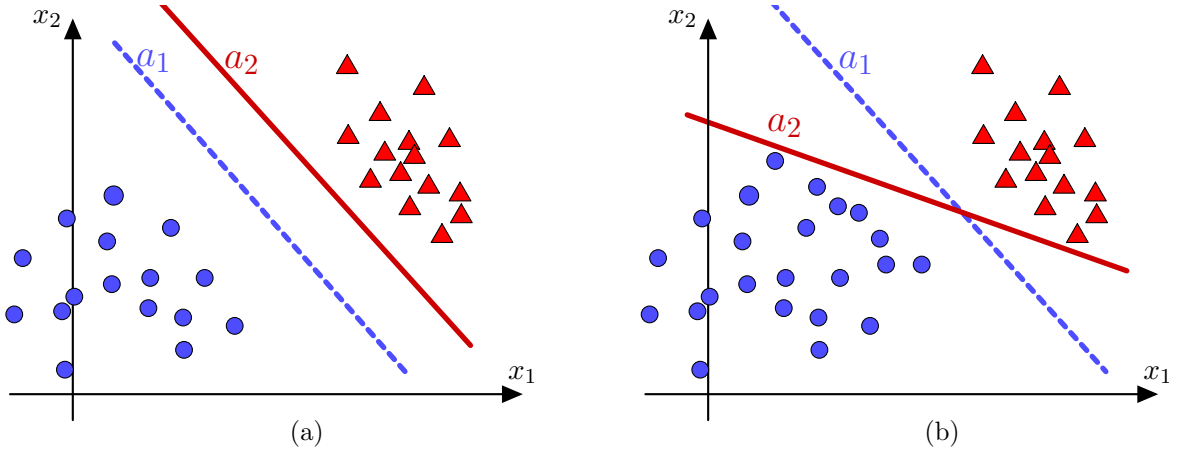


Figure 1: Action rules a_1 and a_2 have the same in-sample empirical utility, but a_1 is expected to have a smaller out-of-sample generalization error than a_2 ($d_x = 2$).

In the presence of incomplete separation and a covariate-specific payoff, the notion of margin has to be redefined. The optimal decision boundary now takes the form of $\text{sign}\left(P\left(X\right) - c\left(X\right)\right)$,

where $P(X)$ is the probability of $Y = 1$ given X, and $c(X)$ measures the relative cost of false decisions under X (to be defined more precisely later in (1)). Hence, the optimal decision boundary involves a cutoff function $c(X)$ that depends on the covariate. This dependence renders the standard support vector machine (SVM, e.g., Boser et al. (1992) and Cortes and Vapnik (1995)) invalid. To address this problem, we treat the function $c_-(X) = -c(X)$ as a separate covariate. By combining $c_-(X)$ with the original covariate vector X, we obtain the augmented covariate vector $W = (X', c_-(X))'$, which may be referred to as the generalized attributes. With this conceptual change, the original sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ becomes the new sample $\{(W_i, Y_i), i = 1, \ldots, n\}$. This opens the door to using SVM, enabling us to capture the nuanced, covariate-dependent cutoff function at the same time. We then apply the SVM framework to the new sample, treating $c_-(X)$ as a special covariate whose coefficient is constrained to be nonnegative. At the same time, we account for the case-specific loss implied by the dependence of the payoff function on the outcome, the action, and the covariate. We call our approach the maximum utility SVM (MU-SVM).

Invented by Vapnik and Chervonenkis in 1963, the SVM has attracted tremendous attention, resulting in a vast body of literature. The MU-SVM extends the standard SVM in three aspects. First, it internalizes the cutoff function as a separate and special covariate. Second, it accommodates the sign restriction on this special covariate. Third, it accounts for the dependence of the utility-induced loss on the covariate (X) and the binary outcome (Y).

From a broad perspective, the contributions of this paper relative to the MU approach are analogous to those of the standard SVM relative to the maximum score approach of Manski (1975, 1985). Neither the MU approach nor the maximum score approach incorporates the notion of a margin, which is a central feature of both the standard SVM and the approach we are proposing. On the other hand, the standard SVM lacks the concept of a utility function, which is a key feature of our framework.

Like the standard SVM, the MU-SVM exhibits sparsity. The entire sample collectively determines the "support" subsample, which in turn defines the decision boundary. The set of W_i's corresponding to the support subsample is called the support vectors. Once the set of support vectors, denoted by $\{W_i : i \in \mathcal{S}\}$ for some index set $\mathcal{S}$, is determined, the decision rule for a new, out-of-sample attribute w is based on the similarity of w to each support vector W_i for $i \in \mathcal{S}$. Our proposed decision-making method can then be referred to as support vector decision making. It can also be interpreted as a voting rule, with the support subsample acting as the weighted "representatives" who cast votes, with each vote carrying a different weight.

We extend our approach to accommodate nonparametric specifications of the decision boundary. The "kernel trick" based on the theory of reproducing kernel Hilbert spaces (RKHS) can be employed, but the kernel function we use will have an additional component that captures the case-specific cutoff function.

We show that any decision rule that is Bayes-consistent under our MU-SVM criterion is also Bayes-consistent under the MU criterion of Elliott and Lieli (2013) and Su (2021).[2] We establish a generalization bound that can be used, in principle, to construct a finite sample confidence interval for the average out-of-sample utility obtained from using our support vector decision rule.

A variant of the traditional SVM closely related to this paper is the cost-sensitive SVM. See, for example, Lin et al. (2002), Bach et al. (2006), and Fernández et al. (2018). These papers allow

[2]If the excess risk of a decision rule under a certain risk measure goes to zero as the sample size increases, then we say that the decision rule is Bayes-consistent under this risk measure. See Definition 3.

the misclassification cost to depend on the binary outcome variable (Y) but not on the covariate (X). In comparison, we allow the misclassification cost to depend on both Y and X. Despite its importance, the general case we consider here has received little attention in the SVM literature, with only a few exceptions. One exception is Brefeld et al. (2003) which proposes an SVM-type learning rule for the general case. However, as discussed following Proposition 3 in that paper, the proposed rule may not be Bayesian optimal. Another exception is Iranmehr et al. (2019) where the authors use the idea of probability elicitation to design a new loss function. However, their proposed decision rule depends on the individual costs associated with false positives and false negatives separately. This is not a desirable property, as the decision problem remains the same as long as the ratio of the two costs does not change. An ideal decision rule should exhibit invariance to proportional changes in these two costs. In contrast to the existing literature, our decision rule possesses this invariance property and is also Bayesian optimal. This represents a novel contribution to the SVM literature, as it allows for a more general cost function and provides a more principled approach to cost-sensitive decisions.

In our simulation studies, we compare the MU-SVM method to existing methods, using the out-of-sample utility achieved as the performance criterion. Our simulation results show that the MU-SVM outperforms existing methods in an overall sense. First, it outperforms the maximum likelihood method when the model is misspecified. Second, it outperforms the standard SVM method whenever the utility function depends on the outcome or the covariate, and it reduces to the standard SVM method when there is no such dependence. Third, it outperforms the cost-sensitive SVM, as described in Lin et al. (2002), Bach et al. (2006), and Fernández et al. (2018), when the utility function is covariate-dependent. Fourth, it outperforms the penalized MU method of Su (2021) that uses the simulated maximum discrepancy as the penalty.

While the framework presented in this paper covers a wide range of applications, there are certain binary prediction and decision problems that fall outside its scope. As pointed out by Elliott and Lieli (2013), one such case arises when the forecaster and the decision maker are not the same entity. For example, meteorologists use weather data to predict whether a storm will occur, but different users may use the prediction differently. City planners might rely on this prediction to decide whether to prepare emergency shelters, while average citizens might use it to determine whether to stay at home. In this scenario, each user of the prediction would have a distinct utility function that influences their decision-making, but meteorologists do not take these utility functions into account when making predictions.

Another case outside the scope of this framework arises when the action affects the eventual outcome. In such cases, the action can be interpreted broadly as a treatment that directly influences the observed outcome. Our framework does not accommodate this type of relationship, setting it apart from the econometric literature on empirical welfare maximization (e.g., Kitagawa and Tetenov (2018)) and the related statistical literature on individualized treatment rules (e.g., Zhao et al. (2012), Zhou et al. (2017), Liu et al. (2018), ITR hereafter). The latter strand of the ITR literature is related to the present paper, as it also employs SVM. However, in these ITR papers, the goal is to assign a treatment based on covariates in order to affect the eventual outcome so as to maximize a welfare objective. Predicting a binary outcome is not part of the design of an ITR. The settings and objectives of our framework and the ITR literature are conceptually different; it is not feasible to treat one as a special case of the other due to their fundamentally distinct assumptions and information structures.

Finally, our framework does not apply to the matching market problem. The matching problem typically involves two distinct sides (e.g., job seekers and employers), where both sides

have preferences over participants on the other side. These preferences usually stem from mutual evaluation between the two sides, and matches are based on these preferences. The goal is often to find a stable or optimal match that satisfies both sides. In contrast, our framework involves a one-sided decision process. In the hiring example, only the employer chooses to hire or reject a candidate. The job seeker does not make a choice in terms of matching preferences. Instead, the employer's decision is based solely on its own predictions of candidate performance, with the goal of hiring only those candidates predicted to perform well. There is no reciprocal matching process between the two sides, as is typical in the classic matching market model.

The rest of the paper is organized as follows. Section 2 lays out the basic setting. Section 3 motivates the support vector decision from first principles and describes the method, its dual problem, and its computational aspects. Section 4 extends the method to allow for a nonlinear and nonparametric decision boundary. Section 5 establishes some theoretical results, including the Bayes consistency of the support vector decision rule under the MU criterion and a generalization bound that accommodates data-driven choices of tuning parameters. Section 6 reports the simulation results. The last section concludes and discusses future research directions. Proofs of the main theoretical results are provided in the Appendix. Further supporting materials are given in the online supplementary appendix, which includes additional proofs and results, along with a list of methods considered in the simulation study.

2 The Basic Setting

2.1 The framework

We adopt the standard decision-theoretic framework. A decision-maker observes a vector of covariates $X \in \mathbb{R}^{d_x}$ and needs to make a decision regarding the binary action $a \in \{-1, 1\}$. Here, an action is defined in a broad sense. For example, it could indicate whether a bank approves a loan, a firm hires a job candidate, public health officials prepare for an emergency response, or a social media platform flags a post for moderation. The payoff or utility function of the decision-maker is $U(a, Y, X)$, where $Y \in \{-1, 1\}$ is a binary outcome that is not observable at the time of decision making. The table below illustrates the payoff function under $X = x$ with different combinations of (a, Y) :

	State of the world	
Action	$Y = 1$	$Y = -1$
$a = 1$	$U(1, 1, x)$	$U(1, -1, x)$
$a = -1$	$U(-1, 1, x)$	$U(-1, -1, x)$

The payoff $U(a, y, x)$ depends on (a, y) and is possibly a nontrivial function of x for each given (a, y).

This standard decision-theoretic setting has also been considered by Granger and Machina (2006), Elliott and Lieli (2013), and Su (2021). One may also regard a as a forecast of the random variable Y whose value will be realized at a future time. Then, $U(a, y, x)$ is the payoff when the forecast is a, the realized value of Y is y, and the covariate vector X is equal to x.

We expect that $U(1, 1, x) > U(-1, 1, x)$ and $U(-1, -1, x) > U(1, -1, x)$ for all x. That is, a correct prediction delivers a higher payoff than an incorrect prediction. We also assume that the payoff function is measurable and bounded. We formalize these conditions as an assumption.

Assumption 1 *(i) For all x in the support $\mathcal{X}$ of X, $U(1,1,x)-U(-1,1,x)>0$ and $U(-1,-1,x)-U(1,-1,x)>0$; (ii) For all $(a,y)\in\{-1,1\}^2$, $U(a,y,\cdot)$ is Borel measurable and*

$$U_{\max} := \sup_{(a,y)\in\{-1,1\}^2, x\in\mathcal{X}} U(a,y,x) < \infty.$$

Conditional on $X=x$, the outcome variable Y follows a Bernoulli distribution with parameter $P(x)$:

$$P(x) = \Pr(Y=1|X=x).$$

Our setting can be cast as a 2×2 game where Nature plays Y and the decision-maker plays a. Nature plays a mixed strategy: for a given $X=x$, Nature plays $Y=1$ with probability $P(x)$. The decision-maker plays a pure strategy by choosing $a=1$ or $a=-1$, with the payoff $U(a,y,x)$. Under Assumption 1(i), there is no dominating strategy for the decision-maker; otherwise, the decision problem becomes trivial.

Note that our framework also allows the following: a proper subvector of x to enter $P(x)$, and another proper subvector to enter $U(a,y,x)$. The covariate vector x can then be regarded as comprising all the covariates that enter either $P(x)$ or $U(a,y,x)$.

The decision-maker does not know $P(x)$, but she observes an i.i.d. sample (X_i,Y_i) for $i\in[n]\equiv\{1,2,\ldots,n\}$.

Assumption 2 *(i) $\{(X_i,Y_i):i\in[n]\}$ is an i.i.d. sample; (ii) $X_i\in\mathcal{X}\subseteq\mathbb{R}^{d_x}$ and $Y_i\in\{-1,1\}$ where d_x is the number of elements in X_i.*

In order to make an optimal decision, the decision-maker needs to learn $P(x)$ from the sample $\{(X_i,Y_i)\}_{i=1}^n$. While the decision-maker is interested in $P(x)$, their ultimate goal is to take the best action to maximize expected utility. When $X=x$ and the decision-maker takes action a, the expected utility that the decision-maker will obtain is $E[U(a,Y,X)|X=x]$, where the expectation is taken with respect to the conditional distribution Y given $X=x$. An action is optimal if it maximizes this expected utility; that is,

$$a^* = \arg\max_a E[U(a,Y,X)|X=x].$$

The optimal a^* depends on the observed covariate value x. To signify such dependence, we write it as $a^*(x)$. Equivalently, we can represent $a^*(x)$ as

$$\begin{aligned} a^*(x) &= \arg\min_a E[U(Y,Y,X)-U(a,Y,X)|X=x] \\ &= \arg\min_a E[\psi(Y,X)1\{a\neq Y\}|X=x], \end{aligned}$$

where

$$\psi(y,x) = U(y,y,x)-U(-y,y,x).$$

We can regard $\psi(y,x)$ as the loss incurred when an incorrect action is taken. More precisely, it represents the loss that arises from taking action $-y$ rather than y when the outcome and covariate are equal to y and x, respectively. Under Assumption 1(i), $\psi(y,x)>0$ for all $y\in\{-1,1\}$ and $x\in\mathcal{X}$. The loss $\psi(y,x)$ may be a nontrivial function of both y and x. In particular, the loss is not symmetric in the sense that false positives and false negatives may incur different losses. In other words, $\psi(1,x)$ may not equal $\psi(-1,x)$ for any $x\in\mathcal{X}$. See Granger and Machina (2006) for more discussion on utility-induced loss and loss functions.

When $X = x$ and $a = 1$, the expected loss (from a false-positive decision) is $(1-P(x))\psi(-1, x)$. When $X = x$ and $a = -1$, the expected loss (from a false-negative decision) is $P(x)\psi(1, x)$. Given $X = x$, the action $a = 1$ is optimal if the expected loss from a false-positive decision is lower than that from a false-negative decision. So $a^*(x) = 1$ if and only if $(1-P(x))\cdot\psi(-1, x) < P(x)\cdot\psi(1, x)$. That is, when $P(x) \neq 1$, it is optimal to take a "positive" action if and only if

$$\frac{P(x)}{1 - P(x)} > \frac{\psi(-1, x)}{\psi(1, x)}.$$

We can also interpret $\psi(-1, x)$ and $\psi(1, x)$ as regrets from taking non-optimal actions. The regret ratio $\psi(-1, x)/\psi(1, x)$ has to be smaller than the odds ratio of the event $Y = 1$ relative to the event $Y = -1$ to justify taking the action $a = 1$.

Define

$$\begin{aligned} b(x) &= \psi(-1, x) + \psi(1, x), \\ c(x) &= \frac{\psi(-1, x)}{\psi(-1, x) + \psi(1, x)}. \end{aligned} \tag{1}$$

By Assumption 1(i), $c(x) \in (0, 1)$. With the above notation, it is easy to see that $a^*(x) = 1$ if and only if $P(x) > c(x)$. That is,

$$a^*(x) = \text{sign}\left[P(x) - c(x)\right], \tag{2}$$

where $\text{sign}(z) = 1$ for $z > 0$ and $\text{sign}(z) = -1$ for $z < 0$. The optimal action involves thresholding the conditional probability $P(x)$ with a covariate-dependent cutoff $c(x)$.

The decision-maker is assumed to know the payoff function $U(a, y, x)$ and, hence, the cutoff function $c(x)$. Given the sample $\{(X_i, Y_i)\}_{i=1}^n$, the decision-maker only needs to learn $P(x)$ in order to implement the optimal strategy defined by $a^*(x)$. The process through which the decision maker arrives at this payoff function is not the focus of the paper, as it is often highly context-dependent. In some cases, the payoff function may be derived from historical data, expert judgment, or other domain-specific considerations.

To clarify this further, consider an illustrative example: suppose a bank needs to decide whether to extend a loan to an applicant, where uncertainty exists over whether the loan will be repaid. In this case, the bank's utility function can be represented by the net present value (NPV) associated with each combination of action and outcome. Specifically, when the bank approves the loan and the loan is repaid (i.e., the borrower does not default), the NPV depends on several factors, such as the loan amount, loan duration, interest rate, and the discount rate. If the loan defaults, the NPV still depends on these same factors, but it also incorporates additional elements such as the time of default and the recovery amount, which may depend on the applicant's characteristics (e.g., creditworthiness, collateral, or assets available for recovery). In the event that the bank rejects the loan application, the NPV would be zero, reflecting that no loan is issued and, therefore, no return is gained or lost.

In this example, the decision maker's utility function is profit-driven and calculated based on the NPV, which depends on both the applicant's characteristics and the loan terms. All relevant factors, along with additional factors that can help predict the loan outcome, are incorporated into the covariate X. In this paper, we abstract away the processes of covariate selection and utility determination, both of which are highly domain-specific, and focus on the problem of binary decision-making for a given utility function.

2.2 Utility Maximizing Actions

Suppose the decision-maker chooses a proxy $m(x;\theta)$ for $P(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ (with d_θ being the dimension of θ), and uses a decision rule of the form:

$$a(x,\theta) = \text{sign}\left[m(x,\theta) - c(x)\right].$$

Then, the best a of the above form is given by

$$a(x,\theta^*) = \text{sign}\left[m(x,\theta^*) - c(x)\right],$$

where

$$\theta^* \in \arg\min_{\theta\in\Theta} E\left[\psi(Y,X)1\left\{a(X,\theta) \neq Y\right\}\right].$$

To implement $a(\cdot,\theta^*)$, the decision-maker can solve the sample version of the above problem:

$$\begin{aligned}\hat{\theta} &\in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i,X_i)1\left\{a(X_i,\theta) \neq Y_i\right\} \\ &= \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} U(a(X_i,\theta),Y_i,X_i),\end{aligned}$$

and then take the action according to

$$a(x,\hat{\theta}) = \text{sign}[m(x,\hat{\theta}) - c(x)].^3$$

The above M-estimator $\hat{\theta}$ is motivated by utility maximization, and we will refer to it as the maximum utility estimator. To highlight the method behind the estimator, we may write it as $\hat{\theta}_{\text{MU}}$. The MU estimator minimizes the empirical average loss from making incorrect decisions. Intuitively, when $a(X_i,\theta) \neq Y_i$, an incorrect decision is made, and the decision-maker incurs a loss of $\psi(Y_i,X_i)$. The MU criterion function is the average of the losses over the sample. The corresponding population criterion function is

$$Q_{\text{MU}}(\theta) = E\left[\psi(Y,X)1\left\{a(X,\theta) \neq Y\right\}\right].$$

It can be shown that the above MU estimator is the same as the estimator of Elliott and Lieli (2013). In the special case where the loss $\psi(y,x)$ does not depend on either y or x, we have $c(x) = 1/2$, and the MU estimator becomes

$$\hat{\theta}_{\text{MU}} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} 1\left\{\text{sign}[m(X_i,\theta) - 0.5] \neq Y_i\right\}.$$

[3] If we focus on the decision at a particular value, say x_o, of X, we can solve a local version of the problem:

$$\hat{\theta}(x_o) \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} U(a(X_i,\theta),Y_i,X_i)k_h(X_i,x_o),$$

where $k_h(\cdot,\cdot)$ is a kernel weighting function with tuning parameter h. Weighting is also necessary when the sample of covariates $\{X_i\}_{i=1}^n$ comes from a different population or subpopulation than the target population for which we are making predictions and taking actions. In this case, we can solve

$$\hat{\theta}_\omega \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} U(a(X_i,\theta),Y_i,X_i)\omega(X_i),$$

where the weighting function $\omega(\cdot)$ is used to reweight the sample so that it matches the target population of interest. In this paper, we allow $m(x,\theta)$ to take a flexible form and defer the additional complication of (local) weighting to future research.

Hence, the MU estimator reduces to the maximum score estimator of Manski (1975, 1985).

The MU estimator is clearly different from the maximum likelihood (ML) estimator. Suppose $m(X_i, \theta) \in (0, 1)$ for all $\theta \in \Theta$.[4] The ML estimator is defined as

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Y_i + 1}{2} \log[m(X_i, \theta)] + \left(1 - \frac{Y_i + 1}{2}\right) \log\left[1 - m\left(X_i, \theta\right)\right] \right\}.$$

The criterion function underlying the ML estimator is motivated by statistical considerations, without accounting for the payoff differences across different actions and states of the world.

While the ML method has been extensively studied, the MU method has received less attention. In particular, model selection within the MU framework has not been thoroughly explored in the literature. To address this gap, Su (2021) considers a penalized MU estimator, where an additive penalty regularizes the complexity of the model class. This model selection approach is similar to AIC and BIC in the likelihood framework, but the penalty is based on a complexity measure, such as the Vapnik–Chervonenkis (VC) dimension, of the model class. There is a large body of statistical literature on model selection via complexity regularization; see, for example, Koltchinskii (2001), Bartlett et al. (2002), and Massart (2007). Su (2021) is an application of this method to the MU framework.

While the MU approach targets directly the expected payoff, it does not account for the margin of a covariate vector from the underlying decision boundary, and it does not work well when the VC dimension is infinite, which often occurs when the decision boundary is allowed to reside in an RKHS. Our proposed approach overcomes these drawbacks. To the best of our knowledge, this paper is the first to study the margin within the MU framework and to consider margin-maximizing decision rules in this context. Another limitation of the MU approach is that, due to the presence of an indicator function in its criterion function, the underlying optimization problem is NP-hard and thus computationally challenging. In contrast, our approach is computationally efficient, making it an attractive alternative.

3 Support-vector Decision Making

In this section, we introduce the concept of the margin and choose model parameters to maximize it. This method addresses some of the limitations of the MU approach.

3.1 Complete Separation

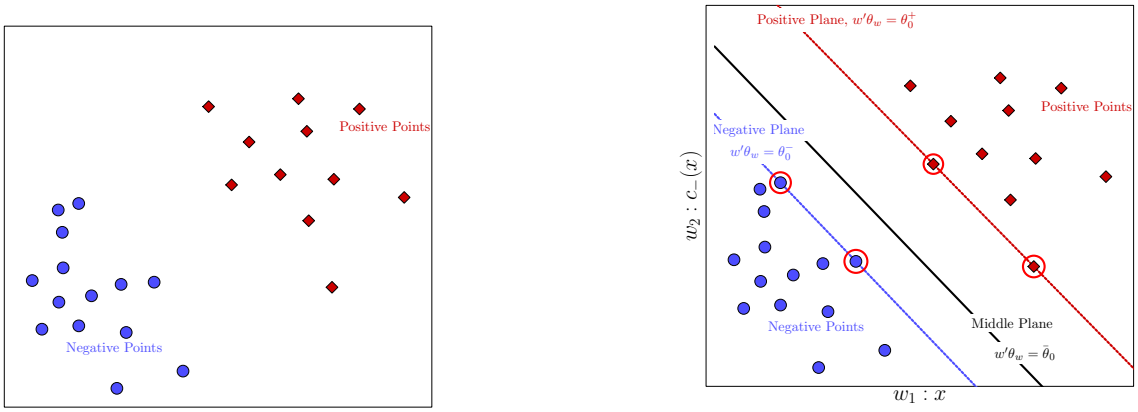
We start with the case of complete separation, where a hyperplane can completely separate the two classes of points. Figure 2a provides a visual representation of this scenario.[5] While complete separation may not be realistic in practice, it provides a useful starting point for introducing the main concepts and ideas. We will address the case of incomplete separation in Section 3.2.

Due to the presence of the covariate-specific cutoff function $c(\cdot)$, the standard SVM cannot be applied directly. To address this, we define

$$c_{-}\left(x\right) := -c(x)$$

[4] If this is not the case, we can apply a transformation, such as the logistic transformation, to ensure that the transformed version falls within the range of $(0, 1)$.

[5] The figure is provided for illustrative purposes only. The specific space in which complete separation is achieved will be defined shortly.



(a) Completely separable data.

(b) Three planes in $\mathbb{R}^{d_w}$.

Figure 2: Completely separable data and three planes.

and view $c_-(X)$ as a separate and special covariate whose coefficient is restricted to be 1. Because of this constraint, we still cannot rely on the standard SVM arguments to obtain the support-vector decision directly. Instead, we develop our method from first principles.

Letting

$$W_i = \begin{pmatrix} X_i \\ c_-(X_i) \end{pmatrix},$$

we map each point $X_i \in \mathbb{R}^{d_x}$ into a point $W_i \in \mathbb{R}^{d_w}$, where $d_w = d_x + 1$ is the number of elements in W_i. Under this mapping, the original sample $\{(X_i, Y_i)\}_{i=1}^n$ is effectively transformed into a new sample $\{(W_i, Y_i)\}_{i=1}^n$. We now focus on the set of points $\left\{W_i \in \mathbb{R}^{d_w}\right\}$ in a higher-dimensional space, along with their "labels" $\{Y_i\}$, and refer to W_i as a vector of generalized attributes. In this subsection, we assume that complete separation is achieved in $\mathbb{R}^{d_w}$.

Denote

$$w = \begin{pmatrix} x \\ c_-(x) \end{pmatrix} \text{ and } \theta_w = \begin{pmatrix} \theta_x \\ \theta_c \end{pmatrix} \text{ for } \theta_c = 1.$$

Under the assumption of complete separation in $\mathbb{R}^{d_w}$, there exist θ_0^+, θ_0^-, and θ_x with $\theta_0^+ > \theta_0^-$ such that the two parallel hyperplanes $w'\theta_w = \theta_0^+$ and $w'\theta_w = \theta_0^-$ in $\mathbb{R}^{d_w}$ completely separate the "positive" points $\{W_i : Y_i = +1\}$ from the "negative" points $\{W_i : Y_i = -1\}$.

For a given θ_w, we choose θ_0^+ to be as large as possible, subject to the constraint that $W_i'\theta_w \geq \theta_0^+$ for all i such that $Y_i = +1$. This constraint requires that all positive points lie on or above the "positive" hyperplane $w'\theta_w = \theta_0^+$. Similarly, for a given θ_w, we choose θ_0^- to be as small as possible, subject to the constraint that $W_i'\theta_w \leq \theta_0^-$ for all i such that $Y_i = -1$. This constraint requires that all negative points lie on or below the "negative" hyperplane $w'\theta_w = \theta_0^-$.[6]

Based on these two hyperplanes, we define the "middle" hyperplane:

$$w'\theta_w = \bar{\theta}_0 \text{ for } \bar{\theta}_0 = \rho\theta_0^+ + (1-\rho)\,\theta_0^-,$$

where $\rho \in (0,1)$. The "middle" hyperplane will serve as the *decision boundary.* For an out-of-sample point x with $w = (x', c_-(x)')'$, we take the action according to whether the point w is

[6]Without loss of generality, we assume that the points with $Y_i = +1$ lie above the positive hyperplane and the points with $Y_i = -1$ lie below the negative hyperplane. If this is not the case, we can simply switch the labels.

above or below the middle hyperplane, that is,

$$a(x) = \text{sign}\left[w'\theta_w - \bar{\theta}_0\right].$$

See Figure 2b for an illustration of the three hyperplanes when $d_x = 1$, $d_w = 2$ so that $w_1 = x$ and $w_2 = c_-(x)$.[7]

The choice of ρ is influenced by a number of factors, such as the relative costs of false positives and false negatives, as well as the distribution of positive and negative points, especially near the theoretically optimal decision boundary. If the cost of a false positive is generally higher than that of a false negative and there are more negative points than positive points, then, as a rule of thumb, we might choose ρ to be greater than 1/2, so that the "middle" hyperplane is closer to the positive hyperplane. This would reduce the likelihood of making false positive decisions. Otherwise, we might choose ρ to be less than 1/2. A default choice could be $\rho = 1/2$, in which case the "middle" hyperplane is equidistant from the positive and negative hyperplanes. To avoid introducing new terminology, we will use the term "middle hyperplane," even if the hyperplane is not exactly in the middle of the positive and negative hyperplanes.

The following lemma characterizes important distances that will be used throughout the remainder of the paper. The proof can be found in Supplementary Appendix S.1.

Lemma 1 *Let* $\|\theta_w\| = \sqrt{\theta_c^2 + \|\theta_x\|^2}$. *Then:*

(i) the geometric distance from a point W_i *to the middle hyperplane* $w'\theta_w = \bar{\theta}_0$ *is* $d_i \equiv d_i\left(\theta_w, \bar{\theta}_0\right) = Y_i\left(W_i'\theta_w - \bar{\theta}_0\right)/\|\theta_w\|$;

(ii) the geometric distance between the positive and negative hyperplanes is $\left(\theta_0^+ - \theta_0^-\right)/\|\theta_w\|$.

In the case of complete separation, the geometric distance between the positive and negative hyperplanes, as given in Lemma 1, is referred to as the hard margin in the SVM literature, as there is no point lying between these two hyperplanes. The positive and negative hyperplanes are then referred to as the *margin boundaries*, and the middle hyperplane is called the *decision boundary*.

A maximum margin decision rule seeks the values of θ_0^+, θ_0^-, and θ_x to maximize the margin. That is, it solves

$$\begin{aligned}
&\max_{\theta_0^+, \theta_0^-, \theta_x} \frac{\theta_0^+ - \theta_0^-}{\sqrt{\theta_c^2 + \|\theta_x\|^2}} \text{ subject to} \\
&X_i'\theta_x + c_-(X_i)\,\theta_c \le \theta_0^- \text{ for all } i \text{ with } Y_i = -1, \\
&X_i'\theta_x + c_-(X_i)\,\theta_c \ge \theta_0^+ \text{ for all } i \text{ with } Y_i = +1, \\
&\theta_0^+ - \theta_0^- > 0.
\end{aligned} \tag{3}$$

To obtain an alternative yet equivalent representation of the above problem, we define

$$\begin{aligned}
\theta_0^\Delta &= \min(\rho, 1-\rho)\left(\theta_0^+ - \theta_0^-\right), \\
q^+ &= \frac{1-\rho}{\min(\rho, 1-\rho)},\ q^- = \frac{\rho}{\min(\rho, 1-\rho)},
\end{aligned}$$

and

$$q_i = q^+ \cdot 1\{Y_i = +1\} + q^- \cdot 1\{Y_i = -1\}.$$

[7] In the figure, we use the term "plane" instead of "hyperplane," and we will use these two terms interchangeably.

If $\rho \leq 1/2$, then $q^- = 1$. If $\rho \geq 1/2$, then $q^+ = 1$. Note that $q_i \geq 1$ for all $i \in [n] := \{1, 2, \ldots, n\}$. With these definitions of $\bar{\theta}_0, \theta_0^\Delta$, and q_i, the separation constraints in (3) can be written compactly as

$$Y_i \left(X_i' \theta_x + c_- (X_i) \theta_c - \bar{\theta}_0 \right) \geq q_i \theta_0^\Delta \text{ for all } i \in [n].$$

Define the normalized parameters:

$$\kappa_x = \frac{\theta_x}{\theta_0^\Delta}, \kappa_c = \frac{\theta_c}{\theta_0^\Delta} \geq 0, \kappa_0 = -\frac{\bar{\theta}_0}{\theta_0^\Delta}, \text{ and } \kappa_w = (\kappa_x', \kappa_c')'.$$

The separation requirement then becomes

$$Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] \geq q_i \text{ for all } i \in [n].$$

Note that when $\rho \neq 1/2$, for one class of points, $q_i = 1$, and for the other class of points, $q_i > 1$. We have, therefore, effectively normalized the smaller of the two lower bounds to be 1. The maximization problem in (3) is then transformed into the following minimization problem:

$$\begin{aligned} &\min_{\kappa_0, \kappa_x, \kappa_c} \frac{1}{2} (\|\kappa_x\|^2 + \kappa_c^2) \text{ subject to} \\ &Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] \geq q_i \text{ for all } i \in [n], \\ &\kappa_c \geq 0. \end{aligned} \tag{4}$$

This is a standard quadratic programming problem and can be easily solved using commonly available software packages.

The minimization problem resembles the minimization problem in the standard SVM, but there are three key differences. First, we have included the cutoff $c(\cdot)$ as a separate covariate, and this constitutes a conceptual innovation. Second, there is a sign restriction $\kappa_c \geq 0$, which is not present in the standard SVM. Third, unlike the standard SVM, our method allows for $q^+ \neq q^-$, so the middle hyperplane may not be equidistant from the positive and negative hyperplanes. While the standard SVM requires $Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] \geq 1$ for all observations, we require $Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] \geq q_i$ for a class-specific lower bound q_i. Therefore, even if there is no sign restriction, our minimization problem is more general than the standard SVM minimization problem.

For the examples in Figure 1, our minimization problem will favor a_1 over a_2. For Figure 1a, the decision boundary associated with a_1 is closer to the maximum margin decision boundary. For Figure 1b, the margin associated with a_1 is larger than that associated with a_2.

To develop the dual to the minimization problem in (4), we form the Lagrangian:

$$L_P(\kappa; \lambda, \lambda_c) = \frac{1}{2} \left(\|\kappa_x\|^2 + \kappa_c^2 \right) - \sum_{i=1}^n \lambda_i \left\{ Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] - q_i \right\} - \lambda_c \kappa_c,$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)'$ and λ_c are the Lagrangian multipliers, all of which are nonnegative. In the above, the subscript "P" stands for "primal". The Karush-Kuhn-Tucker (KKT) stationary and complementary slackness conditions are

$$\kappa_x = \sum_{i=1}^n \lambda_i Y_i X_i, \tag{5}$$

$$\kappa_c = \sum_{i=1}^n \lambda_i Y_i c_-(X_i) + \lambda_c, \tag{6}$$

$$0 = \lambda_i \left\{ Y_i \left[\kappa_0 + X_i' \kappa_x + c_- (X_i) \kappa_c \right] - q_i \right\}, \tag{7}$$

$\sum_{i=1}^{n} \lambda_i Y_i = 0$, and $\lambda_c \kappa_c = 0$.

Define $\hat{\kappa}_w = \left(\hat{\kappa}_x', \hat{\kappa}_c\right)'$ and $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)'$. Let $(\hat{\kappa}_0, \hat{\kappa}_w, \hat{\lambda}, \hat{\lambda}_c)$ be a solution to the full set of KKT conditions. From the KKT dual complementarity in (7), $\hat{\lambda}_i > 0$ implies that $Y_i \left[\hat{\kappa}_0 + W_i' \hat{\kappa}_w\right] = q_i$. Define

$$S_+ = \{i : \hat{\lambda}_i > 0\}. \tag{8}$$

It follows from (5) and (6) that the solution satisfies

$$\begin{aligned} \hat{\kappa}_x &= \sum_{i \in S_+} \hat{\lambda}_i Y_i X_i, \\ \hat{\kappa}_c &= \sum_{i \in S_+} \hat{\lambda}_i Y_i c_- (X_i) + \hat{\lambda}_c. \end{aligned} \tag{9}$$

The above equations show that only the subsample $\{(W_i, Y_i) : i \in S_+\}$ will ultimately determine $\hat{\kappa}_w$. We call this the *support subsample*, as it supports the positive and negative hyperplanes in the sense that it determines the locations and orientations of these hyperplanes. We refer to the corresponding attributes $\{W_i : i \in S_+\}$ as the support vectors. Each support vector lies on either the positive or negative hyperplane: $\hat{\kappa}_0 + w' \hat{\kappa}_w = q^+$ or $\hat{\kappa}_0 + w' \hat{\kappa}_w = -q^-$. As an illustration, the circled points in Figure 2b are the support vectors.[8]

Once the support subsample has been identified, the "non-support" subsample $\{(W_i, Y_i) : i \notin S_+\}$ can be discarded without altering the decision boundary. However, it would be erroneous to conclude that the non-support subsample is irrelevant to the decision problem, since the set of the support subsample is jointly and collectively determined by the whole sample.

Plugging (5) and (6) into $L_P(\kappa; \lambda, \lambda_c)$, we obtain the dual problem:

$$\begin{aligned} \max_{\lambda, \lambda_c} L_D(\lambda, \lambda_c) &= \sum_{i=1}^{n} \lambda_i q_i - \frac{1}{2} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j W_i' W_j + 2\lambda_c \sum_{i=1}^{n} \lambda_i Y_i c_- (X_i) + \lambda_c^2 \right] \\ \text{s.t. } & \sum_{i=1}^{n} \lambda_i Y_i = 0, \ \lambda_i \geq 0, \ \text{for all } i \in [n], \ \text{and } \lambda_c \geq 0. \end{aligned} \tag{10}$$

The dual problem is another quadratic programming problem.

The dual problem reveals that the points $\{W_i\}$ interact with each other only via their cross product $W_i' W_j$, which can be written as the Euclidean inner product in $\mathbb{R}^{d_w}$: $W_i' W_j = \langle W_i, W_j \rangle$. This opens the door to possible generalizations when other inner products are used. We consider such an extension in Section 4.

Once we find the solution $(\hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\lambda}_c)'$ to the dual problem, we can plug it into (9) to obtain $\hat{\kappa}_x$ and $\hat{\kappa}_c$. To find $\hat{\kappa}_0$, we note that for $i \in S_+$, we have $Y_i \left(\hat{\kappa}_0 + W_i' \hat{\kappa}_w\right) = q_i$ and so $\hat{\kappa}_0 + W_i' \hat{\kappa}_w = q_i Y_i$ as $Y_i^2 = 1$. Thus, we can recover $\hat{\kappa}_0$ by taking an average of $\{q_i Y_i - W_i' \hat{\kappa}_w : i \in S_+\}$, leading to

$$\hat{\kappa}_0 = \frac{1}{|S_+|} \sum_{i \in S_+} \left(q_i Y_i - W_i' \hat{\kappa}_w\right),$$

[8] Let $S = \{i : Y_i \left(W_i' \hat{\kappa}_w + \hat{\kappa}_0\right) = q_i\}$. According to the definition of support vectors given here, not every point in $\{W_i : i \in S\}$ is a support vector, but for easy interpretation and geometric intuition, we may refer to all points in $\{W_i : i \in S\}$ as support vectors. The difference in the definitions of support vectors does not affect our theoretical formulation. However, when $\hat{\lambda}_i$ is available from the dual problem, we take the sum over $i \in S^+$, as given in (9), to compute $\hat{\kappa}_w$. We may call $\{W_i : i \in S\}$ the set of geometric support vectors and $\left\{W_i : i \in S^+\right\}$ the set of computational support vectors. What we illustrate in Figure 2b are the geometric support vectors.

where $|S_+|$ denotes the number of elements in the set S_+.[9] This shows that a solution to the dual problem completely determines a corresponding solution to the primal problem.

For an out-of-sample point x with $w = (x', c_-(x))'$, we take the action according to the empirical optimal decision boundary:

$$\begin{aligned}\hat{a}(x) &= \text{sign}(w'\hat{\kappa}_w + \hat{\kappa}_0) \\ &= \text{sign}\left\{\sum_{i=1}^{n} x'\left[\hat{\lambda}_i Y_i X_i\right] + c_-(x)\left[\sum_{i=1}^{n}\hat{\lambda}_i Y_i c_-(X_i) + \hat{\lambda}_c\right] + \hat{\kappa}_0\right\} \\ &= \text{sign}\left\{\sum_{i\in S_+}\hat{\lambda}_i Y_i\left[\langle x, X_i\rangle + c_-(x)c_-\left(X_i\right)\right] + \hat{\lambda}_c c_-(x) + \hat{\kappa}_0\right\},\end{aligned}$$

where $\langle x, X_i\rangle$ is the Euclidean inner product in $\mathbb{R}^{d_x}$. In the last line above, we have reduced the summation over the whole sample to a summation over only the subsample corresponding to the support vectors. Since the number of support vectors, namely, the size of the set S_+, can be much smaller than the sample size, such a reduction can yield significant computational savings, especially in settings where the sample size n is much larger than the number of support vectors.

It is now clear that our decision rule is fully characterized by the support vectors $\{W_i : i \in S_+\}$ and their labels $\{Y_i : i \in S_+\}$. Our approach can therefore be called support-vector decision making. In a nutshell, the support vectors are jointly determined by all observations and may be regarded as "representatives" for $\{W_i : i \in [n]\}$. Once the support vectors and their λ_i's are given, the subsequent decision for an out-of-sample point x can be made based solely on how w is related to the support vectors $\{W_i : i \in S_+\}$.

Under complete separation, the utility function plays the important role of supplying an additional attribute $c_-(x)$ to the decision-making problem. In the next subsection, we will see that under incomplete separation, the utility function also changes our decision-making problem in other significant ways.

3.2 Incomplete Separation

In this subsection, we consider the case of incomplete separation, which is more realistic in economic applications. See Figure 3 for a visual illustration of inseparable data. For now, we maintain a linear decision boundary, though this constraint will be relaxed in Section 4.

3.2.1 The Primal Problem

In the presence of incomplete separation, it is not possible to find two parallel hyperplanes that completely separate the two classes. Nevertheless, we still stipulate a "positive" hyperplane: $\kappa_0 + w'\kappa_w = q^+$ and a "negative" hyperplane: $\kappa_0 + w'\kappa_w = -q^-$. We relax the separation constraint $Y_i(\kappa_0 + W_i'\kappa_w) \geq q_i$ by introducing the smallest possible "slack" variable $\xi_i \geq 0$ such that

$$Y_i\left(\kappa_0 + W_i'\kappa_w\right) + \xi_i \geq q_i,$$

[9] Here, we have implicitly used the result that S^+ is not empty. We can prove this by contradiction. If S^+ is empty, then $\hat{\lambda}_i = 0$ for all $i \in [n]$, from which we can deduce that $\hat{\kappa}_x = 0$ and $\hat{\kappa}_c = 0$. However, under this choice of $(\hat{\kappa}_x, \hat{\kappa}_c)$, the primal feasibility constraints $Y_i[\kappa_0 + X_i'\kappa_x + c_-(X_i)\kappa_c] \geq q_i$ for all $i \in [n]$ become $Y_i\kappa_0 \geq q_i$ for all $i \in [n]$. Since Y_i's do not have the same sign across $i \in [n]$, $Y_i\kappa_0 \geq q_i$ cannot hold for all $i \in [n]$, leading to a contradiction.

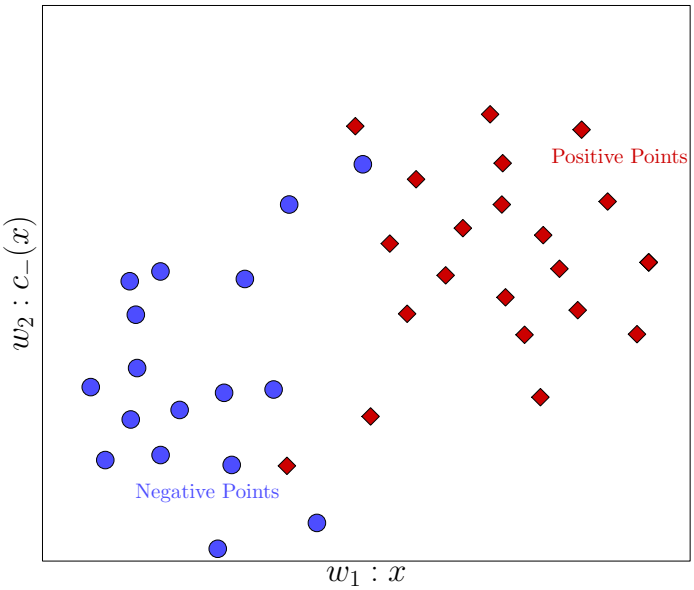


Figure 3: Inseparable data.

for all $i \in [n]$. By definition, $\xi_i = [q_i - Y_i(\kappa_0 + W_i'\kappa_w)]_+$, where $\varsigma_+ \equiv \max(0, \varsigma)$. The slack variable ξ_i measures how far W_i lies on the wrong side of its associated hyperplane. By taking the magnitude of ξ_i into consideration, we account for the distance of a point to the decision boundary. Our method is thus distinctly different from the MU method, where such distances are completely ignored.

When a point lies on the wrong side of the middle hyperplane (i.e., $\kappa_0 + w'\kappa_w = 0$), we have $Y_i(\kappa_0 + W_i'\kappa_w) < 0$ and thus $\xi_i > q_i$. Since $q_i \geq 1$ by definition, it follows that $\xi_i \geq 1$ for all points on the wrong side of the middle hyperplane. Therefore, $n^{-1}\sum_{i=1}^n \xi_i$ is an upper bound for the error rate, i.e., the fraction of false decisions.

For each false decision, we incur a loss $\psi(Y_i, X_i)$. A loss-weighted version of the error rate can be defined as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\psi(Y_i, X_i)}{\bar{\psi}}\xi_i,$$

where $\bar{\psi} = n^{-1}\sum_{i=1}^n \psi(Y_i, X_i)$. This loss-weighted error rate better reflects our ultimate objective. We are concerned not only with the direction and margin of decisions but also with the losses associated with incorrect decisions, which may vary across individual observations.

By Lemma 1, the geometric distance between the negative and positive hyperplanes (i.e., $\kappa_0 + w'\kappa_w = q^+$ and $\kappa_0 + w'\kappa_w = -q^-$) is $(q^+ + q^-)/\|\kappa_w\| = 1/[\min(\rho, 1-\rho)\|\kappa_w\|]$. In seeking a maximum margin decision boundary, we maximize the (soft) margin (equivalently, minimize the squared reciprocal of the soft margin) subject to the control of the loss-weighted error rate:

$$\begin{aligned}
&\min_{\kappa_0, \kappa_w, \xi} \|\kappa_w\|^2 \text{ subject to} \\
&\frac{1}{n}\sum_{i=1}^{n}\frac{\psi(Y_i, X_i)}{\bar{\psi}}\xi_i \leq \mathcal{B}, \\
&Y_i\left(\kappa_0 + W_i'\kappa_w\right) + \xi_i \geq q_i,\ \xi_i \geq 0, \text{ for } i \in [n] \text{ and } \kappa_c \geq 0,
\end{aligned}$$

where $\xi = (\xi_1, \ldots, \xi_n)'$ and $\mathcal{B}$ is a user-chosen upper bound for the loss-weighted error rate. With an appropriately chosen μ, the above problem is equivalent to

$$\min_{\kappa_0, \kappa_w, \xi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\psi(Y_i, X_i)}{\bar{\psi}} \xi_i + \frac{\mu}{2} \|\kappa_w\|^2 \right\} \text{ subject to}$$
$$Y_i \left(\kappa_0 + W_i' \kappa_w\right) + \xi_i \geq q_i,\ \xi_i \geq 0, \text{ for } i \in [n], \text{ and } \kappa_c \geq 0. \tag{11}$$

The minimization problem resembles the standard SVM problem, and we will call it the Maximum Utility SVM (MU-SVM). However, there are four key differences. First, as a conceptual innovation, we allow the cutoff function $c(\cdot)$ to directly affect the decision boundary. Second, the MU-SVM imposes a sign restriction $\kappa_c \geq 0$. Third, since q^+ may not equal q^-, the MU-SVM allows for different margins for different classes of observations. Finally, the MU-SVM accounts for different losses $\psi(Y_i, X_i)$ for different points lying on the wrong side of the positive or negative hyperplane (i.e., for points with $\xi_i > 0$). In effect, the standard SVM problem assumes that $\psi(\cdot, \cdot)$ is a constant function and imposes the same loss for all incorrect decisions. In contrast, we allow the loss to be heterogeneous across observations.

In the minimization problem given in (11), the parameter μ is a regularization parameter that balances the size of the margin and the loss-weighted error rate. Note that the margin is given by $(q^+ + q^-) / \|\kappa_w\|$. A larger value of μ encourages a smaller $\|\kappa_w\|^2$ and hence a larger margin, and it may also give rise to larger slack variables and a higher loss-weighted error rate. However, a larger margin often leads to a smaller generalization error; see Section 5.2 for a theoretical analysis.

As an equivalent formulation, the MU-SVM minimizes

$$Q_{n,\text{MU-SVM}}(\kappa) = \frac{1}{n} \sum_{i=1}^{n} \left(1\{Y_i = 1\} \left[q^+ - V_i\right]_+ + 1\{Y_i = -1\} \left[q^- - V_i\right]_+ \right) \psi(Y_i, X_i) + \frac{\mu \bar{\psi}}{2} \|\kappa_w\|^2, \tag{12}$$

where $V_i := Y_i (\kappa_0 + W_i' \kappa_w)$. Using the same notation for the variables to optimize over, the standard MU (e.g., Elliott and Lieli (2013)) minimizes

$$Q_{n,\text{MU}}(\kappa) = \frac{1}{n} \sum_{i=1}^{n} 1\{V_i \leq 0\} \psi(Y_i, X_i).$$

Comparing the two criterion functions, we observe two key differences. First, in the standard MU criterion function, the loss is weighted by the zero-one function $1\{V_i \leq 0\}$, whereas in the MU-SVM criterion function, the loss is weighted by the hinge function $(q^+ - V_i)_+$ or $(q^- - V_i)_+$. See Figure 4 for an illustration of $(1 - V)_+$ compared to the zero-one loss. The hinge function $(1 - V)_+$ is a convex function that lies above the zero-one function and can be viewed as a convex surrogate for it. Using the hinge loss function simplifies the optimization problem, reducing it to a quadratic programming problem and avoiding the need for complex algorithms such as simulated annealing. Moreover, employing the hinge function enables us to account for the margin in our decision problem.

Second, the MU-SVM criterion function includes an additional term to regularize the margin of the decision rule. From a modern perspective, this term is simply a regularization term that serves to control the size of the coefficients. In the present setting, the regularizer has a margin interpretation, which may not be readily available in a general regularization problem. Note that, for the standard MU approach, adding regularization on the size of the coefficients has no effect on

the decision rule, as the zero-one loss is scale-invariant in the sense that $1\{V \leq 0\} = 1\{\varsigma V \leq 0\}$ for any constant $\varsigma > 0$.

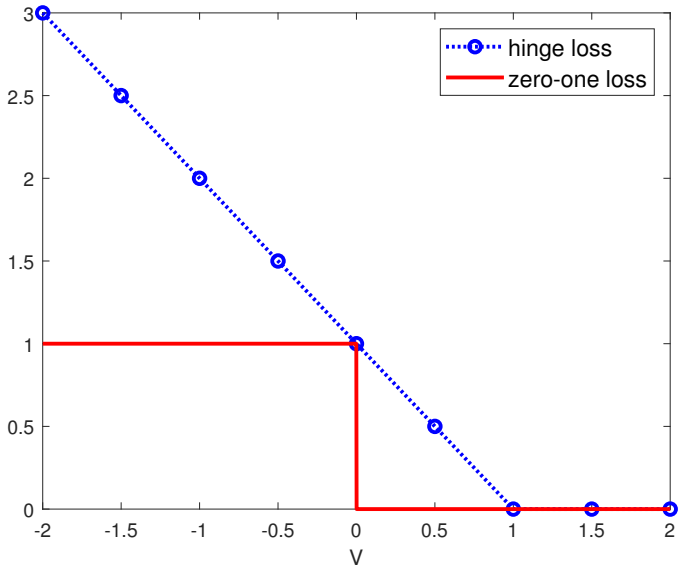


Figure 4: The hinge loss $(1-V)_+$ and zero-one loss $1\{V \leq 0\}$.

Instead of using hinge loss, non-convex loss functions, such as ramp loss (Collobert et al. (2006)) or smoothed ramp loss (Zhou et al. (2017)), can be employed within the maximum utility framework. However, these non-convex losses introduce increased computational complexity due to the potential for local minima. While they may be less sensitive to outliers, we have opted for the modified hinge loss in this paper because its convexity ensures simpler optimization and a unique global solution.

3.2.2 The Dual Problem

We now derive the dual problem to the MU-SVM given in (11). The Lagrangian function is

$$L_P(\kappa, \xi; \lambda, \lambda_c, r) = \frac{1}{n}\sum_{i=1}^{n} \xi_i \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\kappa_w\|^2 - \sum_{i=1}^{n} \lambda_i \left\{Y_i\left[\kappa_0 + W_i'\kappa_w\right] + \xi_i - q_i\right\} - \sum_{i=1}^{n} r_i\xi_i - \lambda_c\kappa_c, \tag{13}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n) \geq 0$, $\lambda_c \geq 0$, and $r = (r_1, \ldots, r_n) \geq 0$ are the Lagrangian multipliers for the three sets of inequality constraints. The KKT conditions are

$$\begin{aligned}
\kappa_x &= \frac{1}{\mu}\left(\sum_{i=1}^{n} \lambda_i Y_i X_i\right), \\
\kappa_c &= \frac{1}{\mu}\left(\sum_{i=1}^{n} \lambda_i Y_i c_-(X_i) + \lambda_c\right), \\
&\sum_{i=1}^{n} \lambda_i Y_i = 0, \\
&\frac{1}{n}\frac{\psi(Y_i, X_i)}{\bar{\psi}} - \lambda_i - r_i = 0,
\end{aligned}$$

and

$$\begin{aligned}
&\lambda_i \left[Y_i\left(\kappa_0 + W_i'\kappa_w\right) + \xi_i - q_i\right] = 0,\ r_i\xi_i = 0,\ \lambda_c\kappa_c = 0, \\
&Y_i\left(\kappa_0 + W_i'\kappa_w\right) + \xi_i \geq q_i,\ \kappa_c \geq 0, \\
&\lambda_i \geq 0, \lambda_c \geq 0,\ r_i \geq 0.
\end{aligned}$$

The first block above consists of the stationarity conditions, and the second block consists of the complementarity conditions and the primal and dual feasibility conditions. A new KKT condition, not present in the case of complete separation, is

$$\frac{\psi(Y_i, X_i)}{n\bar{\psi}} - \lambda_i - r_i = 0.$$

We can interpret $\lambda_i + r_i$ as the shadow price of relaxing ξ_i, and this equation states that the shadow price equals the loss from such a relaxation.

Since the MU-SVM is a convex problem that satisfies Slater's condition, the above KKT conditions are necessary and sufficient for κ_0, κ_w, ξ to be a solution to the primal problem. Let $(\hat{\kappa}_0, \hat{\kappa}_w, \hat{\xi}, \hat{\lambda}, \hat{\lambda}_c, \hat{r})$, where $\hat{\kappa}_w = (\hat{\kappa}_x', \hat{\kappa}_c')'$, be a solution to the KKT conditions. From the KKT complementarity conditions, we can draw the following conclusions:

(i) If $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) > q_i$, then $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) + \hat{\xi}_i > q_i$, and so $\hat{\lambda}_i = 0$.

(ii) If $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) < q_i$, then $\hat{\xi}_i > 0$, and thus $\hat{r}_i = 0$, which, combined with the last stationarity condition, implies that $\hat{\lambda}_i = \psi(Y_i, X_i)/(n\bar{\psi})$.

(iii) If $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) = q_i$, then $0 \leq \hat{\lambda}_i \leq \psi(Y_i, X_i)/(n\bar{\psi})$.

Figure 5 illustrates the value of $\hat{\lambda}_i$ for positive points, which varies depending on whether they lie above, below, or on the positive hyperplane.

Define

$$S_+ = \left\{i \in [n] : \hat{\lambda}_i > 0\right\}.^{10}$$

It then follows from the stationarity conditions that

$$\begin{aligned}
\hat{\kappa}_x &= \frac{1}{\mu}\sum_{i \in S_+} \hat{\lambda}_i Y_i X_i, \\
\hat{\kappa}_c &= \frac{1}{\mu}\left(\sum_{i \in S_+} \hat{\lambda}_i Y_i c_-(X_i) + \hat{\lambda}_c\right).
\end{aligned} \tag{14}$$

[10] As in the case of complete separation, we can prove by contradiction that S_+ is not empty.

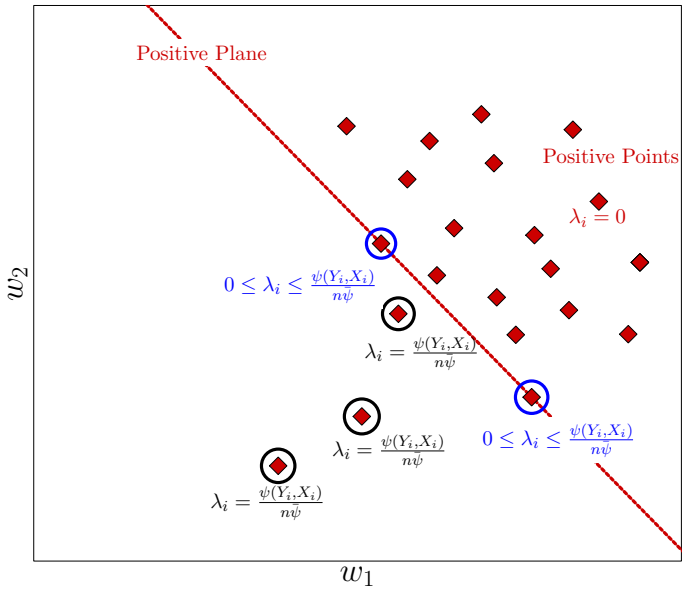


Figure 5: Lagrangian multipliers $\hat{\lambda}_i$ for positive points.

The above representations show that $\hat{\kappa}_x$ is a linear combination of $\{X_i : i \in S_+\}$, and $\hat{\kappa}_c$ is a linear combination of $\{c_-(X_i) : i \in S_+\}$, up to a constant adjustment. After the set S_+ is determined through the dual problem, only the subsample $\{(W_i, Y_i) : i \in S_+\}$ ultimately determines the decision boundary. We refer to this subsample as the support subsample, as it dictates the locations and orientations of the three hyperplanes.

To characterize the support subsample, we note that when $\hat{\lambda}_i > 0$, the constraint $Y_i(\kappa_0 + W_i'\kappa_w) + \xi_i \ge q_i$ in the primal problem holds with equality, yielding $Y_i(\kappa_0 + W_i'\kappa_w) + \xi_i = q_i$. Since $\xi_i \ge 0$, this implies that $Y_i(\kappa_0 + W_i'\kappa_w) \le q_i$ for all $i \in S_+$. Consequently, each positive (or negative) point in $\{W_i : i \in S_+\}$ lies on the positive (or negative) hyperplane, or on the wrong side of the respective hyperplane. We call $\{W_i : i \in S_+\}$ the set of support vectors. See Figure 6 for an illustration.

As in the case of complete separation, the number of support vectors can be much smaller than the number of observations. Numerical work based on the support subsample, rather than the whole sample, can yield substantial computational savings. As before, once the support subsample is identified, removing the non-support subsample *ex post* does not change the solution. However, we do not know in advance which observations belong to the support subsample. We have to use all observations together to determine the membership of the support subsample.

Using the KKT conditions to eliminate κ, ξ, and r in $L_P(\kappa, \xi; \lambda, \lambda_c, r)$, we obtain the criterion function for the dual problem:

$$L_D(\lambda, \lambda_c) = \sum_{i=1}^{n} \lambda_i q_i - \frac{1}{2\mu}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j W_i' W_j + 2\lambda_c \sum_{i=1}^{n} \lambda_i Y_i c_-(X_i) + \lambda_c^2\right].$$

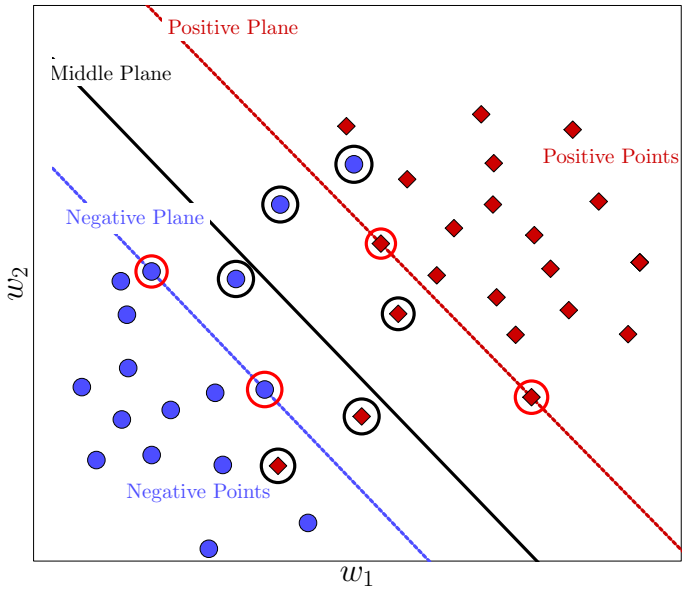


Figure 6: All circled data points are (geometric) support vectors.

For a given μ, the dual problem is then:

$$\begin{aligned}
&\max_{\lambda,\lambda_c} L_D(\lambda,\lambda_c) \text{ subject to} \\
&\sum_{i=1}^{n} \lambda_i Y_i = 0,\ \ \lambda_c \geq 0,\ \text{and} \\
&0 \leq \lambda_i \leq \frac{1}{n}\frac{\psi(Y_i,X_i)}{\bar{\psi}} \text{ for all } i \in [n]\,.
\end{aligned} \tag{15}$$

Like the primal problem, the dual problem is also a quadratic programming problem. The dual problem is almost identical to the one for the case with complete separation (cf. (10)). One difference is that the original constraint $\lambda_i \geq 0$ has now become $0 \leq \lambda_i \leq \psi(Y_i, X_i)/(n\bar{\psi})$. Another difference between incomplete separation and complete separation is that the quadratic term in (λ, λ_c) in the objective function has now been weighted by $1/\mu$. When $\mu = 1$, the objective function for the dual problem in (15) matches that for the complete separation case.

Let $(\hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\lambda}_c)$ be a solution to the dual problem. Then, we can recover the solution for κ_w to the primal problem using (14). To find the primal solution for κ_0, we use the fact that the inequality $0 < \hat{\lambda}_i < \psi(Y_i, X_i)/(n\bar{\psi})$ implies the equality $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) = q_i$, or equivalently, $\hat{\kappa}_0 = q_i Y_i - W_i'\hat{\kappa}_w$. Let

$$S_+^{\psi} = \left\{ i : 0 < \hat{\lambda}_i < \psi(Y_i, X_i)/\left(n\bar{\psi}\right) \right\},$$

which is a subset of the points on the positive and negative hyperplanes (the margin boundaries). If S_+^{ψ} is not empty, we can recover $\hat{\kappa}_0$ as follows:

$$\hat{\kappa}_0 = \frac{1}{|S_+^{\psi}|} \sum_{i \in S_+^{\psi}} \left(q_i Y_i - W_i'\hat{\kappa}_w \right). \tag{16}$$

If S_+^ψ is empty, we revert to the primal problem to determine $\hat{\kappa}_0$. More specifically, given $\hat{\kappa}_w$, we solve

$$\hat{\kappa}_0 \in \arg\min_{\kappa_0} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[q_i - Y_i \left(\kappa_0 + W_i' \hat{\kappa}_w \right) \right]_+ \psi\left(Y_i, X_i\right) \right\}.$$

This is a univariate convex minimization problem that can be solved efficiently. If multiple solutions exist, we select the one closest to the origin. Section S.3 in the supplementary appendix contains some discussion on the uniqueness of $\hat{\kappa}_w$ and $\hat{\kappa}_0$.

As in the case of complete separation, for an out-of-sample point x with $w = (x', c_-(x))'$, we take the action according to the estimated decision boundary:

$$\begin{aligned}
\hat{a}(x) &= \operatorname{sign}(w' \hat{\kappa}_w + \hat{\kappa}_0) \\
&= \operatorname{sign} \left\{ \frac{1}{\mu} \sum_{i \in S_+} \hat{\lambda}_i Y_i x' X_i + \frac{1}{\mu} c_-(x) \left[\sum_{i \in S_+} \hat{\lambda}_i Y_i c_-(X_i) + \hat{\lambda}_c \right] + \hat{\kappa}_0 \right\} \\
&= \operatorname{sign} \left\{ \sum_{i \in S_+} \frac{\hat{\lambda}_i}{\mu} Y_i \langle w, W_i \rangle + \frac{\hat{\lambda}_c}{\mu} c_-(x) + \hat{\kappa}_0 \right\}, \qquad (17)
\end{aligned}$$

where $\langle w, W_i \rangle$ is the usual Euclidean inner product in $\mathbb{R}^{d_w}$.

3.2.3 Interpretation of the Support-vector Decision

As in the case of complete separation, the action rule in (17) depends on the sample only through the support subsample. To facilitate interpretation, consider the case where $\hat{\lambda}_c = 0$ and $q_i = 1$ for all $i \in [n]$ as an example. To make the decision for an out-of-sample point x with $w = (x', c_-(x))'$, we first compute the Euclidean inner product $\langle w, W_i \rangle$ and obtain the score $Y_i \langle w, W_i \rangle$ for each support vector W_i. Then, we aggregate the scores according to the weighted formula:

$$\mathbb{S}(w) = \sum_{i \in S_+} \frac{\hat{\lambda}_i}{\mu} Y_i \langle w, W_i \rangle .$$

If w is in the same direction as W_i so that $\langle w, W_i \rangle > 0$ and $Y_i = +1$, then the score $Y_i \langle w, W_i \rangle$ will be positive, tipping the scale toward the positive action $a = +1$. If w is in the same direction as W_i but $Y_i = -1$, then the score $Y_i \langle w, W_i \rangle$ will be negative, tipping the scale toward the negative action $a = -1$. The same intuition applies to other scenarios when w is in the opposite direction to W_i.

The individual scores are weighted by $\{\hat{\lambda}_i/\mu, i \in S_+\}$. For a given μ, a support vector with higher loss $\psi(Y_i, X_i)$ tends to have a larger $\hat{\lambda}_i$ and hence receives a higher weight. This is quite reasonable, as we should assign relatively higher weights to observations with higher potential loss. The final decision rule is based on the sign of the aggregate score after a location adjustment of $\hat{\kappa}_0$, that is, $\operatorname{sign}\{\mathbb{S}(w) + \hat{\kappa}_0\}$.

If we view the scores as votes, the support vector decision rule can be regarded as a voting rule. Each support vector carries a weight of $\hat{\lambda}_i/\mu$ and casts a vote based on its outcome (Y_i) and similarity to the out-of-sample target point w measured by $\langle w, W_i \rangle$. The total vote, $\mathbb{S}(w)$, is a weighted sum of individual votes from the support vectors. When this voting rule is applied to

the *subset* of support vectors $\{W_i\colon\ i \in S_+^\psi\}$, we get

$$\mathbb{S}(W_i) = \sum_{j \in S_+} \frac{\hat{\lambda}_j}{\mu} Y_i \langle W_i, W_j \rangle \text{ for each } i \in S_+^\psi.$$

The average difference between Y_i and the vote $\mathbb{S}(W_i)$ over $i \in S_+^\psi$ is $(|S_+^\psi|)^{-1} \sum_{i \in S_+^\psi} [Y_i - \mathbb{S}(W_i)]$. This is just equal to $\hat{\kappa}_0$ defined in (16).

For an out-of-sample point w, let $Y(w)$ be the expected value of Y conditional on $W = w$. We expect the difference between $Y(w)$ and the vote $\mathbb{S}(w)$ it receives to be comparable to the average difference over $i \in S_+^\psi$:

$$Y(w) - \mathbb{S}(w) \approx \frac{1}{|S_+^\psi|} \sum_{i \in S_+^\psi} [Y_i(W_i) - \mathbb{S}(W_i)],$$

where $Y_i(W_i) := Y_i$. This resembles the parallel trend assumption in the difference-in-differences literature: the difference between $Y(w)$ and $\mathbb{S}(w)$ is expected to be the same as the difference between Y_i and $\mathbb{S}(W_i)$ averaged over $i \in S_+^\psi$, the subset of (computational) support vectors. It then follows that $Y(w) \approx \mathbb{S}(w) + \frac{1}{|S_+^\psi|} \sum_{i \in S_+^\psi} [Y_i - \mathbb{S}(W_i)]$. Hence, it is reasonable to predict the binary outcome by

$$\operatorname{sign}\left\{ \mathbb{S}(w) + \frac{1}{|S_+^\psi|} \sum_{i \in S_+^\psi} [Y_i - \mathbb{S}(W_i)] \right\} = \operatorname{sign}\left\{ \sum_{i \in S_+} \frac{\hat{\lambda}_i}{\mu} Y_i \langle w, W_i \rangle + \hat{\kappa}_0 \right\}.$$

This is exactly our support-vector decision rule for the case where $\hat{\lambda}_c = 0$ and $q_i = 1$ for all $i \in [n]$.

4 Series and Kernel Support Vector Decision Making

The previous section assumes a linear decision boundary. In this section, we introduce a more flexible approach by incorporating a nonlinear decision boundary using the widely adopted "kernel trick" from the machine learning literature.

4.1 Series Support Vector Decision Making

We first assume that the nonlinear decision boundary can be approximated by a series expansion. The proposed decision boundary is now:

$$\kappa_0 + \phi(x)' \kappa_\phi + c_-(x)\kappa_c,$$

where $\phi(x) = (\phi_1(x), \ldots, \phi_J(x))'$ and $\{\phi_j(\cdot) : j = 1, \ldots, J\}$ is a sequence of J basis functions. For example, we can take $J = 3$ and $\phi(x) = \left(x, x^2, x^3\right)'$. In the machine learning literature, $\phi(\cdot)$ is often referred to as the feature mapping, which maps "attributes" $x \in \mathbb{R}^{d_x}$ to "features" $\phi(x) \in \mathbb{R}^J$. Note that the number of features, J, may be much larger than the number of covariates, d_x. We assume that J is finite in this subsection and will consider the case of an infinite J in the next subsection.

Replacing $X_i'\kappa_x$ with $\phi(X_i)'\kappa_\phi$ in the MU-SVM given in (11), we obtain the following minimization problem:

$$\min_{\kappa_0,\kappa_\phi,\kappa_c,\xi}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\psi(Y_i,X_i)}{\bar{\psi}}\xi_i+\frac{\mu}{2}\left(\|\kappa_\phi\|^2+\kappa_c^2\right)\right\}\text{ subject to}$$
$$Y_i\left(\kappa_0+\phi(X_i)'\kappa_\phi+c_-(X_i)\kappa_c\right)+\xi_i\geq q_i,\ \xi_i\geq 0,\text{ for } i\in[n],\text{ and }\kappa_c\geq 0.$$

Similar to (15), the dual problem is

$$\max_{\lambda_1,\ldots,\lambda_n,\lambda_c} L_D(\lambda,\lambda_c)\text{ subject to}$$
$$\sum_{i=1}^{n}\lambda_i Y_i=0,\ \lambda_c\geq 0,\text{ and}$$
$$0\leq\lambda_i\leq\frac{1}{n}\frac{\psi(Y_i,X_i)}{\bar{\psi}}\text{ for } i\in[n],$$

where

$$L_D(\lambda,\lambda_c)=\sum_{i=1}^{n}q_i\lambda_i$$
$$-\frac{1}{2\mu}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_iY_j\left[\langle\phi(X_i),\phi(X_j)\rangle+c_-(X_i)'c_-(X_j)\right]+2\lambda_c\sum_{i=1}^{n}\lambda_i Y_i c_-(X_i)+\lambda_c^2\right],$$

and

$$\langle\phi(x),\phi(\tilde{x})\rangle=\sum_{j=1}^{J}\phi_j(x)\phi_j(\tilde{x}).$$

Let $(\hat{\kappa}_0,\hat{\kappa}_\phi,\hat{\kappa}_c,\hat{\xi})$ be the solution to the primal problem, and $(\hat{\lambda}_1,\ldots,\hat{\lambda}_n,\hat{\lambda}_c)$ be the solution to the dual problem. For an out-of-sample point x such that $w=(x',c_-(x))'$, we take the action according to

$$\hat{a}(x)=\text{sign}\left(\phi(x)\hat{\kappa}_\phi+c_-(x)\hat{\kappa}_c+\hat{\kappa}_0\right)$$
$$=\text{sign}\left(\frac{1}{\mu}\sum_{i:\hat{\lambda}_i>0}\hat{\lambda}_i Y_i\left[\langle\phi(x),\phi(X_i)\rangle+c_-(x)c_-(X_i)\right]+\frac{1}{\mu}\cdot\hat{\lambda}_c\cdot c_-(x)+\hat{\kappa}_0\right).$$

Since a series expansion is used to approximate $P(x)$, we refer to the above rule as the series support vector decision rule. In particular, if we use a polynomial approximation such that $\phi(x)=\left(x,x^2,x^3,\ldots,x^J\right)'$, we refer to the method as "Poly-MU-SVM".

4.2 Kernel Support Vector Decision Making

In series support vector decision making, the dual objective function depends on the features only via the inner product $\langle\phi(x),\ \phi(\tilde{x})\rangle$ in $\mathbb{R}^{d_J}\subseteq\ell^2$, where ℓ^2 is the standard sequence space consisting of square-summable sequences. The action rule is also solely determined by this inner product. To capture the inner product more compactly, we define a kernel function:

$$\mathcal{K}^J(x,\tilde{x})=\langle\phi(x),\ \phi(\tilde{x})\rangle:=\sum_{j=1}^{J}\phi_j(x)\phi_j(\tilde{x})\tag{18}$$

for any $x \in \mathcal{X}$ and $\tilde{x} \in \mathcal{X}$. If $\phi(\cdot)$ is continuous, then $\mathcal{K}^J(\cdot,\cdot)$ is continuous, symmetric, and positive-definite. Importantly, our series support vector decision rule can be expressed entirely in terms of the kernel function $\mathcal{K}^J(\cdot,\cdot)$ defined above.

Conversely, for any general kernel function $\mathcal{K}(\cdot,\cdot)$ that is continuous, symmetric, and strictly positive-definite, Mercer's theorem allows us to represent it as:

$$\mathcal{K}(x,\tilde{x}) = \sum_{j=1}^{\infty} \alpha_j^* \phi_j^*(x) \phi_j^*(\tilde{x}), \tag{19}$$

where $\{\alpha_j^*\}_{j=1}^{\infty}$ are a sequence of non-increasing eigenvalues of $\mathcal{K}(\cdot,\cdot)$, and $\{\phi_j^*(\cdot)\}$ are the corresponding eigenfunctions:

$$\int_{\mathcal{X}} \mathcal{K}(x,\tilde{x}) \phi_j^*(\tilde{x})\, d\tilde{x} = \alpha_j^* \phi_j^*(x) \text{ with } \alpha_j^* > 0, \text{ for each } x \in \mathcal{X},$$

and the right-hand side of (19) converges uniformly on compact subsets of $\mathcal{X}$. See, for example, Steinwart and Christmann (2008) (Section 4.5) for further discussion on Mercer's theorem.[11]

To develop a support vector decision rule using the general kernel function $\mathcal{K}(\cdot,\cdot)$, we choose the feature mapping to be

$$\phi(x) = (\sqrt{\alpha_1^*}\phi_1^*(x), \ldots, \sqrt{\alpha_j^*}\phi_j^*(x), \ldots)' \in \ell_2, \tag{20}$$

for each $x \in \mathcal{X}$. Then the primal problem becomes

$$\min_{\kappa_0,\kappa_\phi,\kappa_c,\xi} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\psi(Y_i,X_i)}{\bar{\psi}} \xi_i + \frac{\mu}{2}\left(\|\kappa_\phi\|^2 + \kappa_c^2 \right) \right\} \text{ subject to}$$

$$Y_i \left(\kappa_0 + \sum_{j=1}^{\infty} \left[\sqrt{\alpha_j^*}\phi_j^*(X_i) \right] \kappa_{\phi,j} + c_-(X_i)\kappa_c \right) + \xi_i \geq q_i,\ \xi_i \geq 0, \text{ for } i \in [n], \text{ and } \kappa_c \geq 0,$$

where $\|\kappa_\phi\|^2 = \sum_{j=1}^{\infty} \kappa_{\phi,j}^2$. We can regard $\sum_{j=1}^{\infty} \sqrt{\alpha_j^*}\phi_j^*(\cdot)\kappa_{\phi,j}$ as an element of the RKHS generated by the kernel $\mathcal{K}(\cdot,\cdot)$. By definition, its squared RKHS norm is

$$\sum_{j=1}^{\infty} \frac{(\sqrt{\alpha_j^*}\kappa_{\phi,j})^2}{\alpha_j^*} = \sum_{j=1}^{\infty} \kappa_{\phi,j}^2,$$

which is exactly the same as $\|\kappa_\phi\|^2$. Thus, the margin regularization becomes a control of the RKHS norm.

Note that, for the feature mapping given in (20), we have

$$\langle \phi(x), \phi(\tilde{x}) \rangle_{\ell^2} = \sum_{j=1}^{\infty} \alpha_j^* \phi_j^*(x) \phi_j^*(x) = \mathcal{K}(x,\tilde{x}).$$

[11] If $\mathcal{K}(\cdot,\cdot)$ is positive-definite but not strictly positive-definite, then $\alpha_j^* \geq 0$. In this case, if $\mathcal{K}(\cdot,\cdot)$ has an infinite number of positive eigenvalues, then the representation in (19) still holds by dropping the summands associated with zero eigenvalues, if any. If $\mathcal{K}(\cdot,\cdot)$ has a finite number of positive eigenvalues such that for some integer $J > 1$, $\alpha_j = 0$ for all $j > J$, then Mercer's expansion can be written in the form given in (18), and this case has been covered in the previous subsection.

The dual problem then becomes:

$$\begin{aligned}
&\max_{\lambda_1,\ldots,\lambda_n,\lambda_c} L_D(\lambda,\lambda_c) \text{ subject to} \\
&\sum_{i=1}^{n}\lambda_i Y_i = 0,\ \lambda_c \geq 0, \text{ and} \\
&0 \leq \lambda_i \leq \frac{1}{n}\frac{\psi(Y_i,X_i)}{\bar{\psi}}, \text{ for } i \in [n], \qquad (21)
\end{aligned}$$

where

$$L_D(\lambda,\lambda_c) = \sum_{i=1}^{n} q_i\lambda_i - \frac{1}{2\mu}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \mathcal{K}_c(X_i,X_j) + 2\lambda_c\sum_{i=1}^{n}\lambda_i Y_i c_-(X_i) + \lambda_c^2\right],$$

and

$$\mathcal{K}_c(x,\tilde{x}) = \mathcal{K}(x,\tilde{x}) + c_-(x)'c_-(\tilde{x})$$

is an augmented kernel function. The dual problem is as easy to solve as the problem with a linear decision boundary.

With the solution $(\hat{\lambda}_1,\ldots,\hat{\lambda}_n,\hat{\lambda}_c)$ to the dual problem, we can recover $\hat{\kappa}_\phi$, $\hat{\kappa}_c$, and $\hat{\kappa}_0$ as follows:

$$\begin{aligned}
\hat{\kappa}_\phi &= \frac{1}{\mu}\sum_{i\in S_+}\hat{\lambda}_i Y_i \phi(X_i), \\
\hat{\kappa}_c &= \frac{1}{\mu}\left(\sum_{i\in S_+}\hat{\lambda}_i Y_i c_-(X_i) + \hat{\lambda}_c\right), \\
\hat{\kappa}_0 &= \frac{1}{\left|S_+^\psi\right|}\sum_{i\in S_+^\psi}\left(q_i Y_i - \frac{1}{\mu}\left[\sum_{j\in S_+}\hat{\lambda}_j Y_j \mathcal{K}_c(X_i,X_j) + \hat{\lambda}_c c_-(X_i)\right]\right),
\end{aligned}$$

where $S_+ = \{i : \hat{\lambda}_i > 0\}$ and $S_+^\psi = \{i : 0 < \hat{\lambda}_i < \psi(Y_i,X_i)/(n\bar{\psi})\}$. For an out-of-sample point $w = (x', c_-(x))'$, we take the action according to

$$\begin{aligned}
\hat{a}(x) &= \text{sign}\left(\phi(x)'\hat{\kappa}_\phi + c_-(x)\hat{\kappa}_c + \hat{\kappa}_0\right) \\
&= \text{sign}\left(\frac{1}{\mu}\left\{\sum_{i\in S_+}\hat{\lambda}_i Y_i\left[\phi(x)'\phi(X_i) + c_-(x)c_-(X_i)\right] + \hat{\lambda}_c c_-(x)\right\} + \hat{\kappa}_0\right) \\
&= \text{sign}\left(\frac{1}{\mu}\left[\sum_{i\in S_+}\hat{\lambda}_i Y_i \mathcal{K}_c(x,X_i) + \hat{\lambda}_c c_-(x)\right] + \hat{\kappa}_0\right).
\end{aligned}$$

The last expression depends only on the augmented kernel function $\mathcal{K}_c(\cdot,\cdot)$. There is no need to know $\phi(\cdot)$ to compute the action rule. The main differences between our procedure and the standard kernel SVM are the sign constraint, the class-dependent margin requirement, the case-specific loss, and the augmented kernel function.

For ease of reference, we call the support vector decision rule built on a kernel function with an infinite number of positive eigenvalues as the kernel support vector decision rule. A

widely used example of such kernels in the SVM literature is the Gaussian Radial Basis Function (RBF) kernel, defined as $\mathcal{K}(x,\tilde{x}) = \exp(-\tau\|x-\tilde{x}\|^2)$, which is continuous, symmetric, and strictly positive-definite. Using this kernel function is equivalent to choosing $m(x,\theta) = \sum_{i=1}^{n}\theta_i \exp(-\tau\|x-X_i\|^2)$ to approximate $P(x)$. Buhmann (2003) provides a comprehensive introduction to the theory and applications of radial basis functions.

5 Theoretical results

5.1 Bayes Consistency

In this subsection, we consider the population decision problem where the true data distribution is known. We show that the excess risk under the MU approach is bounded above by the excess risk under MU-SVM approach. Hence, a decision rule that is Bayes-consistent under the MU-SVM is also Bayes-consistent under the MU.[12]

Ignoring the regularization term in (12), which is relevant only for finite samples, we can see that the population support vector decision solves

$$f^*_{\text{MU-SVM}}(\cdot) = \arg\min_{f\in\mathcal{M}} Q_{\text{MU-SVM}}(f), \tag{22}$$

where $Q_{\text{MU-SVM}}(f) = E\ell_{\text{MU-SVM}}(Y, f(X))$, with

$$\ell_{\text{MU-SVM}}(y, f(x)) = \left\{1\{y=+1\}\left[q^+ - f(x)\right]_+ + 1\{y=-1\}\left[q^- + f(x)\right]_+\right\}\psi(y,x),$$

and $\mathcal{M}$: $\mathcal{X}\to\mathbb{R}$ is the class of all measurable functions defined on $\mathcal{X}$.[13] Since the MU-SVM loss is nonnegative, we can solve for the optimal function value $f^*_{\text{MU-SVM}}(x)$ pointwise for each point $x\in\mathcal{X}$.

Now, for a given point $x\in\mathcal{X}$, we define $\alpha := f(x)\in\mathbb{R}$. Then, the pointwise minimization problem is

$$\begin{aligned}\alpha^*_{\text{MU-SVM}} &= \arg\min_{\alpha\in\mathbb{R}} E\left[\ell_{\text{MU-SVM}}(Y,\alpha)\,|X=x\right] \\ &= \arg\min_{\alpha\in\mathbb{R}}\left\{P(x)\left[q^+-\alpha\right]_+\psi(1,x) + [1-\alpha]\left[q^-+\alpha\right]_+\psi(-1,x)\right\}. \end{aligned}\tag{23}$$

The above minimization problem is equivalent to the one in (22), but with a different point of view. In (22), we adopt a function view and search for the entire function $f(\cdot)$ in a function space, while in (23), we adopt a scalar view and search for a scalar value in the real line for a given value x of X.

In the rest of this subsection, our analysis will be conditional on $X=x$, unless stated otherwise. For notional simplicity, we suppress the conditioning and write $P_1 = P = P(x)$, $P_{-1} = 1-P(x)$, $\psi_1 = \psi_1(x) = \psi(1,x)$, $\psi_{-1} = \psi_{-1}(x) = \psi(-1,x)$. We then have

$$\alpha^*_{\text{MU-SVM}} = \arg\min_{\alpha\in\mathbb{R}} Q_{\text{MU-SVM}}(\alpha),$$

where, with some abuse of notation, we define

$$Q_{\text{MU-SVM}}(\alpha) = P_1\psi_1\left[q^+-\alpha\right]_+ + P_{-1}\psi_{-1}\left[q^-+\alpha\right]_+ \tag{24}$$

[12] In the machine learning literature, Bayes consistency, as defined here, is often referred to as Fisher consistency.

[13] Sufficient conditions for the measurability of $f^*_{\text{MU-SVM}}(\cdot)$ can be found in Lemma A.3.18 of Steinwart and Christmann (2008).

as the conditional risk (at $X = x$).

Based on $Q_{\text{MU-SVM}}(\cdot)$, we define the (conditional) excess risk under the MU-SVM:

$$R_{\text{MU-SVM}}(\alpha) = Q_{\text{MU-SVM}}(\alpha) - \inf_{\alpha \in \mathbb{R}} Q_{\text{MU-SVM}}(\alpha).$$

The excess risk $R_{\text{MU-SVM}}(\alpha)$ measures how far $Q_{\text{MU-SVM}}(\alpha)$ is from its infimum over $\alpha \in \mathbb{R}$. If the infimum is achieved at $\alpha^*_{\text{MU-SVM}}$, then for a given α, $R_{\text{MU-SVM}}(\alpha)$ measures how far α is from $\alpha^*_{\text{MU-SVM}}$ in terms of their $Q_{\text{MU-SVM}}$-risks. Intuitively, we can regard the gap $Q_{\text{MU-SVM}}(\alpha) - Q_{\text{MU-SVM}}(\alpha^*_{\text{MU-SVM}})$ as a measure of the distance between α and $\alpha^*_{\text{MU-SVM}}$.

Next, we define the corresponding quantities under the MU approach. Let

$$\ell_{\text{MU}}(y, f(x)) = 1\{\text{sign}(yf(x)) < 0\}\,\psi(y, x).$$

The conditional risk function under the MU approach, conditional on $X = x$, is

$$Q_{\text{MU}}(\alpha) = E[\ell_{\text{MU}}(Y, \alpha) \,|X = x] = P_1\psi_1 \cdot 1\{\text{sign}(\alpha) = -1\} + P_{-1}\psi_{-1} \cdot 1\{\text{sign}(\alpha) = 1\}. \quad (25)$$

This is well-defined unless $\alpha = 0$, in which case, $\text{sign}(0)$ is not well-defined and we will treat this shortly.

To define the excess risk, we first find the value of α that minimizes $Q_{\text{MU}}(\alpha)$. Let $c = \psi_{-1}/(\psi_1 + \psi_{-1})$. Clearly, when $P \neq c$,

$$\alpha^*_{\text{MU}} = P - c \in \arg\inf_{\alpha \in \mathbb{R}} Q_{\text{MU}}(\alpha).$$

Hence, when $\alpha \neq 0$ and $P \neq c$, the conditional excess risk

$$R_{\text{MU}}(\alpha) = Q_{\text{MU}}(\alpha) - \inf_{\alpha \in \mathbb{R}} Q_{\text{MU}}(\alpha),$$

is well-defined.

The case with $P = c$ or $\alpha = 0$ requires special treatment, as $\text{sign}(0)$ is not well-defined. Regardless of how $\text{sign}(0)$ may be defined, we assume that the decision $\text{sign}(0)$ incurs the maximum risk when $P \neq c$. That is,

$$Q_{\text{MU}}(0) = \max\{P_1\psi_1, P_{-1}\psi_{-1}\} \text{ if } P \neq c,$$

and so

$$R_{\text{MU}}(0) = \max\{P_1\psi_1, P_{-1}\psi_{-1}\} - Q_{\text{MU}}(P - c) \text{ if } P \neq c.$$

On the other hand, we treat any decision as a correct decision when $P = c$. Hence, $Q_{\text{MU}}(\alpha) = 0$ and $R_{\text{MU}}(\alpha) = 0$ for any $\alpha \in \mathbb{R}$ if $P = c$.

Combining all cases, we define $R_{\text{MU}}(\cdot)$ as follows:

$$R_{\text{MU}}(\alpha) = \begin{cases} Q_{\text{MU}}(\alpha) - Q_{\text{MU}}(P - c), & \text{if } P \neq c, \alpha \neq 0; \\ \max\{P_1\psi_1, P_{-1}\psi_{-1}\} - Q_{\text{MU}}(P - c), & \text{if } P \neq c, \alpha = 0; \\ 0, & \text{if } P = c. \end{cases}$$

In the statistical literature, excess risks are also referred to as regrets. The definition of excess risks, or regrets, depends on the risk measures being used. Here, we have two risk measures: Q_{MU} and $Q_{\text{MU-SVM}}$. The first measure is directly tied to what we care about, while the second measure is motivated by margin considerations. From a computational perspective, it is much easier to

optimize the sample analogue of the second measure; however, this measure is not directly linked to the expected utility we aim to maximize. Our goal is to show that a minimizer of the surrogate excess risk function also minimizes the excess risk function induced by the expected utility, thereby maximizing the expected utility. More precisely, we aim to show that $R_{\text{MU}}(\alpha) \leq R_{\text{MU-SVM}}(\alpha)$ for any $\alpha \in \mathbb{R}$, such that whenever $R_{\text{MU-SVM}}(\alpha)$ is close to zero, $R_{\text{MU}}(\alpha)$ is also close to zero.

For any $\epsilon > 0$, define the "ϵ-worse" set:

$$\mathcal{S}_{\text{MU}}(\epsilon) = \{\alpha \in \mathbb{R} : R_{\text{MU}}(\alpha) > \epsilon\}.$$

For any $\alpha \in \mathcal{S}_{\text{MU}}(\epsilon)$, its excess Q_{MU}-risk is greater than ϵ. Intuitively, $\mathcal{S}_{\text{MU}}(\epsilon)$ consists of α values that perform worse than the best choice α^*_{MU} by at least ϵ. On this set, we show in the next proposition that the excess $Q_{\text{MU-SVM}}$-risk is also greater than ϵ, which then implies that $R_{\text{MU}}(\alpha) \leq R_{\text{MU-SVM}}(\alpha)$ for any $\alpha \in \mathbb{R}$.

Proposition 2 *For any finite $\epsilon > 0$, we have*

(i) $\inf_{P\in[0,1]} \inf_{\alpha\in\mathcal{S}_{\text{MU}}(\epsilon)} R_{\text{MU-SVM}}(\alpha) \geq \epsilon$ *where, for a null set $\varnothing$,* $\inf_{\alpha\in\varnothing} R_{\text{MU-SVM}}(\alpha)$ *is defined to be $+\infty$.*

(ii) $R_{\text{MU}}(\alpha) \leq R_{\text{MU-SVM}}(\alpha)$ *for any $\alpha \in \mathbb{R}$.*

(iii) the unconditional excess risks satisfy

$$E[R_{\text{MU}}(f(X))] \leq E[R_{\text{MU-SVM}}(f(X))],$$

for any $f \in \mathcal{M}$ and any distribution of X such that $E[R_{\text{MU-SVM}}(f(X))]$ is well-defined.

Proposition 2 extends a well-known result in the SVM literature (e.g., Chapter 3 of Steinwart and Christmann (2008)) by allowing for case-dependent losses and class-specific margins. Propositions 2(i) and 2(ii) are pointwise results, each of which holds for every $x \in \mathcal{X}$, while Proposition 2(iii) is an integrated version of Proposition 2(ii). The proposition holds for any positive values of q^+ and q^-. This gives us the flexibility to use different margins for different classes of observations.

Proposition 2(ii) shows that the minimizer of $R_{\text{MU-SVM}}(\alpha)$ is necessarily also a minimizer of $R_{\text{MU}}(\alpha)$. According to the proof of Proposition 2, we have: if $P_1\psi_1 < P_{-1}\psi_{-1}$ (equivalently, $P < c$),

$$R_{\text{MU}}(\alpha) = \begin{cases} 0, & \text{if } \alpha < 0 \\ P_{-1}\psi_{-1} - P_1\psi_1, & \text{if } \alpha \geq 0 \end{cases}$$

$$R_{\text{MU-SVM}}(\alpha) = \begin{cases} -(q^- + \alpha) P_1\psi_1, & \text{if } \alpha \leq -q^- \\ (q^- + \alpha)\left(P_{-1}\psi_{-1} - P_1\psi_1\right), & \text{if } -q^- < \alpha < q^+ \\ (q^- + \alpha) P_{-1}\psi_{-1} - (q^+ + q^-) P_1\psi_1, & \text{if } \alpha \geq q^+ \end{cases}$$

and if $P_1\psi_1 \geq P_{-1}\psi_{-1}$ (equivalently, $P \geq c$),

$$R_{\text{MU}}(\alpha) = \begin{cases} P_1\psi_1 - P_{-1}\psi_{-1}, & \text{if } \alpha \leq 0 \\ 0, & \text{if } \alpha > 0 \end{cases}$$

$$R_{\text{MU-SVM}}(\alpha) = \begin{cases} (q^+ - \alpha) P_1\psi_1 - (q^+ + q^-) P_{-1}\psi_{-1}, & \text{if } \alpha \leq -q^- \\ (q^+ - \alpha)\left(P_1\psi_1 - P_{-1}\psi_{-1}\right), & \text{if } -q^- < \alpha < q^+ \\ (\alpha - q^+) P_{-1}\psi_{-1}, & \text{if } \alpha \geq q^+ \end{cases}$$

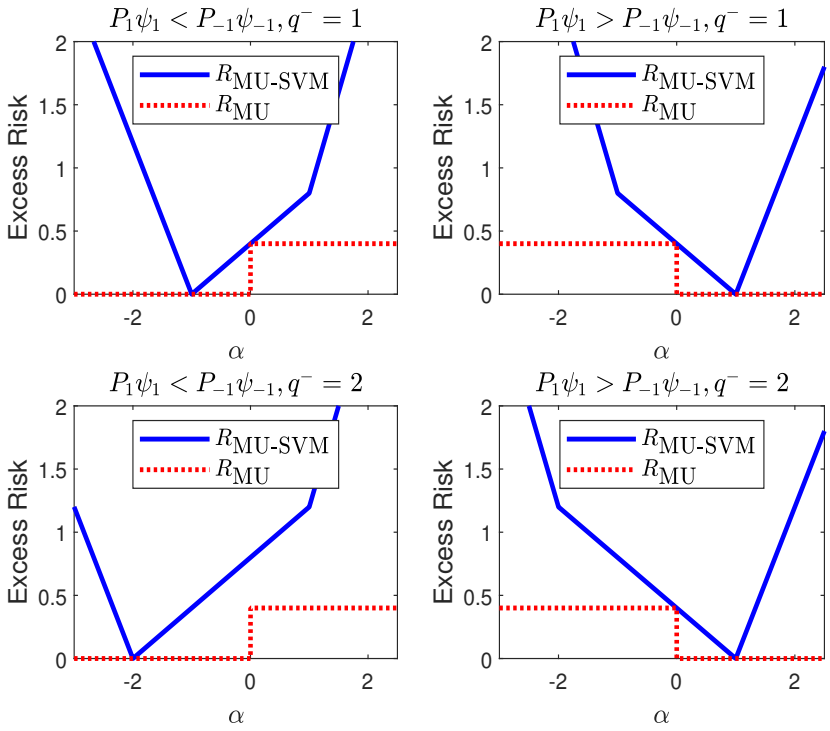


Figure 7: The excess risk as a function of α for the cases $(q^+, q^-) = (1, 1)$ and $(1, 2)$.

Figure 7 plots each of $R_{\text{MU-SVM}}(\alpha)$ and $R_{\text{MU}}(\alpha)$ as a function of α for some fixed $P_1\psi_1$ and $P_{-1}\psi_{-1}$ when $q^+ = q^- = 1$ and when $q^+ = 1, q^- = 2$. Given P and ψ, the figure clearly shows that the minimizer of $R_{\text{MU-SVM}}(\alpha)$ also minimizes $R_{\text{MU}}(\alpha)$. The figure also demonstrates the convexity of $R_{\text{MU-SVM}}(\alpha)$ as a function of α.

Proposition 2(iii) compares the *unconditional* excess risks, as the conditional covariate has been integrated out. It shows that for any $f \in \mathcal{M}$, the unconditional excess Q_{MU}-risk of f is bounded above by its unconditional excess $Q_{\text{MU-SVM}}$-risk. Thus, we can use the $Q_{\text{MU-SVM}}$-risk as a surrogate for the Q_{MU}-risk.

Proposition 2 allows us to study the transferability of Bayes consistency, a notion we now define. To this end, consider a general loss function $\ell_{\text{G}}(\cdot,\cdot) \geq 0$, with the subscript "G" indicating it is a general loss. Define the corresponding *conditional* Q_{G}-risk, conditional on $X = x$, by $Q_{\text{G}}(\alpha) = E[\ell_{\text{G}}(Y, \alpha)|X = x]$, and the *conditional* excess risk by $R_{\text{G}}(\alpha) = Q_{\text{G}}(\alpha) - \inf_{\alpha\in\mathbb{R}} Q_{\text{G}}(\alpha)$. Here, $Q_{\text{G}}(\alpha)$ and $R_{\text{G}}(\alpha)$ are defined in the same way as $Q_{\text{MU-SVM}}(\alpha)$ and $R_{\text{MU-SVM}}(\alpha)$ (or $Q_{\text{MU}}(\alpha)$ and $R_{\text{MU}}(\alpha)$) are defined. The unconditional risk and excess risk of f under the loss $\ell_{\text{G}}(\cdot,\cdot)$ are then given by $E[Q_{\text{G}}(f(X))]$ and $E[R_{\text{G}}(f(X))]$, respectively. Let $\hat{f}_n$ be a sequence of estimators of the target function that minimizes $E[Q_{\text{G}}(f(X))]$ over $f \in \mathcal{M}$. For example, we can take $\hat{f}_n$ as an approximate minimizer of $n^{-1}\sum_{i=1}^n \ell_{\text{G}}(Y_i, f(X_i))$, the empirical version of $E[Q_{\text{G}}(f(X))]$, over a certain function space.

Definition 3 *Let $F_X(\cdot)$ be the CDF of X. If $E\left[R_{\text{G}}(\hat{f}_n(X))\right] := \int_{\mathcal{X}} R_{\text{G}}(\hat{f}_n(x))dF_X(x) \to 0$ in probability as $n \to \infty$, then we say that $\hat{f}_n$ is Bayes-consistent with respect to the risk measure Q_{G}.*

Clearly, Bayes consistency is a concept inherently linked to a specific risk measure of interest. By definition, if $\hat{f}_n$ is Bayes-consistent with respect to a particular risk measure Q_{G}, then the Q_{G}-risk of $\hat{f}_n$ will approach the smallest possible Q_{G}-risk as n increases. In other words, the performance of $\hat{f}_n$ in terms of the Q_{G}-risk improves, eventually converging to the smallest possible Q_{G}-risk. Mathematically, $Q_{\mathrm{G}}(\hat{f}_n) \to \inf_{f \in \mathcal{M}} Q_{\mathrm{G}}(f)$ in probability as $n \to \infty$.

Now, suppose that $\hat{f}_{\text{MU-SVM},n}$ is Bayes-consistent under $Q_{\text{MU-SVM}}$ so that

$$E[R_{\text{MU-SVM}}(\hat{f}_{\text{MU-SVM},n}(X))] = o_p(1).$$

By Proposition 2(iii), which is an algebraic result that holds for any $f \in \mathcal{M}$, including a sample-dependent function $\hat{f}_{\text{MU-SVM},n}$, we have

$$E[R_{\text{MU}}(\hat{f}_{\text{MU-SVM},n}(X))] \leq E[R_{\text{MU-SVM}}(\hat{f}_{\text{MU-SVM},n}(X))].$$

Hence, $E[R_{\text{MU}}(\hat{f}_{\text{MU-SVM},n}(X))] = o_p(1)$. In fact, if

$$E[R_{\text{MU-SVM}}(\hat{f}_{\text{MU-SVM},n}(X))] = O_p(\delta_n)$$

for a sequence $\delta_n \to 0$, then, by the same argument, we have: $E[R_{\text{MU}}(\hat{f}_{\text{MU-SVM},n}(X))] = O_p(\delta_n)$. Therefore, any decision rule that is Bayes-consistent under $Q_{\text{MU-SVM}}$ is also Bayes-consistent under Q_{MU} with the same rate of convergence. To a great extent, using $Q_{\text{MU-SVM}}$ as the risk measure combines the best of both worlds: the computational feasibility and the transferability of Bayes consistency.

5.2 Finite Sample Generalization Bound

In this subsection, we establish a generalization bound for the support vector decision rule. We consider the general case where the decision boundary lies in an RKHS, which can be either finite or infinite dimensional, covering all cases considered in the previous sections.

Let $\mathcal{F}_0$ be a set of constant functions, $\mathcal{F}_{\mathcal{K}}$ be the RKHS corresponding to a kernel $\mathcal{K}(\cdot,\cdot)$, which can be $\mathcal{K}^J(\cdot,\cdot)$ as defined in (18) or the kernel considered in (19), and $\mathcal{F}_c$ be the linear subspace spanned by $c_-(\cdot)$. Recall that the MU-SVM solves

$$\min_{\kappa_\phi, \kappa_c, \kappa_0, \xi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu_n}{2} \left(\|\kappa_\phi\|^2 + \kappa_c^2 \right) \right\},$$

subject to the constraints:

$$Y_i f(X_i) + \xi_i \geq q_i, \ \xi_i \geq 0, \ \text{for } i \in [n], \ \text{and } \kappa_c \geq 0.$$

In the above, μ_n is a positive constant that may depend on the sample size, and $f(x) = f_0(x) + f_{\mathcal{K}}(x) + f_c(x)$ for

$$f_0(x) \in \mathcal{F}_0, \ f_{\mathcal{K}}(x) = \sum_{j=1}^{J} \sqrt{\alpha_j^*} \phi_j^*(x) \kappa_{\phi,j} \in \mathcal{F}_{\mathcal{K}}, \ \text{and } f_c(x) = c_-(x) \kappa_c \in \mathcal{F}_c,$$

where J can be finite or infinite in this subsection, depending on whether $\mathcal{K}(\cdot,\cdot)$ has a finite number of positive eigenvalues or not.

We assume that $\mathcal{F}_0 = \{f : f(x) \equiv \kappa_0 \text{ for all } x \in \mathcal{X} \text{ and some } \kappa_0 \in \mathbb{K}_0\}$, where $\mathbb{K}_0 \subset \mathbb{R}$ is a compact set, and that any pairwise intersection of $\mathcal{F}_0, \mathcal{F}_\mathcal{K}$, and $\mathcal{F}_c$ contains only the zero function.[14] Let $\mathcal{F} = \mathcal{F}_0 \oplus \mathcal{F}_\mathcal{K} \oplus \mathcal{F}_c$ be the direct sum of $\mathcal{F}_0, \mathcal{F}_\mathcal{K}$, and $\mathcal{F}_c$. For any $f(x) = f_0(x) + f_\mathcal{K}(x) + c_-(x)\kappa_c \in \mathcal{F}$, we define the "$\mathcal{K}c$" norm $\|\cdot\|_{\mathcal{K}c}$ as

$$\|f\|_{\mathcal{K}c} = \left(\|f_\mathcal{K}\|_\mathcal{K}^2 + \kappa_c^2\right)^{1/2},$$

where $\|\cdot\|_\mathcal{K}$ stands for the RKHS norm. Equipped with the above norm $\|\cdot\|_{\mathcal{K}c}$, $\mathcal{F}$ is a Banach space.

For a solution $\hat{f} \in \mathcal{F}$ of the MU-SVM problem, we are interested in obtaining a high-probability upper bound for $Q_{\text{MU}}(\hat{f})$ where

$$Q_{\text{MU}}(f) = E\left[\psi(Y,X)1\left\{Y \neq \text{sign}\left(f(X)\right)\right\}\right] = E\left[\psi(Y,X) \cdot 1\left\{Yf(X) \leq 0\right\}\right],$$

assuming X is a continuous random variable. The upper bound for $Q_{\text{MU}}(\hat{f})$ informs us about the out-of-sample performance of the decision rule $a(x) = \text{sign}(\hat{f}(x))$. To this end, for some $s > 0$, we define the ramp loss:

$$\tilde{h}_s(r) = \begin{cases} 1, & \text{if } r \leq 0, \\ 1 - r/s, & \text{if } 0 < r < s, \\ 0, & \text{if } r \geq s, \end{cases}$$

which is a truncated and rescaled version of the hinge loss $[1-r]_+$. The ramp loss is not convex, but it is Lipschitz continuous. Note that $1\{r \leq 0\} \leq \tilde{h}_s(r) \leq 1\{r \leq s\}$. So $\tilde{h}_s(r)$ can be regarded as a smooth interpolation between the usual 0-1 loss $1\{r < 0\}$ and the margin-sensitive 0-1 loss $1\{r < s\}$.

For some $\gamma^+ \in (0, q^+]$ and $\gamma^- \in (0, q^-]$, denote $\gamma = (\gamma^+, \gamma^-)$ and define

$$h_\gamma(r,y) = \tilde{h}_{\gamma^+}(r)\, 1\{y = +1\} + \tilde{h}_{\gamma^-}(r)\, 1\{y = -1\}.$$

In view of the fact that $1\{r \leq 0\} \leq \tilde{h}_{\min(\gamma^+,\gamma^-)}(r) \leq h_\gamma(r,y)$ for both $y = +1$ and $y = -1$, we have

$$Q_{\text{MU}}(f) \leq E\left[h_\gamma(Yf(X), Y)\psi(Y,X)\right] := Q_{h_\gamma}(f)$$

for any $f \in \mathcal{F}$. In particular, $Q_{\text{MU}}(\hat{f}) \leq Q_{h_\gamma}(\hat{f})$. Instead of establishing an upper bound for $Q_{\text{MU}}(\hat{f})$, we will focus on deriving an upper bound for $Q_{h_\gamma}(\hat{f})$. This approach proves to be more manageable since, for each fixed y, $h_\gamma(\cdot, y)$ is both bounded and Lipschitz continuous.

Our upper bound for $Q_{h_\gamma}(\hat{f})$ is based on its empirical version $Q_{n,h_\gamma}(\hat{f})$, where

$$Q_{n,h_\gamma}(f) = \frac{1}{n}\sum_{i=1}^n \left[h_\gamma\left(Y_i f(X_i), Y_i\right)\psi(Y_i, X_i)\right].$$

Let

$$\psi_{\max}^+ = \sup_{x \in \mathcal{X}_+} \psi(1,x), \psi_{\max}^- = \sup_{x \in \mathcal{X}_-} \psi(-1,x),$$

where $\mathcal{X}_+$ and $\mathcal{X}_-$ are the supports of X conditional on $Y = +1$ and -1, respectively.

[14] For example, if we take $\mathcal{F}_\mathcal{K}$ to be the popular RKHS generated by a Gaussian RBF, then $\mathcal{F}_0 \cap \mathcal{F}_\mathcal{K}$ contains only the zero function, as the Gaussian RKHS does not include any non-zero constant functions; see Corollary 4.44 of Steinwart and Christmann (2008) (p. 141).

Theorem 4 *Let Assumptions 1 and 2 hold. Assume that* $\psi_{\max} = \max\left(\psi_{\max}^{+}, \psi_{\max}^{-}\right) < \infty$ *and* $\sup_{\kappa_0 \in \mathbb{K}_0} |\kappa_0| < \infty$.

(i) With probability one, $\hat{f} \in \mathcal{F}_n$ *for*

$$\mathcal{F}_n := \mathcal{F}^{\mu_n} = \left\{ f \in \mathcal{F} : \|f\|_{\mathcal{K}c}^2 \leq \frac{q^+ + q^-}{\mu_n} \right\}.$$

(ii) With probability at least $1-\delta$, *for any* $\gamma = (\gamma^+, \gamma^-) \in (0, q^+] \otimes (0, q^-]$, *we have*

$$Q_{\mathrm{MU}}(\hat{f}) \leq Q_{n,h_\gamma}(\hat{f}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^{+}}{\gamma^+}, \frac{\psi_{\max}^{-}}{\gamma^-} \right) \mathcal{V}\left(\mathcal{F}_n, n, \delta\right) + \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2} \log \frac{2}{\delta}},$$

where

$$\mathcal{V}\left(\mathcal{F}_n, n, \delta\right) = \sqrt{\frac{q^+ + q^-}{\mu_n}} \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_c\left(X_i, X_i\right)} + \sqrt{\max_{x \in \mathcal{X}} \mathcal{K}_c\left(x, x\right)} \sqrt{\frac{1}{2} \log \frac{2}{\delta}} \right) + \sup_{\kappa_0 \in \mathbb{K}_0} |\kappa_0|.$$

Theorem 4(ii) remains valid if $Q_{n,h_\gamma}(\hat{f})$ is replaced by the following empirical measure:

$$\frac{1}{n} \sum_{i=1}^{n} \left[1\{Y_i \hat{f}\left(X_i\right) < \gamma^+\} 1\left\{Y_i = +1\right\} + 1\{Y_i \hat{f}\left(X_i\right) < \gamma^-\} 1\left\{Y_i = -1\right\} \right] \psi(Y_i, X_i), \tag{26}$$

which is larger than $Q_{n,h_\gamma}(\hat{f})$. The above measure is constructed based on the margin-sensitive 0-1 losses with γ^+ and γ^- as the margin parameters. The upper bound in Theorem 4(ii) is tighter and is therefore preferred from a theoretical perspective. However, one might find the measure in (26) more intuitively appealing.

The generalization bound in Theorem 4(ii) depends on μ_n, which controls the margin of the action rule. On the one hand, as μ_n increases, the first term (i.e., $Q_{n,h_\gamma}(\hat{f})$) in the generalization bound becomes larger because there is a stronger restriction on the RKHS norm and a weaker restriction on the margin, leading to a larger margin and hence more misclassified points. This causes the first term in the generalization bound to increase.

For intuition, we can refer to Figure 6 and consider the problem from the perspectives of the extensive and intensive margins. From the perspective of the extensive margin, a larger μ_n implies that the MU-SVM solution has a smaller norm. As a result, the "positive" and "negative" hyperplanes in the figure are further apart, and more points will fall on the wrong side of their respective hyperplanes. This leads to more points having a positive value of $h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)$, which increases $Q_{n,h_\gamma}(\hat{f})$. From the perspective of the intensive margin, a larger μ_n implies a smaller value of $Y_i f\left(X_i\right)$, a weakly larger value of $h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)$, and thus a larger value of $Q_{n,h_\gamma}(\hat{f})$.

On the other hand, as μ_n increases, the second term, which reflects the size of the function space $\mathcal{F}_n$, becomes smaller. Therefore, there is an opportunity to choose μ_n to trade off these two terms, and it can be selected via cross-validation.

Suppose we use the MU approach to obtain

$$\hat{f}_{\mathrm{MU}} \in \arg\min_{f \in \mathcal{F}_n} Q_{n,\mathrm{MU}}(f), \text{ for } Q_{n,\mathrm{MU}}(f) = \frac{1}{n} \sum_{i=1}^{n} 1\left\{Y_i f(X_i) \leq 0\right\} \psi(Y_i, X_i).$$

Then a generalization bound based on VC theory typically takes the following form: with probability at least $1-\delta$,

$$Q_{\text{MU}}(\hat{f}_{\text{MU}}) \leq Q_{n,\text{MU}}(\hat{f}_{\text{MU}}) + \frac{\psi_{\max}}{\sqrt{n}}\mathcal{C}\left(\sqrt{\text{VC}[\text{sign}(\mathcal{F}_n)]\log(n)} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right),$$

where $\mathcal{C}>0$ is a constant and $\text{VC}[\text{sign}(\mathcal{F}_n)]$ is the VC dimension of the class $\text{sign}(\mathcal{F}_n) := \{\text{sign}(f): f \in \mathcal{F}_n\}$. See, for example, Proposition 1 in Su (2021). However, the above bound is not useful if $\text{VC}[\text{sign}(\mathcal{F}_n)]$ is infinite. Indeed, $\text{VC}[\text{sign}(\mathcal{F}_n)] = \infty$ when $\mathcal{F}_{\mathcal{K}}$ is an infinite-dimensional RKHS generated by a universal kernel.[15] In such a case, we have to replace $\text{VC}[\text{sign}(\mathcal{F}_n)]\log(n)$ by the logarithm of 2^n, the growth function of $\text{sign}(\mathcal{F}_n)$, and we obtain: with probability at least $1-\delta$,

$$Q_{\text{MU}}(\hat{f}_{\text{MU}}) \leq Q_{n,\text{MU}}(\hat{f}_{\text{MU}}) + \psi_{\max}\mathcal{C}\sqrt{\log 2} + \frac{\psi_{\max}}{\sqrt{n}}\mathcal{C}\sqrt{\log\left(\frac{1}{\delta}\right)}.$$

Note that $\psi_{\max}\mathcal{C}\sqrt{\log 2}$ is a constant that does not decay to zero. The high-probability bound for the generalization error $Q_{\text{MU}}(\hat{f}_{\text{MU}}) - Q_{n,\text{MU}}(\hat{f}_{\text{MU}})$ does not decay to zero as n increases. In contrast, the generalization error bound in Theorem 4 goes to zero as long as $\mu_n \to 0$ such that $n\mu_n \to \infty$. In particular, if μ_n goes to zero arbitrarily slowly, then $Q_{\text{MU}}(\hat{f}) - Q_{n,h_\gamma}(\hat{f})$ goes to zero at a rate arbitrarily close to $1/\sqrt{n}$.

The generalization bound in Theorem 4 also depends on γ. The first term in the generalization bound increases with γ^+ and γ^-, while the second term decreases with γ^+ and γ^-. In principle, we can choose γ to optimize these two terms. However, this would lead to a random $\hat{\gamma}$, but the above bound can only accommodate a fixed γ. To address this, we employ an idea from Bartlett (1998) to establish an upper bound that holds for a random $\hat{\gamma}$. The main departure is that we have a two-dimensional parameter $\gamma \in (0, q^+] \otimes (0, q^-]$, while the parameter in Bartlett (1998) is one-dimensional (i.e., a scalar).

Proposition 5 *Let the assumptions in Theorem 4 hold. For any $\hat{\gamma} = \left(\hat{\gamma}^+, \hat{\gamma}^-\right) \in (0, q^+] \otimes (0, q^-]$ that may be random and data-dependent, we have: with probability at least $1-\delta$,*

$$Q_{\text{MU}}(\hat{f}) \leq Q_{n,h_{\hat{\gamma}}}(\hat{f}) + \frac{4}{\sqrt{n}}\max\left(\frac{\psi^+_{\max}}{\hat{\gamma}^+}, \frac{\psi^-_{\max}}{\hat{\gamma}^-}\right) \cdot \mathcal{V}\left(\mathcal{F}_n, n, \frac{\delta\hat{\gamma}^+\hat{\gamma}^-}{4q^+q^-}\right) + \frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{8q^+q^-}{\delta\hat{\gamma}^+\hat{\gamma}^-}}.$$

The proposition allows us to search for a $\hat{\gamma}$ to minimize the generalization bound. We note that there is a search cost: compared to the generalization bound in Theorem 4(ii), both the second and third terms in the above bound have extra inflation factors.

To accommodate a data-driven choice of μ_n, we need to establish an upper bound that is uniform over all choices of μ_n. Let $\{\mu_{n,\ell},\ \ell = 1, \ldots, L_n\}$ be the set of the candidate values of μ_n. Both $\mu_{n,\ell}$ and L_n can depend on and grow with n, but they are not allowed to depend on the sample. Let $\{p_\ell \geq 0, \ell = 1, \ldots, L_n\}$ be a nonnegative sequence with $\sum_{\ell=1}^{L_n} p_\ell \leq 1$. Let $\hat{f}_{\mu_n,\ell}$ be the estimator of $f \in \mathcal{F}^{\mu_{n,\ell}}$, and let $\hat{\ell}$ represent the data-driven cross-validated choice over $\ell = 1, \ldots, L_n$.

[15] In the context of machine learning and kernel methods, a universal kernel refers to a type of kernel function whose RKHS is dense in the space of continuous functions under the sup norm.

Corollary 6 *Let the assumptions in Theorem 4 hold.*

(i) With probability at least $1-\delta$*, we have*

$$Q_{\mathrm{MU}}(\hat{f}_{\mu_{n,\hat{\ell}}}) < Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\hat{\ell}}}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\gamma^+}, \frac{\psi_{\max}^-}{\gamma^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, p_{\hat{\ell}}\delta\right) + \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2}\log\frac{2}{p_{\hat{\ell}}\delta}}$$

for any fixed $\gamma = (\gamma^+, \gamma^-) \in (0, q^+] \otimes (0, q^-]$.

(ii) For any $\hat{\gamma} = \left(\hat{\gamma}^+, \hat{\gamma}^-\right) \in (0, q^+] \otimes (0, q^-]$ *that may be random and data-dependent, such as*

$$\left(\hat{\gamma}^+, \hat{\gamma}^-\right) \in \arg\min_{\gamma \in (0,q^+]\otimes(0,q^-]} \left[Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\hat{\ell}}}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\gamma^+}, \frac{\psi_{\max}^-}{\gamma^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, p_{\hat{\ell}}\delta\right)\right],$$

we have, with probability at least $1-\delta$*,*

$$Q_{\mathrm{MU}}(\hat{f}_{\mu_{n,\hat{\ell}}}) \le Q_{n,h_{\hat{\gamma}}}(\hat{f}_{\mu_{n,\hat{\ell}}}) + \frac{4}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\hat{\gamma}^+}, \frac{\psi_{\max}^-}{\hat{\gamma}^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, \frac{\delta p_{\hat{\ell}} \hat{\gamma}^+ \hat{\gamma}^-}{4q^+q^-}\right) + \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2}\log\frac{8q^+q^-}{\delta p_{\hat{\ell}} \hat{\gamma}^+ \hat{\gamma}^-}}.$$

The above upper bound can be translated directly into a lower bound on the out-of-sample mean utility. Consider the case with a fixed γ as an example. Denote the out-of-sample mean utility as

$$\begin{aligned} \mathbb{U}(f) &= EU(\mathrm{sign}\,(f), Y, X) = EU(Y, Y, X) - E\psi(Y, X) \cdot 1\{Y \neq \mathrm{sign}\,(f)\} \\ &= EU(Y, Y, X) - Q_{\mathrm{MU}}(f). \end{aligned} \tag{27}$$

Then, by Theorem 4(ii), we have, with probability at least $1-\delta/2$,

$$\mathbb{U}(\hat{f}_{\mu_{n,\hat{\ell}}}) > EU(Y,Y,X) - Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\hat{\ell}}}) - \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\gamma^+}, \frac{\psi_{\max}^-}{\gamma^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, \frac{p_{\hat{\ell}}\delta}{2}\right) - \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2}\log\frac{4}{p_{\hat{\ell}}\delta}}.$$

The unknown $EU(Y,Y,X)$ in the above can be estimated by $\frac{1}{n}\sum_{i=1}^n U(Y_i, Y_i, X_i)$. Using Hoeffding's lemma, we have

$$\begin{aligned} &\Pr\left(\frac{1}{n}\sum_{i=1}^n U(Y_i, Y_i, X_i) - EU(Y,Y,X) > U_{\max}\sqrt{\frac{2}{n}\log\frac{2}{\delta}}\right) \\ &\le \exp\left\{-\frac{4n^2U_{\max}^2}{4nU_{\max}^2}\frac{1}{n}\log\frac{2}{\delta}\right\} = \frac{\delta}{2}. \end{aligned}$$

Hence, with probability at least $1-\delta$,

$$\begin{aligned} \mathbb{U}(\hat{f}_{\mu_{n,\hat{\ell}}}) > \frac{1}{n}\sum_{i=1}^n U(Y_i, Y_i, X_i) - Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\hat{\ell}}}) - \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\gamma^+}, \frac{\psi_{\max}^-}{\gamma^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, \frac{1}{2}p_{\hat{\ell}}\delta\right) \\ - \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2}\log\frac{4}{p_{\hat{\ell}}\delta}} - U_{\max}\sqrt{\frac{2}{n}\log\frac{2}{\delta}}. \end{aligned}$$

The case with a data-driven γ can be handled similarly.

6 Simulation Study

6.1 Utility Normalization

To make our simulation results comparable to those in Su (2021), we normalize the utility function in the same way. By direct calculations, we have

$$U(a,y,x) = b(x)\left[\frac{y+1}{2} - c(x)\right]\left(\frac{a+1}{2}\right) + U(-1,y,x).$$

This shows that the binary decision problem does not depend on $U(-1,y,x)$. Hence, we can normalize $U(-1,y,x)$ to be any value. If we normalize it according to

$$U(-1,y,x) = -\frac{1}{2}b(x)\left(\frac{y+1}{2} - c(x)\right),$$

then

$$U(a,y,x) = \frac{1}{2}b(x)\left[\frac{y+1}{2} - c(x)\right]a.$$

This normalization amounts to using the payoff functions in the table below:

	$y=1$	$y=-1$
$a=1$	$+\frac{1}{2}b(x)[1-c(x)]$	$-\frac{1}{2}b(x)c(x)$
$a=-1$	$-\frac{1}{2}b(x)[1-c(x)]$	$+\frac{1}{2}b(x)c(x)$
loss from "incorrect actions"	$b(x)[1-c(x)]$	$b(x)c(x)$

Under the above normalization, the loss function becomes

$$\begin{aligned}\ell(a,y,x) &= U(y,y,x) - U(a,y,x) = \frac{1}{2}b(x)\left[\frac{y+1}{2} - c(x)\right](y-a)\\ &= \psi(y,x)\cdot 1\{y \neq a\},\end{aligned}$$

where

$$\psi(y,x) = b(x)\left[\frac{y+1}{2} - yc(x)\right] = \frac{1}{2}b(x)\{y[1-2c(x)]+1\} > 0. \tag{28}$$

We will use the above normalization in our simulation study; see (29) and (30). The normalization is the same as that in Su (2021), so our simulation results are directly comparable to those in Su (2021).

6.2 Simulation Design

Following Elliott and Lieli (2013), we consider two data generating processes (DGPs), each of which is combined with two preferences. These simulation designs are also considered by Su (2021). In the first data generating process, X is a scalar random variable following the location-shifted and rescaled beta distribution:

$$X \sim 5B(1,1.3) - 2.5,$$

where $B(1,1.3)$ is the standard beta distribution with parameter $(1,1.3)$. Since $B(1,1.3)$ is supported on $[0,1]$, X is supported on $[-2.5,2.5]$. Conditional on $X=x$, Y is generated according to

$$P(Y=1|X=x) = P(x) \text{ and } P(Y=-1|X=x) = 1-P(x),$$

where

$$P(x) = \Lambda\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right) := \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)\right]}$$

is the true conditional probability and $\beta =: (\beta_0, \beta_1, \beta_2, \beta_3)' = (0, -0.5, 0, 0.2)'$. The choice of β, which is the same as that in Elliott and Lieli (2013) and Su (2021), makes $P(x)$ sufficiently close to the linear logit model $\Lambda(\beta_0 + \beta_1 x)$ in the following sense: the 5% Lagrangian multiplier test of the null of a linear logit model against the alternative of the cubic logit model $\Lambda\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)$ has power of about 35% only.

For this DGP, we consider two preferences, leading to two different configurations of $b(\cdot)$ and $c(\cdot)$:

Preference 1: $b(x) = 20$ and $c(x) = 0.5$;

Preference 2: $b(x) = 20$ and $c(x) = 0.5 + 0.025x$.

Given this DGP, the cubic ML under the specification $m_{\mathrm{ML}}(x, \theta) = \Lambda\left(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3\right)$ is clearly correctly specified. On the other hand, the cubic MU with the specification that $m_{\mathrm{MU}}(x, \theta)$ is a cubic polynomial is correctly (sign) specified. To see this, note that $P(x) - c(x) = 0$ has three real solutions under both Preferences 1 and 2. See Figure 8 below for an illustration. For each preference, we can find a cubic polynomial $\mathcal{P}_3(x, \theta)$ that has the same three solutions as $P(x) - c(x)$ and satisfies $\operatorname{sign}(\mathcal{P}_3(x, \theta)) = \operatorname{sign}(P(x) - c(x))$.[16] Note that $\operatorname{sign}(\mathcal{P}_3(x; \theta)) = \operatorname{sign}([\mathcal{P}_3(x, \theta) + c(x)] - c(x))$ and $\mathcal{P}_3^c(x, \theta) := \mathcal{P}_3(x, \theta) + c(x)$ is also a cubic polynomial. From the perspective of the action rule, the specification of $m_{\mathrm{MU}}(x, \theta)$ as a cubic polynomial (i.e., $m_{\mathrm{MU}}(x, \theta) = \mathcal{P}_3^c(x; \theta) = \mathcal{P}_3(x, \theta) + c(x)$) is correct, as only the sign of $m_{\mathrm{MU}}(x, \theta) - c(x)$ matters. Unlike $m_{\mathrm{ML}}(x, \theta)$, which is constrained to be in the unit interval $[0, 1]$, $m_{\mathrm{MU}}(x, \theta)$ does not have such a restriction and can take values outside this interval.

In the second data generating process, X consists of two variables: $X = (X_1, X_2)'$, where X_1 and X_2 are independent and uniformly distributed on $[-3.5, 3.5]$. The true conditional distribution of Y given $X = x$ for $x = (x_1, x_2)'$ is

$$P(x) = \Lambda(Q(v)) \text{ for } v = 1.5x_1 + 1.5x_2,$$

where $Q(v) = (1.5 - 0.1v)\exp\left[-\left(0.25v + 0.1v^2 - 0.04v^3\right)\right]$ is not a polynomial. We also consider two preferences under this DGP:

Preference 3: $b(x) = 20$ and $c(x) = 0.75$;

Preference 4: $b(x) = 20 + 40 \cdot 1\{|x_1 + x_2| < 1.5\}$ and $c(x) = 0.75$.

Relative to Preference 3, Preference 4 makes observations closer to the center of the distribution more important. This will have an effect on the decision rule, as the choice of $b(x)$ affects the loss $\psi(y, x)$; see equation (28).

For this DGP, the cubic ML under the specification that $m_{\mathrm{ML}}(x, \theta) = \Lambda(\mathcal{P}_3(x; \theta))$ is not correctly specified. In fact, $\Lambda(\mathcal{P}_j(x; \theta))$ is not correctly specified for any finite order polynomial $\mathcal{P}_j(x; \theta)$, because $Q(v)$ is not a polynomial. However, the cubic MU, which specifies $m_{\mathrm{MU}}(x, \theta)$ as a cubic polynomial, is correctly (sign) specified. To see this, we note that $\Lambda(Q(v)) - 0.75 = 0$ has three solutions, say v_1, v_2, v_3; see Figure 9. For each solution v_j, the pair $(x_1, x_2)'$ satisfying $1.5x_1 + 1.5x_2 = v_j$ is a solution to $P(x) - c(x) = 0$. But there is a cubic

[16] We use $\mathcal{P}_j(x, \theta)$ to represent a polynomial in variable x with degree j and coefficient θ. The polynomial can be different for different occurrences.

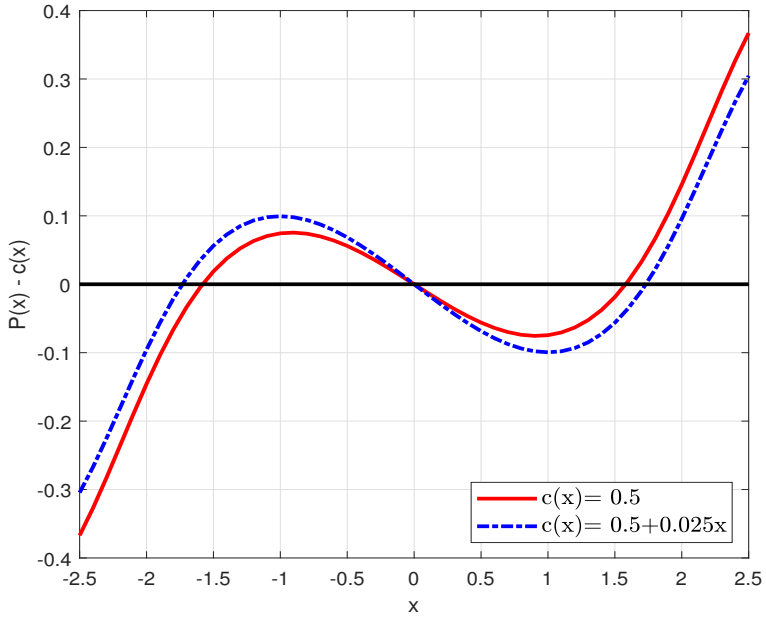


Figure 8: Graph of $P(x) - c(x)$ under DGP 1 and Preferences 1 and 2.

polynomial $\mathcal{P}_3(v;\theta_v)$ with coefficients θ_v that has the same three roots as $\Lambda(Q(v)) - 0.75$, and $\text{sign}(\mathcal{P}_3(v,\theta_v)) = \text{sign}(\Lambda(Q(v)) - 0.75)$. Hence $\text{sign}(\mathcal{P}_3(1.5x_1 + 1.5x_2;\theta_v)) = sign(P(x) - c(x))$. Given that $\mathcal{P}_3(1.5x_1 + 1.5x_2;\theta_v) := \mathcal{P}_3^x(x,\theta)$ is a cubic polynomial in x, and $c(x)$ is constant, $\mathcal{P}_3^x(x,\theta) + c(x) = \mathcal{P}_3^{x,c}(x;\theta)$ is also a cubic polynomial in x. Therefore, if we specify $m_{\text{MU}}(x,\theta)$ as a cubic polynomial (i.e., $m_{\text{MU}}(x,\theta) = \mathcal{P}_3^{x,c}(x,\theta)$), then for some θ, the sign of $m_{\text{MU}}(x,\theta) - c(x)$ matches exactly with the sign of $P(x) - c(x)$.

For each data generating process, we consider five groups of methods (see Section S.4 in the supplementary appendix for a summary of these methods). The first group consists of the ML method with the logit polynomial specification $m_{\text{ML}}(x,\theta) = \Lambda(\mathcal{P}_j(x,\theta))$ of the conditional probability $P(x)$ for $j = 1, 2, 3$. The second group consists of the standard MU method with the polynomial specification $m_{\text{MU}}(x,\theta) - c(x) = \mathcal{P}_j(x,\theta)$. The third group consists of the penalized MU method that uses the simulated maximal discrepancy (SMD) as the data-dependent complexity penalty.[17] We denote this as MU-SMD. The SMD is a data-driven measure of model complexity. We let the maximum polynomial order J be 3, 4, 5. For each maximum polynomial order J, the MU-SMD selects the best polynomial order over $j = 1, 2, \ldots, J$. In terms of achieving higher utilities, Su (2021) shows that the MU-SMD method dominates the MU method of (Elliott and Lieli (2013)) that uses pretesting to select the polynomial order. It also dominates the MU method that uses cross-validation and the MU method that uses AIC and BIC types of penalty to select the polynomial order. Hence, the third group consists of the most competitive procedure in the literature.[18]

The fourth group consists of "Lp-SVMs", which ignore the loss heterogeneity, and the cost-

[17]Su (2021) also considers other data-dependent complexity penalties, but alternative penalties do not deliver better performances than SMD. Here, we have also ignored a technical term that is detrimental to the performance of the penalized MU with data-dependent complexity penalties. See Su (2021) for more details.

[18]In estimating the SMD, we follow Su (2021) to set the parameter m (the number of simulations defined in that paper) to be 10. Setting $m = 100$ produces similar simulation results. We thank Dr. Su for sharing the programs.

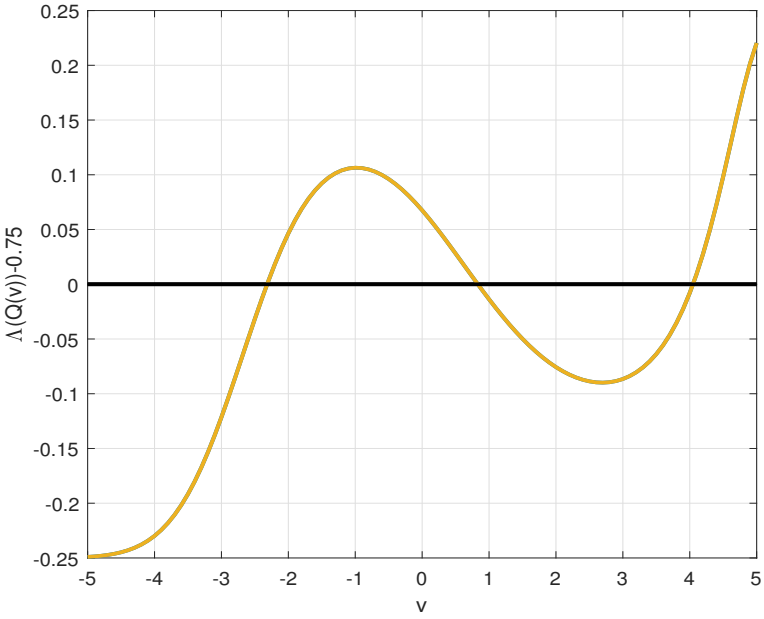


Figure 9: Graph of $\Lambda(Q(v)) - 0.75$ (cf. DGP 2 and Preferences 3 and 4).

sensitive SVM (CS-SVM), which partially accounts for the loss heterogeneity. As a prominent example of an "Lp-SVM", the standard "L2-SVM" employs the squared L_2-norm $\|\kappa_w\|_2^2$ as the regularizer. We also consider using the L_1-norm $\|\kappa_w\|_1$ as the regularizer, leading to the "L1-SVM". The CS-SVM we consider here is a modified version of the L2-SVM (see, for example, Lin et al. (2002), Bach et al. (2006), and Fernández et al. (2018)). It involves using different costs for false positives and false negatives. However, the costs do not depend on any covariate. In the simulation, we use the averages of $\psi(-1, X_j)$ and $\psi(1, X_j)$ over the simulated covariate values as the costs for false positives and false negatives, respectively. The CS-SVM can be regarded as a misspecified version of the proposed polynomial MU-SVM: instead of using correctly specified covariate-specific costs, the CS-SVM employs the averaged version, thereby only partially accounting for the loss heterogeneity.

For all SVM-based methods in the fourth group, we use $\kappa_0 + \phi(x)' \kappa_\phi = 0$ as the decision boundary and specify $\kappa_0 + \phi(x)' \kappa_\phi$ as a polynomial of order $j = 3, 4$, or 5. We consider polynomials of higher order than those for the ML and standard MU methods because the SVM is a regularized method, while the ML method and the standard MU method are not. We implement the L2-SVM and CS-SVM using quadratic programming based on the primal problem, and we implement the L1-SVM using the `lpsvm` algorithm provided by Fung and Mangasarian (2004). For the hyperparameter μ in the L1-SVM and L2-SVM, we employ the rate of correct actions as the criterion and use ten-fold cross validation to choose it from $\left(2^{-12}, 2^{-10}, \ldots, 2^{12}\right)/n$. This method of selecting μ is compatible with the recommendation of Hsu et al. (2003). For the hyperparameter μ in the CS-SVM, we use the same selection method as that used for the Poly-MU-SVM, described next.

The last group consists of the methods proposed in this paper with three different values of $\rho = 1/4, 1/2, 3/4$. The first procedure in this group is the polynomial series MU-SVM method (Poly-MU-SVM) in Section 4.1 with $\kappa_0 + \phi(x)' \kappa_\phi = 0$ as the decision boundary. As in the case with L1-SVM and L2-SVM, we specify $\kappa_0 + \phi(x)' \kappa_\phi$ as polynomials of order $j = 3, 4, 5$.

For example, when x is a scalar, we have $\phi(x) = \left(x, x^2, \ldots, x^j\right)$ so that all monomials are treated equally. We also use ten-fold cross validation to choose the hyperparameter μ, but the cross-validation criterion is now the average utility achieved for each μ. We solve the polynomial MU-SVM using quadratic programming based on the primal problem. See Section S.2 for details.

The second procedure in the last group is the kernel-based MU-SVM method in Section 4.2. We use the radial basis function (RBF) kernel $\mathcal{K}(x, \tilde{x}) = \exp(-\tau \|x - \tilde{x}\|^2)$. We note that for the MU or Poly-MU-SVM methods, $c(x)$ can be absorbed into the polynomial specification, and there is no need to include the extra covariate $c_-(x)$ in formulating the decision boundary. In contrast, for the RBF-based MU-SVM, $c(x)$ is included as a separate variable. We implement the RBF-based kernel MU-SVM (denoted by RBF-MU-SVM) via quadratic programming based on the dual problem (cf. Section S.2). The hyperparameters (μ, τ) are chosen via ten-fold cross validation over the grid $\left(2^{-12}, 2^{-10}, \ldots, 2^{12}\right)/n \otimes \left(2^{-12}, 2^{-10}, \ldots, 2^{12}\right)$.

We use the average relative out-of-sample utility as the performance criterion to compare different methods. For each method, let $\hat{a}(\cdot)$ be the action rule constructed using the sample $\{(X_i, Y_i)\}_{i=1}^n$ where n is the size of the estimation sample (i.e., the training sample). The average out-of-sample utility is computed as

$$\bar{U}_{out,\hat{a}} = \frac{1}{n_{out}} \sum_{j=n+1}^{n+n_{out}} \left\{ \frac{1}{2} b(X_j) \left[\frac{Y_j + 1}{2} - c(X_j) \right] \hat{a}(X_j) \right\}, \tag{29}$$

where $\{(X_j, Y_j) : j = n+1, \ldots, n+n_{out}\}$ is the set of out-of-sample observations (i.e., the testing sample). We normalize $\bar{U}_{out,\hat{a}}$ by the oracle out-of-sample utility assuming that $P(x)$ is known:

$$U_{oracle} = \frac{1}{n_{out}} \sum_{j=n+1}^{n+n_{out}} \left\{ \frac{1}{2} b(X_j) \left[\frac{Y_j + 1}{2} - c(X_j) \right] \operatorname{sign}(P(X_j) - c(X_j)) \right\}. \tag{30}$$

For each simulation replication, we compute the utility ratio $\bar{U}_{out,\hat{a}}/U_{oracle}$ and report its average over 500 simulation replications. In the experiments, we set $n = 500$, 1000 and $n_{out} = 5000$.

6.3 Simulation Results

Table 1 reports the utility ratio for DGP 1 when $n = 500$. For this DGP, the cubic polynomial specification (i.e., $j = 3$ in the table) under the ML method is correct, and it is not surprising that it outperforms all other methods. However, the ML method with misspecified lower-order polynomials (i.e., $j = 1$ and 2) is dominated by all other methods. This is especially true under Preference 2, where $c(\cdot)$ is not a constant function. The penalized MU-SMD method outperforms the standard MU method. The L1-SVM and L2-SVM methods perform well with L1-SVM dominating L2-SVM, which is the standard SVM. The performance of the CS-SVM is close to that of the L2-SVM, which is expected, as $c(\cdot)$ is not significantly different from a constant function.

Our proposed Poly-MU-SVM with $\rho = 1/2$ performs as well as the standard L2-SVM under Preference 1, where both $b(\cdot)$ and $c(\cdot)$ are constant functions and $c(x) = 0.5$. It clearly outperforms the L2-SVM under Preference 2, where $c(x)$ depends on x. This provides strong evidence that the Poly-MU-SVM dominates the standard SVM when the utility-induced loss depends on covariates. On the other hand, the Poly-MU-SVM with $\rho = 1/2$ is numerically indistinguishable from the CS-SVM under Preference 1 where both $b(\cdot)$ and $c(\cdot)$ are constant functions. In this case, these two methods are theoretically identical. However, the Poly-MU-SVM with $\rho = 1/2$

clearly dominates the CS-SVM under Preference 2, where $c(\cdot)$ is not a constant function. This demonstrates the advantage of using the Poly-MU-SVM over the CS-SVM when the cost of a false decision is covariate-dependent.

Among the MU-SVM methods, the RBF kernel version and the polynomial series version have comparable performances. In an overall sense, each of our proposed MU-SVM methods outperforms the MU-SMD, often by a large margin.

For the proposed MU-SVM methods, Table 1 shows that the choice of ρ matters. In particular, $\rho = 1/2$ yields better results than the other two alternatives.

Table 1: Average out-of-sample utility ratios (in percentage) under DGP 1 and Preferences 1 and 2 for $n = 500$.

	$P(x) = \Lambda\left(-0.5x + 0.2x^3\right)$					
	$b(x) = 20, c(x) = .5$			$b(x) = 20, c(x) = .5 + .025x$		
Polynomial order	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
ML	34.21	31.47	93.10	8.66	11.67	94.33
MU	54.87	44.68	64.47	33.36	45.06	53.81
Max Poly order	$J = 3$	$J = 4$	$J = 5$	$J = 3$	$J = 4$	$J = 5$
MU-SMD	59.25	62.07	69.40	56.62	59.07	67.19
Polynomial order	$j = 3$	$j = 4$	$j = 5$	$j = 3$	$j = 4$	$j = 5$
L1-SVM	85.85	84.20	87.52	82.10	79.50	81.72
L2-SVM	69.18	70.47	81.32	62.00	63.53	75.24
CS-SVM	69.21	70.48	81.19	62.48	64.94	75.47
Poly-MU-SVM ($\rho = 1/4$)	59.81	63.23	75.33	65.72	66.92	75.45
Poly-MU-SVM ($\rho = 1/2$)	69.21	70.48	81.19	76.42	76.50	83.17
Poly-MU-SVM ($\rho = 3/4$)	66.96	67.09	76.09	74.36	72.00	76.97
	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
RBF-MU-SVM	74.98	77.14	74.30	74.90	77.83	72.76

Table 2 reports the utility ratio for DGP 2 when $n = 500$. For the ML, all polynomial specifications are incorrect. As a result, the ML method does not perform well. The L1-SVM and L2-SVM perform even worse. The reason is that these two SVM methods do not account for the dependence of the utility on covariates. Even though neither $b(x)$ nor $c(x)$ depends on the covariate x under Preference 3, the fact that $c(x) = 0.75$ rather than 0.5 captures the dependence of the utility on the action and the outcome. The L1-SVM and L2-SVM effectively use 0.5 as the decision threshold for the conditional probability. Such a threshold does not capture the loss asymmetry. Theoretically, when $b(x) = 20$ and $c(x) = 0.75$, we have $\psi(y, x) = b(x)\left[(y+1)/2 - y \times c(x)\right] = 10 - 5y$ and $U(a, y, x) = 0.5b(x)\left[(y+1)/2 - c(x)\right] a = (5y - 2.5)\, a$. Hence, the loss under $y = -1$ (i.e., the loss from a false positive action) is three times as large as the loss under $y = 1$ (i.e., the loss from a false negative action). That is, a false-positive decision costs three times as large as a false-negative decision. Ignoring the loss asymmetry, both L1-SVM and L2-SVM result in very bad performances. In addition, $\bar{U}_{out,\hat{a}}$ could be negative if too many false decisions are made. In the extreme case when $a = -y$, we have $U(a, y, x) = -5 + 2.5y \in$

$\{-7.5, -2.5\}$.

The Poly-MU-SVM with $\rho = 1/2$ performs the same as the CS-SVM under Preference 3, where both $b(\cdot)$ and $c(\cdot)$ are constant functions. However, the former outperforms the latter under Preference 4, where $b(\cdot)$ is not a constant function even though $c(\cdot)$ is. Note that when the covariate is equal to x, the costs of false positives and false negatives are $b(x)c(x)$ and $b(x)[1-c(x)]$, respectively. Therefore, the cost of a false decision depends on x when $b(x)$ depends on x, while $c(x)$ does not. In such cases, the CS-SVM is inferior to the Poly-MU-SVM because it fails to capture the cost heterogeneity across different covariate values.

The performance of the Poly-MU-SVM is comparable to or better than that of the MU-SMD when the maximum polynomial order is 4 or 5. When the maximum polynomial order is 3, the MU-SMD does better than the Poly-MU-SVM, but not by a large margin. The performance of the RBF-MU-SVM is much better than that of the MU-SMD for all polynomial specifications under consideration.

As in Table 1, we observe that the choice of ρ affects the performance of the proposed MU-SVM methods, with $\rho = 1/2$ and $\rho = 3/4$ yielding better results than $\rho = 1/4$.

Table 2: Average out-of-sample utility ratios (in percentage) under DGP 2 and Preferences 3 and 4 for $n = 500$.

	$P(x) = \Lambda(Q(1.5(x_1+x_2)))$, $Q(v) = \frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
	$b(x) = 20$ $c(x) = 0.75$			$b(x) = 20 + 40\{x_1 + x_2 < 1.5\}$ $c(x) = 0.75$		
Polynomial order	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
MLE	59.16	58.53	59.48	29.04	27.71	33.43
MU	69.45	50.11	68.60	54.63	32.91	50.66
Max Poly order	$J=3$	$J=4$	$J=5$	$J=3$	$J=4$	$J=5$
MU-SMD	68.16	67.84	68.95	51.10	51.06	52.90
Polynomial order	$j=3$	$j=4$	$j=5$	$j=3$	$j=4$	$j=5$
L1-SVM	0.37	7.42	8.32	14.11	19.90	19.72
L2-SVM	-6.51	0.71	2.90	9.40	14.72	16.35
CS-SVM	64.94	69.54	78.44	38.48	49.46	64.85
Poly-MU-SVM ($\rho = 1/4$)	61.70	64.79	69.84	38.79	52.94	62.59
Poly-MU-SVM ($\rho = 1/2$)	64.94	69.54	78.44	43.57	59.63	69.32
Poly-MU-SVM ($\rho = 3/4$)	65.60	70.40	79.56	43.51	59.23	67.68
	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
RBF-MU-SVM	74.90	76.24	76.58	63.76	64.75	64.09

Tables 3 and 4 report the results when $n = 1000$. Relative to the sample size $n = 500$, the performance of the ML method improves under the correct specification, but this is not the case under misspecifications. This is expected. On the one hand, under correct specifications, the MLE converges more quickly to the true parameter value, as it becomes more efficient. On the other hand, under misspecifications, the MLE converges more quickly to a value that is different

from the true parameter value. In contrast, when the sample size increases from 500 to 1000, the performance of each MU-based method improves. The L1-SVM and L2-SVM do not necessarily have better performances for a larger sample size. This is because these SVM methods are not tailored to the specific problem at hand.

The relative performances of all methods in Table 3 are comparable to those in Table 1. The qualitative observations made for Table 1 are applicable to Table 3. Similarly, the qualitative observations made for Table 2 are applicable to Table 4.

To sum up, the ML method is not suitable for utility-based decision making when there is a risk of model misspecifications. The standard L1-SVM and L2-SVM do not work well when the loss function depends on either the outcome variable or the covariates. The CS-SVM works well when the loss function depends only on the outcome variable, but its performance deteriorates when the loss also depends on the covariates. The penalized MU-SMD works reasonably well, but its performance is often dominated by the proposed support vector decision rules. In particular, the RBF-MU-SVM outperforms the penalized MU-SMD in all cases, and often by a large margin. Simulation results not reported here show that, for the MU-SVM method, it is important to include the cutoff function as a separate covariate if it is not multicollinear with other covariates or features used in formulating the decision boundary.

Table 3: Average out-of-sample utility ratios (in percentage) under DGP 1 and Preferences 1 and 2 for $n = 1000$.

	$P(x) = \Lambda\left(-0.5x + 0.2x^3\right)$					
	$b(x) = 20, c(x) = .5$			$b(x) = 20, c(x) = .5 + .025x$		
Polynomial order	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
ML	31.06	30.80	97.23	6.78	6.72	97.80
MU	58.94	48.53	68.45	34.71	48.02	60.31
Max Poly order	$J = 3$	$J = 4$	$J = 5$	$J = 3$	$J = 4$	$J = 5$
MU-SMD	63.43	67.62	77.07	64.10	67.07	77.47
Polynomial order	$j = 3$	$j = 4$	$j = 5$	$j = 3$	$j = 4$	$j = 5$
L1-SVM	94.13	92.70	94.63	91.25	88.64	91.12
L2-SVM	73.38	75.69	91.85	67.56	69.77	87.98
CS-SVM	73.51	75.70	91.79	68.81	71.32	86.77
Poly-MU-SVM, $\rho = 1/4$	65.98	69.33	83.41	76.75	75.32	85.67
Poly-MU-SVM, $\rho = 1/2$	73.51	75.70	91.79	88.05	87.02	94.69
Poly-MU-SVM, $\rho = 3/4$	71.50	72.23	85.13	83.17	81.91	85.69
	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
RBF-MU-SVM	83.66	87.11	83.55	85.31	90.06	83.54

Table 4: Average out-of-sample utility ratios (in percentage) under DGP 2 and and Preferences 3 and 4 for $n = 1000$.

	$P(x) = \Lambda\left(Q(1.5(x_1+x_2))\right)$, $Q(v) = \frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
	$b(x) = 20$ $c(x) = 0.75$			$b(x) = 20 + 40\{x_1 + x_2 < 1.5\}$ $c(x) = 0.75$		
Polynomial order	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
ML	58.22	56.99	59.70	26.81	24.07	31.83
MU	72.24	56.67	71.72	58.79	39.00	56.42
Max Poly order	$J=3$	$J=4$	$J=5$	$J=3$	$J=4$	$J=5$
MU-SMD	71.37	71.45	72.95	57.40	58.00	61.07
Polynomial order	$j=3$	$j=4$	$j=5$	$j=3$	$j=4$	$j=5$
L1-SVM	-2.16	7.12	7.92	13.46	19.48	20.04
L2-SVM	-8.94	-3.96	-1.80	7.81	11.50	13.10
CS-SVM	66.06	71.29	85.10	37.16	49.51	74.46
Poly-MU-SVM, $\rho = 1/4$	64.66	66.78	76.30	42.18	57.49	70.88
Poly-MU-SVM, $\rho = 1/2$	66.06	71.29	85.10	46.79	65.34	80.22
Poly-MU-SVM, $\rho = 3/4$	67.07	71.81	84.46	46.83	66.07	79.42
	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
RBF-MU-SVM	83.54	84.40	85.40	75.66	75.80	75.23

7 Conclusion

The paper considers a binary decision-making problem. Given the training sample $(X_i, Y_i)_{i=1}^n$ and an out-of-sample covariate $X = x$, the decision-maker takes a binary action $a(x)$ to maximize the expected utility $E[U(a, Y, x)|X = x]$ where the utility function $U(a, Y, x)$ depends on the action a taken, the covariate value x, and the outcome variable Y to be realized after the covariate is observed and the action is taken. Had the decision-maker known the conditional distribution $P(x) = \Pr(Y = 1|X = x)$ and the optimal cutoff function $c(x)$, which depends on the utility function, the decision-maker would have taken the optimal action $a(x) = \text{sign}(P(x) - c(x))$. However, $P(x)$ is not known, and the optimal action rule has to be learned from the training sample. This paper proposes a learning method that accounts for the covariate-specific cutoff function and the distance of the training points to the decision boundary. The method is motivated by the literature on support vector machines. However, the presence of a covariate-specific cutoff function calls for a conceptual change, leading to an augmented attribute space and a new learning method. A simulation study shows that the proposed method outperforms the most recent methods in the literature and the ML method when the model is misspecified.

In this paper, the margin and the regularizer take the form of a squared ℓ_2/L_2 norm or a squared RKHS norm. While these norms have delivered promising results, it may be worthwhile to investigate regularizers of different forms, such as the ℓ_1/L_1 norm, especially when the covariate space is of high dimension. By doing so, we could potentially achieve double sparsity, resulting in both a small number of support vectors and a limited number of covariates entering the

support vector. Another intriguing avenue for research is the development of a simple rule to select the parameter ρ that largely captures the optimal location of the middle hyperplane. To accomplish this, a more extensive simulation study may be necessary, incorporating new DGPs and preferences that have not been considered in the existing literature. A data-driven approach to choosing ρ, such as cross-validation, is also a possibility. The current focus of the paper is on binary decision problems, but it would be valuable to extend our method to handle multi-class decisions, where both Y and a can take more than two values. We leave these extensions for future research.

8 Appendix of Proofs

Proof of Proposition 2. Part (i). We follow a few steps to prove this part. First, we characterize the set $\mathcal{S}_{\mathrm{MU}}(\epsilon)$. Note that $\alpha^*_{\mathrm{MU}} = P - c$, so we have $\operatorname{sign}(\alpha^*_{\mathrm{MU}}) = \operatorname{sign}\left(P_1\psi_1 - P_{-1}\psi_{-1}\right)$ and

$$\inf_{\alpha\in\mathbb{R}} Q_{\mathrm{MU}}(\alpha) = Q_{\mathrm{MU}}(\alpha^*_{\mathrm{MU}}) = \min\left\{P_1\psi_1, P_{-1}\psi_{-1}\right\}.$$

When $P \neq c$, we have, for $\alpha \neq 0$,

$$\begin{aligned}
R_{\mathrm{MU}}(\alpha) &= Q_{\mathrm{MU}}(\alpha) - Q_{\mathrm{MU}}(\alpha^*_{\mathrm{MU}}) \\
&= P_1\psi_1\left\{1\left\{\operatorname{sign}(\alpha) = -1\right\} - 1\left\{\operatorname{sign}(\alpha^*_{\mathrm{MU}}) = -1\right\}\right\} \\
&+ P_{-1}\psi_{-1}\left\{1\left\{\operatorname{sign}(\alpha) = 1\right\} - 1\left\{\operatorname{sign}(\alpha^*_{\mathrm{MU}}) = 1\right\}\right\} \\
&= \left(P_1\psi_1 - P_{-1}\psi_{-1}\right)\cdot 1\left\{\alpha < 0, \alpha^*_{\mathrm{MU}} > 0\right\} \\
&+ \left(P_{-1}\psi_{-1} - P_1\psi_1\right)\cdot 1\left\{\alpha > 0, \alpha^*_{\mathrm{MU}} < 0\right\} \\
&= \left(P_1\psi_1 - P_{-1}\psi_{-1}\right)\cdot 1\left\{\alpha < 0, P > c\right\} \\
&+ \left(P_{-1}\psi_{-1} - P_1\psi_1\right)\cdot 1\left\{\alpha > 0, P < c\right\}.
\end{aligned}$$

When $P \neq c$, we have, for $\alpha = 0$,

$$\begin{aligned}
R_{\mathrm{MU}}(\alpha) &= \max\left\{P_1\psi_1, P_{-1}\psi_{-1}\right\} - P_1\psi_1\cdot 1\left\{\operatorname{sign}(\alpha^*_{\mathrm{MU}}) = -1\right\} - P_{-1}\psi_{-1}\cdot 1\left\{\operatorname{sign}(\alpha^*_{\mathrm{MU}}) = 1\right\} \\
&= P_1\psi_1\cdot 1\left\{P > c\right\} + P_{-1}\psi_{-1}\cdot 1\left\{P < c\right\} - P_1\psi_1\cdot 1\left\{P < c\right\} - P_{-1}\psi_{-1}\cdot 1\left\{P > c\right\} \\
&= \left(P_1\psi_1 - P_{-1}\psi_{-1}\right)1\left\{P > c\right\} + \left(P_{-1}\psi_{-1} - P_1\psi_1\right)1\left\{P < c\right\},
\end{aligned}$$

where the first line holds because $Q_{\mathrm{MU}}(0)$ is defined to be $\max\left\{P_1\psi_1, P_{-1}\psi_{-1}\right\}$ when $P \neq c$. Therefore, when $P \neq c$,

$$\begin{aligned}
&\mathcal{S}_{\mathrm{MU}}(\epsilon) \\
&= \left\{\alpha \leq 0 : P > c \text{ and } P_1\psi_1 - P_{-1}\psi_{-1} > \epsilon\right\} \cup \left\{\alpha \geq 0 : P < c \text{ and } P_{-1}\psi_{-1} - P_1\psi_1 > \epsilon\right\} \\
&= \left\{\alpha \in \mathbb{R} : \alpha \leq 0, P_1\psi_1 - P_{-1}\psi_{-1} > \epsilon\right\} \cup \left\{\alpha \in \mathbb{R} : \alpha \geq 0, P_1\psi_1 - P_{-1}\psi_{-1} < -\epsilon\right\},
\end{aligned}$$

where the second line follows because:

(i) $P_1\psi_1 - P_{-1}\psi_{-1} > \epsilon$ implies that $P_1\psi_1 - P_{-1}\psi_{-1} > 0$, which in turn implies that $P > c$; and

(ii) $P_1\psi_1 - P_{-1}\psi_{-1} < -\epsilon$ implies that $P_1\psi_1 - P_{-1}\psi_{-1} < 0$, which then implies that $P < c$.

When $P = c$, we have $P_1\psi_1 = P_{-1}\psi_{-1}$, and thus

$$R_{\text{MU}}(\alpha) = Q_{\text{MU}}(\alpha) - Q_{\text{MU}}(\alpha^*_{\text{MU}}) = 0 \text{ for any } \alpha \in \mathbb{R}.$$

Hence, $\mathcal{S}_{\text{MU}}(\epsilon) = \varnothing$, that is, $\mathcal{S}_{\text{MU}}(\epsilon)$ is an empty set.

Second, we compute $\inf_{\alpha\in\mathbb{R}} Q_{\text{MU-SVM}}(\alpha)$. We have

$$\inf_{\alpha\in\mathbb{R}} Q_{\text{MU-SVM}}(\alpha) = \inf_{\alpha}\left\{P_1\psi_1\left[q^+ - \alpha\right]_+ + P_{-1}\psi_{-1}\left[q^- + \alpha\right]_+\right\},$$

where

$$\begin{aligned}
&P_1\psi_1\left[q^+ - \alpha\right]_+ + P_{-1}\psi_{-1}\left[q^- + \alpha\right]_+ \\
&= \begin{cases} P_1\psi_1(q^+ - \alpha), & \text{if } \alpha \le -q^-; \\ P_1\psi_1(q^+ - \alpha) + P_{-1}\psi_{-1}(q^- + \alpha), & \text{if } -q^- < \alpha < q^+; \\ P_{-1}\psi_{-1}(q^- + \alpha), & \text{if } \alpha \ge q^+. \end{cases}
\end{aligned}$$

If $P < \psi_{-1}/\left(\psi_1 + \psi_{-1}\right) = c$ (i.e., $P_1\psi_1 < P_{-1}\psi_{-1}$), the infimums over the individual ranges are achieved at $\alpha = -q^-, -q^-$, and q^+, with the infimums given by $(q^+ + q^-)P_1\psi_1$, $(q^+ + q^-)P_1\psi_1$, and $(q^+ + q^-)P_{-1}\psi_{-1}$, respectively. The infimum over the entire range is achieved at $\alpha = -q^-$, with the infimum given by $(q^+ + q^-)P_1\psi_1$.

If $P > \psi_{-1}/\left(\psi_1 + \psi_{-1}\right) = c$ (i.e., $P_1\psi_1 > P_{-1}\psi_{-1}$), the infimums over the individual ranges are achieved $\alpha = -q^-, q^+$, and q^+, with the infimums given by $(q^+ + q^-)P_1\psi_1$, $(q^+ + q^-)P_{-1}\psi_{-1}$, and $(q^+ + q^-)P_{-1}\psi_{-1}$, respectively. The infimum over the entire range is achieved at $\alpha = q^+$, with the infimum given by $(q^+ + q^-)P_{-1}\psi_{-1}$.

If $P = \psi_{-1}/\left(\psi_1 + \psi_{-1}\right) = c$ (i.e., $P_1\psi_1 = P_{-1}\psi_{-1}$), the infimum over the entire range is achieved at any $\alpha \in [-q^-, q^+]$, with the infimum given by $(q^+ + q^-)P_1\psi_1$.

Hence,

$$\inf_{\alpha\in\mathbb{R}} Q_{\text{MU-SVM}}(\alpha) = \left(q^+ + q^-\right)\left(P_1\psi_1 1\{P \le c\} + P_{-1}\psi_{-1}1\{P > c\}\right).$$

Third, we compute $\inf_{\alpha\in\mathcal{S}(\epsilon)} Q_{\text{MU-SVM}}(\alpha)$. If $\left|P_1\psi_1 - P_{-1}\psi_{-1}\right| \le \epsilon$, we have $\mathcal{S}(\epsilon) = \varnothing$. By definition, $\inf_{\alpha\in\mathcal{S}(\epsilon)} Q_{\text{MU-SVM}}(\alpha) = \infty$, and Part (i) of the proposition clearly holds.

If $P_1\psi_1 - P_{-1}\psi_{-1} > \epsilon$, then the infimum is taken over $\alpha \le 0$, under which we have

$$\begin{aligned}
&P_1\psi_1\left[q^+ - \alpha\right]_+ + P_{-1}\psi_{-1}\left[q^- + \alpha\right]_+ \\
&= \begin{cases} P_1\psi_1(q^+ - \alpha) + P_{-1}\psi_{-1}(q^- + \alpha), & \text{if } -q^- < \alpha \le 0; \\ P_1\psi_1(q^+ - \alpha), & \text{if } \alpha \le -q^-. \end{cases}
\end{aligned}$$

The infimums over the individual ranges are achieved at $\alpha = 0$ and $-q^-$, respectively. The infimum over the entire range $\alpha \le 0$ is

$$\min\left(q^+P_1\psi_1 + q^-P_{-1}\psi_{-1}, \left(q^+ + q^-\right)P_1\psi_1,\right) = q^+P_1\psi_1 + q^-P_{-1}\psi_{-1}.$$

If $P_1\psi_1 - P_{-1}\psi_{-1} < -\epsilon$, then the infimum is taken over $\alpha \ge 0$, under which we have

$$\begin{aligned}
&P_1\psi_1\left[q^+ - \alpha\right]_+ + P_{-1}\psi_{-1}\left[q^- + \alpha\right]_+ \\
&= \begin{cases} P_1\psi_1(q^+ - \alpha) + P_{-1}\psi_{-1}(q^- + \alpha), & \text{if } 0 \le \alpha < q^+; \\ P_{-1}\psi_{-1}(q^- + \alpha), & \text{if } \alpha \ge q^+. \end{cases}
\end{aligned}$$

The infimums over the individual ranges are achieved at $\alpha = 0$ and q^+, respectively. The infimum over the entire range $\alpha \geq 0$ is

$$\min\left(q^+ P_1\psi_1 + q^- P_{-1}\psi_{-1}, \left(q^+ + q^-\right) P_{-1}\psi_{-1}\right) = q^+ P_1\psi_1 + q^- P_{-1}\psi_{-1}.$$

Therefore, when $\left|P_1\psi_1 - P_{-1}\psi_{-1}\right| > \epsilon$, we have

$$\inf_{\alpha\in\mathcal{S}(\epsilon)} Q_{\text{MU-SVM}}(\alpha) = q^+ P_1\psi_1 + q^- P_{-1}\psi_{-1}.$$

Fourth, using the results from the previous two steps, we have, when $\left|P_1\psi_1 - P_{-1}\psi_{-1}\right| > \epsilon$:

$$\begin{aligned}
&\inf_{\alpha\in\mathcal{S}(\epsilon)} R_{\text{MU-SVM}}(\alpha) \\
&= q^+ P_1\psi_1 + q^- P_{-1}\psi_{-1} - \left(q^+ + q^-\right)\left(P_1\psi_1 1\{P \leq c\} + P_{-1}\psi_{-1} 1\{P > c\}\right) \\
&= P_1\psi_1\left[q^+ - \left(q^+ + q^-\right) 1\{P \leq c\}\right] + P_{-1}\psi_{-1}\left[q^- - \left(q^+ + q^-\right) 1\{P > c\}\right] \\
&= P_1\psi_1\left[q^+ 1\{P > c\} - q^- 1\{P \leq c\}\right] \\
&+ P_{-1}\psi_{-1}\left[q^- 1\{P \leq c\} - q^+ 1\{P > c\}\right] \\
&= \left(P_1\psi_1 - P_{-1}\psi_{-1}\right)\left[q^+ 1\{P > c\} - q^- 1\{P \leq c\}\right] \\
&= \left|P_1\psi_1 - P_{-1}\psi_{-1}\right|\left[q^+ 1\{P > c\} + q^- 1\{P \leq c\}\right] \\
&\geq \epsilon.
\end{aligned}$$

Since $\inf_{\alpha\in\mathcal{S}(\epsilon)} R_{\text{MU-SVM}}(\alpha) \geq \epsilon$ for any P satisfying $\left|P_1\psi_1 - P_{-1}\psi_{-1}\right| > \epsilon$, and

$$\inf_{\alpha\in\mathcal{S}(\epsilon)} R_{\text{MU-SVM}}(\alpha) = \infty$$

for any P satisfying $\left|P_1\psi_1 - P_{-1}\psi_{-1}\right| \leq \epsilon$, we have

$$\inf_{P\in[0,1]} \inf_{\alpha\in\mathcal{S}(\epsilon)} R_{\text{MU-SVM}}(\alpha) \geq \epsilon.$$

Part (ii). It follows from Part (i) that, for any $\epsilon > 0$, if $R_{\text{MU}}(\alpha) \geq \epsilon$ for any $\alpha \in \mathbb{R}$, then we must have $R_{\text{MU-SVM}}(\alpha) \geq \epsilon$. Upon choosing $\epsilon = R_{\text{MU}}(\alpha)$, so that $R_{\text{MU}}(\alpha) \geq \epsilon$ holds trivially for any $\alpha \in \mathbb{R}$, we obtain:

$$R_{\text{MU-SVM}}(\alpha) \geq R_{\text{MU}}(\alpha)$$

for any $\alpha \in \mathbb{R}$.

Part (iii). We first make the dependence of $R_{\text{MU-SVM}}(\alpha)$ on x explicit. We write

$$R_{\text{MU-SVM}}(\alpha) = E\left[\ell_{\text{MU-SVM}}(Y,\alpha)\,|X = x\right] - E\left[\ell_{\text{MU-SVM}}\left(Y, \alpha^*_{\text{MU-SVM}}(x)\right)|X = x\right],$$

where $\alpha^*_{\text{MU-SVM}}(x)$ is the same as $\alpha^*_{\text{MU-SVM}}$, but its dependence on x is made explicit. For any $f \in \mathcal{M}$ and any $x \in \mathcal{X}$, by plugging in $\alpha = f(x)$, we obtain

$$R_{\text{MU-SVM}}(f(x)) = E\left[\ell_{\text{MU-SVM}}(Y, f(x))\,|X = x\right] - E\left[\ell_{\text{MU-SVM}}\left(Y, \alpha^*_{\text{MU-SVM}}(x)\right)|X = x\right].$$

Similarly, we have

$$R_{\text{MU}}(f(x)) = E\left[\ell_{\text{MU}}(Y, f(x))\,|X = x\right] - E\left[\ell_{\text{MU}}\left(Y, \alpha^*_{\text{MU}}(x)\right)|X = x\right].$$

By Part (ii), we know that

$$R_{\text{MU}}\left(f(x)\right) \leq R_{\text{MU-SVM}}\left(f(x)\right),$$

for any $f \in \mathcal{M}$ and any $x \in \mathcal{X}$. Taking the integral of the above with respect to the distribution $F_X(\cdot)$ of X, we obtain

$$\int_{\mathcal{X}} R_{\text{MU}}\left(f(x)\right) dF_X(x) \leq \int_{\mathcal{X}} R_{\text{MU-SVM}}\left(f(x)\right) dF_X(x),$$

i.e., $E\left[R_{\text{MU}}\left(f\left(X\right)\right)\right] \leq E\left[R_{\text{MU-SVM}}\left(f\left(X\right)\right)\right]$, as long as the integral $\int_{\mathcal{X}} R_{\text{MU-SVM}}\left(f(x)\right) dF_X(x)$ or, equivalently, the expectation $E\left[R_{\text{MU-SVM}}\left(f\left(X\right)\right)\right]$, is well-defined. ■

Proof of Theorem 4. Part (i). We write $\hat{f}(x) = \hat{\kappa}_0 + \hat{f}_{\mathcal{K}}(x) + \hat{f}_c(x)$, where $\hat{f}_{\mathcal{K}}(x) \in \mathcal{F}_{\mathcal{K}}$ and $\hat{f}_c(x) = \hat{\kappa}_c c_-(x) \in \mathcal{F}_{\mathcal{C}}$. By definition, the objective function evaluated at $\hat{f}$ is not greater than that evaluated at the constant function $f(x) \equiv q^+$. So

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[q_i - Y_i\hat{f}\left(X_i\right)\right]_+ \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu_n}{2}\left(\left\|\hat{f}_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \hat{\kappa}_c^2\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left[q_i - Y_i q^+\right]_+ \frac{\psi(Y_i, X_i)}{\bar{\psi}} = \frac{(q^- + q^+)}{n}\sum_{i:Y_i=-1}\frac{\psi(Y_i, X_i)}{\bar{\psi}}.
\end{aligned}$$

Using the same argument with $f(x) \equiv -q^-$ as the candidate function, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[q_i - Y_i\hat{f}\left(X_i\right)\right]_+ \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu_n}{2}\left(\left\|\hat{f}_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \hat{\kappa}_c^2\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left[q_i + Y_i q^-\right]_+ \frac{\psi(Y_i, X_i)}{\bar{\psi}} = \frac{(q^- + q^+)}{n}\sum_{i:Y_i=+1}\frac{\psi(Y_i, X_i)}{\bar{\psi}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[q_i - Y_i\hat{f}\left(X_i\right)\right]_+ \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu_n}{2}\left(\left\|\hat{f}_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \hat{\kappa}_c^2\right) \\
&\leq \left(q^- + q^+\right)\min\left\{\frac{1}{n}\sum_{i:Y_i=+1}\frac{\psi(Y_i, X_i)}{\bar{\psi}}, \frac{1}{n}\sum_{i:Y_i=-1}\frac{\psi(Y_i, X_i)}{\bar{\psi}}\right\} \\
&\leq \left(q^- + q^+\right)\frac{1}{2}\left\{\frac{1}{n}\sum_{i:Y_i=+1}\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{1}{n}\sum_{i:Y_i=-1}\frac{\psi(Y_i, X_i)}{\bar{\psi}}\right\} \\
&= \frac{q^- + q^+}{2}.
\end{aligned}$$

As a result, with probability one,

$$\frac{\mu_n}{2}\left\|\hat{f}\right\|_{\mathcal{K}c}^2 \leq \frac{q^- + q^+}{2},$$

and thus

$$\left\|\hat{f}\right\|_{\mathcal{K}c}^2 \leq \frac{q^- + q^+}{\mu_n}.$$

That is, with probability one, $\hat{f} \in \mathcal{F}_n$, and we can treat $\mathcal{F}_n$ as the parameter space.

Part (ii). We adopt the arguments in Bartlett and Mendelson (2002) and prove the results in two steps.

Step I: Use the Rademacher Analysis to obtain a high-probability upper bound for $Q_{\mathrm{MU}}(\hat{f})$.

To emphasize the dependence of $Q_{h_\gamma}(f)$ on the sample $\mathfrak{S} = \{(X_i, Y_i)\}_{i=1}^n$, we write it as $Q_{n,h_\gamma}(f; \mathfrak{S})$ when needed. To prove the upper bound in the theorem, we first provide an upper bound for the probability:

$$\Pr\left\{\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] - E\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] > \epsilon\right\}$$

using McDiarmid's inequality. The inequality requires that the object of interest does not change too much if only one observation in the sample is replaced. Let $\mathfrak{S}^i = [\cup_{j\neq i}(X_j, Y_j)] \cup (X_i^\circ, Y_i^\circ)$ be another sample, which is the same as $\mathfrak{S}$ but with the i-th pair (X_i, Y_i) replaced by (X_i°, Y_i°) drawn at random from the same population. Then

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] - \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f;\mathfrak{S}^i\right) - Q_{h_\gamma}(f)\right] \\
&= \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] \\
&- \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f) + \frac{h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right), Y_i^\circ\right)\psi(Y_i^\circ, X_i^\circ) - h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)\psi(Y_i, X_i)}{n}\right] \\
&\geq \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] - \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] \\
&- \sup_{f\in\mathcal{F}_n}\frac{h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right), Y_i^\circ\right)\psi(Y_i^\circ, X_i^\circ) - h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)\psi(Y_i, X_i)}{n} \\
&= -\sup_{f\in\mathcal{F}_n}\frac{h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right), Y_i^\circ\right)\psi(Y_i^\circ, X_i^\circ) - h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)\psi(Y_i, X_i)}{n}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f;\mathfrak{S}^i\right) - Q_{h_\gamma}(f)\right] - \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] \\
&\geq -\sup_{f\in\mathcal{F}_n}\frac{h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)\psi(Y_i, X_i) - h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right), Y_i^\circ\right)\psi(X_i^\circ, Y_i^\circ)}{n},
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}(f;\mathfrak{S}) - Q_{h_\gamma}(f)\right] - \sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f;\mathfrak{S}^i\right) - Q_{h_\gamma}(f)\right] \\
&\leq \sup_{f\in\mathcal{F}_n}\frac{h_\gamma\left(Y_i f\left(X_i\right), Y_i\right)\psi(Y_i, X_i) - h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right), Y_i^\circ\right)\psi(Y_i^\circ, X_i^\circ)}{n}.
\end{aligned}$$

Hence,

$$\left|\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f;\mathfrak{S}\right)-Q_{h_\gamma}\left(f\right)\right]-\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f;\mathfrak{S}^i\right)-Q_{h_\gamma}\left(f\right)\right]\right|$$
$$\leq\sup_{f\in\mathcal{F}_n}\frac{\left|h_\gamma\left(Y_if\left(X_i\right)\right)\psi(Y_i,X_i)-h_\gamma\left(Y_i^\circ f\left(X_i^\circ\right),Y_i^\circ\right)\psi(Y_i^\circ,X_i^\circ)\right|}{n}\leq\frac{\psi_{\max}}{n}.$$

Now, by McDiarmid's inequality (Lemma 26.4 in Shalev-Shwartz and Ben-David (2014), p. 328), we have

$$\Pr\left\{\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f\right)-Q_{h_\gamma}\left(f\right)\right]-E\sup_{f\in\mathcal{F}_n}\left[Q_{n,h_\gamma}\left(f\right)-Q_{h_\gamma}\left(f\right)\right]>\epsilon\right\}\leq\exp\left(-\frac{2n\epsilon^2}{\psi_{\max}^2}\right).$$

That is, with probability at least $1-\exp\left(-\frac{2n\epsilon^2}{\psi_{\max}^2}\right)$,

$$\sup_{f\in\mathcal{F}_n}\left[Q_{h_\gamma}\left(f\right)-Q_{n,h_\gamma}\left(f\right)\right]<E\sup_{f\in\mathcal{F}_n}\left[Q_{h_\gamma}\left(f\right)-Q_{n,h_\gamma}\left(f\right)\right]+\epsilon.$$

For any $\delta\in(0,1)$, choose ϵ to satisfy $\exp\left(-\frac{2n\epsilon^2}{\psi_{\max}^2}\right)=\frac{\delta}{2}$ or $\epsilon=\psi_{\max}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$. Then the above implies that, with probability at least $1-\frac{\delta}{2}$, we have

$$\begin{aligned}Q_{\mathrm{MU}}(\hat{f})&\leq Q_{n,h_\gamma}(\hat{f})+Q_{h_\gamma}(\hat{f})-Q_{n,h_\gamma}(\hat{f})\\&\leq Q_{n,h_\gamma}(\hat{f})+E\sup_{f\in\mathcal{F}_n}\left[Q_{h_\gamma}\left(f\right)-Q_{n,h_\gamma}\left(f\right)\right]+\psi_{\max}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.\end{aligned}$$

Using Lemma 26.2 in Shalev-Shwartz and Ben-David (2014) (p. 326), we have

$$E\sup_{f\in\mathcal{F}_n}\left[Q_{h_\gamma}\left(f\right)-Q_{n,h_\gamma}\left(f\right)\right]\leq2E_{\sigma,\mathfrak{S}}\sup_{f\in\mathcal{F}_n}\left[\frac{1}{n}\sum_{i=1}^n\sigma_ih_\gamma\left(Y_if\left(X_i\right),Y_i\right)\psi(Y_i,X_i)\right],$$

where $\{\sigma_i\}$ are independent Rademacher random variables (i.e., $\Pr\left(\sigma_i=1\right)=\Pr\left(\sigma_i=-1\right)=1/2$) and the expectation in the upper bound is taken with respect to the distributions of $\sigma=\{\sigma_i\}$ and the sample $\mathfrak{S}=\{(X_i,Y_i)\}_{i=1}^n$.

Define

$$\mathfrak{H}_{\gamma,i}\left(r\right)=h_\gamma\left(r,Y_i\right)\psi(Y_i,X_i).$$

Then

$$E\sup_{f\in\mathcal{F}_n}\left[Q_{h_\gamma}\left(f\right)-Q_{n,h_\gamma}\left(f\right)\right]\leq2E_{\sigma,\mathfrak{S}}\sup_{f\in\mathcal{F}_n}\left[\frac{1}{n}\sum_{i=1}^n\sigma_i\mathfrak{H}_{\gamma,i}\left(t_i\right)\right]\text{ for }t_i=Y_if(X_i).$$

Note that

$$\begin{aligned}\left|\mathfrak{H}_{\gamma,i}\left(r_1\right)-\mathfrak{H}_{\gamma,i}\left(r_2\right)\right|&=\left|h_\gamma\left(r_1,Y_i\right)\psi(Y_i,X_i)-h_\gamma\left(r_2,Y_i\right)\psi(Y_i,X_i)\right|\\&=\left|h_\gamma\left(r_1,Y_i\right)-h_\gamma\left(r_2,Y_i\right)\right|\psi(Y_i,X_i)\\&\leq\max\left(\frac{\psi_{\max}^+}{\gamma^+},\frac{\psi_{\max}^-}{\gamma^-}\right)\left|r_1-r_2\right|.\end{aligned}$$

Using Lemma 26.9 of Shalev-Shwartz and Ben-David (2014) (p. 331), which is a version of Talagrand's contraction lemma (Ledoux and Talagrand (1991)), we have

$$E_{\sigma,\mathfrak{S}} \sup_{f\in\mathcal{F}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathfrak{H}_{\gamma,i}\left(t_i\right)\right] \leq \max\left(\frac{\psi_{\max}^{+}}{\gamma^{+}},\frac{\psi_{\max}^{-}}{\gamma^{-}}\right) E_{\sigma,\mathfrak{S}} \sup_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(X_i\right).$$

To obtain an upper bound for $E_{\sigma,\mathfrak{S}}\sup_{f\in\mathcal{F}_n} n^{-1}\sum_{i=1}^{n}\sigma_i f\left(X_i\right)$, we define the function class

$$\mathcal{F}_{\mathcal{K}c,n} := \left\{ f_{\mathcal{K}} + f_c : f_{\mathcal{K}}\in\mathcal{F}_{\mathcal{K}}, f_c = c_{-}(x)\kappa_c \in \mathcal{F}_c, \text{ and } \left\|f_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \kappa_c^2 \leq \frac{q^{+}+q^{-}}{\mu_n}\right\}.$$

Then any $f\in\mathcal{F}_n$ can be represented by $f=\kappa_0+f_{\mathcal{K}c}$ where $\kappa_0\in\mathbb{K}_0$ and $f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}$. So

$$\begin{aligned}
E_{\sigma,\mathfrak{S}} \sup_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(X_i\right) &= \left\{ E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n},\kappa_0\in\mathbb{K}_0}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[f_{\mathcal{K}c}\left(X_i\right)+\kappa_0\right]\right\} \\
&\leq E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) + E_{\sigma}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right| \sup_{\kappa_0\in\mathbb{K}_0}\left|\kappa_0\right| \\
&\leq E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) + \frac{1}{\sqrt{n}} \sup_{\kappa_0\in\mathbb{K}_0}\left|\kappa_0\right|,
\end{aligned}$$

where the second inequality holds because

$$E_{\sigma}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right| \leq \left[E_{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)^2\right]^{1/2} = \frac{1}{\sqrt{n}}.$$

Therefore,

$$\begin{aligned}
&E \sup_{f\in\mathcal{F}_n}\left[Q_h\left(f\right)-Q_{n,h}\left(f\right)\right] \\
&\leq 2\max\left(\frac{\psi_{\max}^{+}}{\gamma^{+}},\frac{\psi_{\max}^{-}}{\gamma^{-}}\right)\left(E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) + \frac{1}{\sqrt{n}} \sup_{\kappa_0\in\mathbb{K}_0}\left|\kappa_0\right|\right).
\end{aligned}$$

It then follows that with probability at least $1-\delta/2$, we have

$$\begin{aligned}
&Q_{\mathrm{MU}}(\hat{f}) \leq Q_{n,h_\gamma}(\hat{f}) \\
&\quad + 2\max\left(\frac{\psi_{\max}^{+}}{\gamma^{+}},\frac{\psi_{\max}^{-}}{\gamma^{-}}\right)\left(E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) + \frac{1}{\sqrt{n}} \sup_{\kappa_0\in\mathbb{K}_0}\left|\kappa_0\right|\right) + \psi_{\max}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
\end{aligned}$$

But, by McDiarmid's inequality, we obtain:

$$\Pr\left\{E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) - E_{\sigma} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) > \epsilon\right\} < \exp\left(-\frac{2n\epsilon^2}{f_{\mathcal{K}c,\max}^2}\right),$$

where $\left|f_{\mathcal{K}c,\max}\right| = \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\sup_{x\in\mathcal{X}}\left|f_{\mathcal{K}c}\left(x\right)\right|$ and so

$$\Pr\left\{E_{\sigma,\mathfrak{S}} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) - E_{\sigma} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) > \left|f_{\mathcal{K}c,\max}\right|\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right\} < \frac{\delta}{2}.$$

Combining the above analysis, we obtain, with probability at least $1-\delta$,

$$
\begin{aligned}
Q_{\mathrm{MU}}(\hat{f}) \leq\ & Q_{n,h_\gamma}(\hat{f}) \\
& + 2\max\left(\frac{\psi_{\max}^{+}}{\gamma^{+}}, \frac{\psi_{\max}^{-}}{\gamma^{-}}\right)\left(E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) + \left|f_{\mathcal{K}c,\max}\right|\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{1}{\sqrt{n}}\sup_{\kappa_0\in\mathbb{K}_0}\left|\kappa_0\right|\right) \\
& + \psi_{\max}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
\end{aligned}
$$

Step II: Derive upper bounds for $E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right)$ and $\left|f_{\mathcal{K}c,\max}\right|$. Let $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ be the inner product on the RKHS associated with the kernel $\mathcal{K}\left(\cdot,\cdot\right)$. We have

$$
\begin{aligned}
& E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{K}c}\left(X_i\right) \\
&= E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \left(f_{\mathcal{K}}\left(X_i\right) + c_{-}(X_i)\kappa_c\right) \\
&= \frac{1}{n}E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \left[\left\langle\sum_{i=1}^{n}\sigma_i\mathcal{K}\left(X_i,\cdot\right), f_{\mathcal{K}}\left(\cdot\right)\right\rangle_{\mathcal{K}} + \sum_{i=1}^{n}\sigma_i c_{-}(X_i)\kappa_c\right] \\
&\leq \frac{1}{n}E_\sigma \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \left[\left\|\sum_{i=1}^{n}\sigma_i\mathcal{K}\left(X_i,\cdot\right)\right\|_{\mathcal{K}}\left\|f_{\mathcal{K}}\right\|_{\mathcal{K}} + \left|\sum_{i=1}^{n}\sigma_i c_{-}(X_i)\right|\left|\kappa_c\right|\right] \\
&\leq \frac{1}{n}E_\sigma \left(\left\|\sum_{i=1}^{n}\sigma_i\mathcal{K}\left(X_i,\cdot\right)\right\|_{\mathcal{K}}^2 + \left|\sum_{i=1}^{n}\sigma_i c_{-}(X_i)\right|^2\right)^{1/2} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \left(\left\|f_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \kappa_c^2\right)^{1/2} \\
&= \frac{1}{n}\left(E_\sigma\left\|\sum_{i=1}^{n}\sigma_i\mathcal{K}\left(X_i,\cdot\right)\right\|_{\mathcal{K}}^2 + \left|\sum_{i=1}^{n}\sigma_i c_{-}(X_i)\right|^2\right)^{1/2} \sup_{f_{\mathcal{K}c}\in\mathcal{F}_{\mathcal{K}c,n}} \left(\left\|f_{\mathcal{K}}\right\|_{\mathcal{K}}^2 + \kappa_c^2\right)^{1/2} \\
&\leq \frac{1}{n}\left\{E_\sigma\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j\left[\mathcal{K}\left(X_i,X_j\right) + c_{-}(X_i)c_{-}\left(X_j\right)\right]\right\}^{1/2}\sqrt{\frac{q^{+}+q^{-}}{\mu_n}} \\
&\leq \frac{1}{n}\sqrt{\sum_{i=1}^{n}\mathcal{K}\left(X_i,X_i\right) + c_{-}(X_i)c_{-}\left(X_i\right)}\sqrt{\frac{q^{+}+q^{-}}{\mu_n}} \\
&= \frac{1}{\sqrt{n}}\sqrt{\frac{q^{+}+q^{-}}{\mu_n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{K}_c\left(X_i,X_i\right)}.
\end{aligned}
$$

Next, for any $f_{\mathcal{K}c} \in \mathcal{F}_{\mathcal{K}c,n}$, we have

$$\begin{aligned}
|f_{\mathcal{K}c}(x)| &= |f_{\mathcal{K}}(x) + c_-(x)\kappa_c| \leq |f_{\mathcal{K}}(x)| + |c_-(x)| \cdot |\kappa_c| \\
&= \langle \mathcal{K}(x,\cdot), f_{\mathcal{K}}(\cdot)\rangle_{\mathcal{K}} + |c_-(x)| \cdot |\kappa_c| \\
&\leq \|\mathcal{K}(x,\cdot)\|_{\mathcal{K}} \|f(\cdot)\|_{\mathcal{K}} + |c_-(x)| \cdot |\kappa_c| \\
&= \sqrt{\mathcal{K}(x,x)}\, \|f(\cdot)\|_{\mathcal{K}} + |c_-(x)| \cdot |\kappa_c| \\
&\leq \sqrt{\mathcal{K}(x,x) + c_-(x)^2}\sqrt{\|f(\cdot)\|_{\mathcal{K}} + \kappa_c^2} \\
&\leq \sqrt{\mathcal{K}_c(x,x)}\sqrt{\frac{q^+ + q^-}{\mu_n}}
\end{aligned}$$

for any $x \in \mathcal{X}$. Hence, $|f_{\mathcal{K}c,\max}| \leq \sqrt{\frac{q^+ + q^-}{\mu_n} \max_{x\in\mathcal{X}} \mathcal{K}_c(x,x)}$.

Combining the results in the above two steps, we have

$$Q_{\mathrm{MU}}(\hat{f}) \leq Q_{n,h_\gamma}(\hat{f}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi^+_{\max}}{\gamma^+}, \frac{\psi^-_{\max}}{\gamma^-}\right) \mathcal{V}(\mathcal{F}_n, n, \delta) + \frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{\delta}},$$

with probability at least $1-\delta$. ■

Proof of Proposition 5. For any $\gamma_1 = \left(\gamma_1^+, \gamma_1^-\right), \gamma_2 = \left(\gamma_2^+, \gamma_2^-\right), \delta \in (0,1]$, define the event:

$$\mathcal{E}_n(\gamma_1,\gamma_2,\delta) = \left\{ Q_{\mathrm{MU}}(\hat{f}) > Q_{n,h_{\gamma_1}}(\hat{f}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi^+_{\max}}{\gamma_2^+}, \frac{\psi^-_{\max}}{\gamma_2^-}\right) \cdot \mathcal{V}(\mathcal{F}_n, n, \delta) + \frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{\delta}} \right\}.$$

Note that the lower bound in the definition of $\mathcal{E}_n(\gamma_1,\gamma_2,\delta)$ is increasing in γ_1 and decreasing in γ_2 and δ. We have

$$\begin{aligned}
&\mathcal{E}_n(\gamma_1,\gamma_2,\delta) \subseteq \mathcal{E}_n(\tilde{\gamma}_1,\gamma_2,\delta) \text{ for } \tilde{\gamma}_1 \leq \gamma_1, \\
&\mathcal{E}_n(\gamma_1,\gamma_2,\delta) \subseteq \mathcal{E}_n(\gamma_1,\tilde{\gamma}_2,\delta) \text{ for } \tilde{\gamma}_2 \geq \gamma_2, \\
&\mathcal{E}_n(\gamma_1,\gamma_2,\delta) \subseteq \mathcal{E}_n(\gamma_1,\gamma_2,\tilde{\delta}) \text{ for } \tilde{\delta} \geq \delta,
\end{aligned}$$

where, for example, $\tilde{\gamma}_1 \leq \gamma_1$ is an elementwise inequality. Using the above results, we have

$$\begin{aligned}
&\Pr\left(Q_{\mathrm{MU}}(\hat{f}) > Q_{n,h_\gamma}(\hat{f}) + \frac{4}{\sqrt{n}}\max\left(\frac{\psi^+_{\max}}{\gamma^+}, \frac{\psi^-_{\max}}{\gamma^-}\right)\cdot\mathcal{V}\left(\mathcal{F}_n, n, \frac{\delta\gamma^+\gamma^-}{4q^+q^-}\right)\right.\\
&\left.+\frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{8q^+q^-}{\delta\gamma^+\gamma^-}} \text{ for some } \gamma := \left(\gamma^+,\gamma^-\right) \in (0,q^+]\otimes(0,q^-]\right)\\
&= \Pr\left(\mathcal{E}_n(\gamma, \frac{\gamma}{2}, \frac{\delta}{4}\frac{\gamma^+\gamma^-}{q^+q^-}) \text{ for some } \gamma \in (0,q^+]\otimes(0,q^-]\right)\\
&\leq \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty}\Pr\left\{\mathcal{E}_n(\gamma, \frac{\gamma}{2}, \frac{\delta}{4}\frac{\gamma^+\gamma^-}{q^+q^-}) \text{ for some } \gamma \in \left(\frac{q^+}{2^{\ell+1}}, \frac{q^+}{2^\ell}\right]\otimes\left(\frac{q^-}{2^{m+1}}, \frac{q^-}{2^m}\right]\right\}\\
&\leq \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty}\Pr\left\{\mathcal{E}_n\left(\left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \frac{1}{2}\left(\frac{q^+}{2^\ell}, \frac{q^-}{2^m}\right), \frac{\delta}{4q^+q^-}\frac{q^+}{2^\ell}\frac{q^-}{2^m}\right)\right\}\\
&= \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty}\Pr\left\{\mathcal{E}_n\left(\left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \frac{\delta}{4}\frac{1}{2^\ell}\frac{1}{2^m}\right)\right\}\\
&\leq \frac{1}{4}\sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty}\frac{1}{2^\ell}\frac{1}{2^m}\delta = \delta,
\end{aligned}$$

where the second inequality follows from the fact that

$$\mathcal{E}_n(\gamma, \frac{\gamma}{2}, \frac{\delta}{4}\frac{\gamma^+\gamma^-}{q^+q^-}) \subseteq \mathcal{E}_n\left(\left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \frac{1}{2}\left(\frac{q^+}{2^\ell}, \frac{q^-}{2^m}\right), \frac{\delta}{4q^+q^-}\frac{q^+}{2^\ell}\frac{q^-}{2^m}\right)$$

for any $\gamma \in \left(\frac{q^+}{2^{\ell+1}}, \frac{q^+}{2^\ell}\right]\otimes\left(\frac{q^-}{2^{m+1}}, \frac{q^-}{2^m}\right]$ and the last inequality follows from $\Pr\left(\mathcal{E}_n(\gamma,\gamma,\delta)\right) \leq \delta$ for any $\gamma \in (0,q^+]\otimes(0,q^-]$ and $\delta \in (0,1]$. It then follows that for any $\hat{\gamma}$ that may be data-dependent, we have

$$\begin{aligned}
&\Pr\left(Q_{\mathrm{MU}}(\hat{f}) > Q_{n,h_{\hat{\gamma}}}(\hat{f}) + \frac{4}{\sqrt{n}}\max\left(\frac{\psi^+_{\max}}{\hat{\gamma}^+}, \frac{\psi^-_{\max}}{\hat{\gamma}^-}\right)\cdot\mathcal{V}\left(\mathcal{F}_n, n, \frac{\delta\hat{\gamma}^+\hat{\gamma}^-}{4q^+q^-}\right) + \frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{8q^+q^-}{\delta\gamma^+\gamma^-}}\right)\\
&\leq \Pr\left(Q_{\mathrm{MU}}(\hat{f}) > Q_{n,h_\gamma}(\hat{f}) + \frac{4}{\sqrt{n}}\max\left(\frac{\psi^+_{\max}}{\gamma^+}, \frac{\psi^-_{\max}}{\gamma^-}\right)\cdot\mathcal{V}\left(\mathcal{F}_n, n, \frac{\delta\gamma^+\gamma^-}{4q^+q^-}\right)\right.\\
&\left.+\frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{8q^+q^-}{\delta\gamma^+\gamma^-}} \text{ for some } \gamma = \left(\gamma^+,\gamma^-\right) \in (0,q^+]\otimes(0,q^-]\right)\\
&\leq \delta.
\end{aligned}$$

■

Proof of Corollary 6. Part (i). Define the event:

$$\begin{aligned}
&\mathcal{E}_{n,\ell}(\hat{f}_{\mu_{n,\ell}}, \gamma, \delta)\\
&= \left\{Q_{\mathrm{MU}}(\hat{f}_{\mu_{n,\ell}}) > Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\ell}}) + \frac{2}{\sqrt{n}}\max\left(\frac{\psi^+_{\max}}{\gamma^+}, \frac{\psi^-_{\max}}{\gamma^-}\right)\mathcal{V}\left(\mathcal{F}^{\mu_{n,\ell}}, n, p_\ell\delta\right) + \frac{\psi_{\max}}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{p_\ell\delta}}\right\}.
\end{aligned}$$

Using Theorem 4(ii), we obtain $\Pr[\mathcal{E}_{n,\ell}(\hat{f}_{\mu_{n,\ell}}, \gamma, \delta)] \leq \delta p_\ell$ for all $\ell = 1, 2, \ldots, L_n$. Then

$$\begin{aligned}\Pr\left[\mathcal{E}_{n,\hat{\ell}}(\hat{f}_{\mu_{n,\hat{\ell}}}, \gamma, \delta)\right] &\leq \Pr\left[\mathcal{E}_{n,\ell}(\hat{f}_{\mu_{n,\ell}}, \gamma, \delta) \text{ for at least one } \ell = 1, 2, \ldots, L_n\right] \\ &\leq \sum_{\ell=1}^{L_n} \Pr\left[\mathcal{E}_{n,\ell}(\hat{f}_{\mu_{n,\ell}}; \gamma, \delta)\right] \leq \sum_{\ell=1}^{L_n} p_\ell \delta \leq \delta,\end{aligned}$$

which is equivalent to the result in Corollary 6(i).

Part (ii). For any $\gamma_1 = \left(\gamma_1^+, \gamma_1^-\right), \gamma_2 = \left(\gamma_2^+, \gamma_2^-\right), \delta \in (0, 1]$, define the event:

$$\begin{aligned}&\tilde{\mathcal{E}}_{n,\hat{\ell}}(\gamma_1, \gamma_2, \delta) \\ = \ &\left\{Q_{\mathrm{MU}}(\hat{f}_{\mu_{n,\hat{\ell}}}) > Q_{n,h_{\gamma_1}}(\hat{f}_{\mu_{n,\hat{\ell}}}) + \frac{2}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\gamma_2^+}, \frac{\psi_{\max}^-}{\gamma_2^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, p_{\hat{\ell}}\delta\right) + \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2} \log \frac{2}{p_{\hat{\ell}}\delta}}\right\}.\end{aligned}$$

Using the same argument as in the proof Proposition 5, we can see that the set $\tilde{\mathcal{E}}_{n,\hat{\ell}}(\gamma_1, \gamma_2, \delta)$ becomes larger for a smaller γ_1, a larger γ_2 or a larger δ. So, using Part (i), we have

$$\begin{aligned}&\Pr\left\{Q_{\mathrm{MU}}(\hat{f}_{\mu_{n,\hat{\ell}}}) > Q_{n,h_\gamma}(\hat{f}_{\mu_{n,\hat{\ell}}}) + \frac{4}{\sqrt{n}} \max\left(\frac{\psi_{\max}^+}{\hat{\gamma}^+}, \frac{\psi_{\max}^-}{\hat{\gamma}^-}\right) \mathcal{V}\left(\mathcal{F}^{\mu_{n,\hat{\ell}}}, n, \frac{p_{\hat{\ell}}\hat{\gamma}^+\hat{\gamma}^-}{4q^+q^-}\delta\right)\right. \\ &\left. + \frac{\psi_{\max}}{\sqrt{n}} \sqrt{\frac{1}{2} \log \frac{8q^+q^-}{p_{\hat{\ell}}\hat{\gamma}^+\hat{\gamma}^-\delta}}\right\} \\ &= \Pr\left(\tilde{\mathcal{E}}_{n,\hat{\ell}}(\gamma, \frac{\gamma}{2}, \frac{\delta}{4}\frac{\gamma^+\gamma^-}{q^+q^-}) \text{ for some } \gamma \in (0, q^+] \otimes (0, q^-]\right) \\ &\leq \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty} \Pr\left\{\tilde{\mathcal{E}}_{n,\hat{\ell}}(\gamma, \frac{\gamma}{2}, \frac{\delta}{4}\frac{\gamma^+\gamma^-}{q^+q^-}) \text{ for some } \gamma \in \left(\frac{q^+}{2^{\ell+1}}, \frac{q^+}{2^\ell}\right] \otimes \left(\frac{q^-}{2^{m+1}}, \frac{q^-}{2^m}\right]\right\} \\ &= \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty} \Pr\left\{\tilde{\mathcal{E}}_{n,\hat{\ell}}\left(\left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \left(\frac{q^+}{2^{\ell+1}}, \frac{q^-}{2^{m+1}}\right), \frac{\delta}{4}\frac{1}{2^\ell}\frac{1}{2^m}\right)\right\} \\ &= \sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty} \Pr\left\{\mathcal{E}_{n,\hat{\ell}}\left(\hat{f}_{\mu_{n,\hat{\ell}}}, \left(\frac{1}{2^{\ell+1}}, \frac{1}{2^{m+1}}\right), \frac{\delta}{4}\frac{1}{2^\ell}\frac{1}{2^m}\right)\right\} \\ &\leq \frac{1}{4}\sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty} \frac{1}{2^\ell}\frac{1}{2^m}\delta = \delta,\end{aligned}$$

which implies the desired result. ■

References

Bach, F. R., Heckerman, D., and Horvitz, E. (2006). Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7(63):1713–1741.

Bartlett, P. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1):85–113.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Haussler, D., editor, *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 144–152. ACM Press.

Brefeld, U., Geibel, P., and Wysotzki, F. (2003). Support vector machines with example dependent costs. In Lavrač, N., Gamberger, D., Blockeel, H., and Todorovski, L., editors, *Machine Learning: ECML 2003*, pages 23–34. Springer.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, Cambridge, UK.

Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006). Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 201–208, New York, NY, USA.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20:273–297.

Elliott, G. and Lieli, R. P. (2013). Predicting binary outcomes. *Journal of Econometrics*, 174(1):15–26.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer, 1st edition.

Fung, G. M. and Mangasarian, O. L. (2004). A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications*, 28(2):185–202.

Granger, C. W. and Machina, M. J. (2006). Forecasting and decision theory. volume 1 of *Handbook of Economic Forecasting*, pages 81–98. Elsevier.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification. Technical report. Working paper, Department of Computer Science and Information Engineering, National Taiwan University.

Iranmehr, A., Masnadi-Shirazi, H., and Vasconcelos, N. (2019). Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer.

Lin, Y., Lee, Y., and Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Machine Learning*, 46(1):191–202.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, 37(26):3776–3788.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

Massart, P. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, 1st edition.

Su, J.-H. (2021). Model selection in utility-maximizing binary prediction. *Journal of Econometrics*, 223(1):96–124.

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(499):1106–1118.

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

Online Supplementary Appendix

Title: Support Vector Decision Making
Author: Yixiao Sun

S.1 Additional Proofs

Proof of Lemma 1. Let d_i be the geometric distance from a point W_i to the middle hyperplane $w'\theta_w = \bar{\theta}_0$ and let $W_\circ$ be the point on the middle hyperplane that is closest to W_i. Then

$$W_\circ'\theta_w = \bar{\theta}_0, \tag{S.1}$$

and $W_i - W_\circ$ and θ_w are collinear so that

$$W_i - W_\circ = Y_i \frac{\theta_w}{\|\theta_w\|} d_i. \tag{S.2}$$

It follows from equation (S.2) that

$$W_\circ = W_i - Y_i \frac{\theta_w}{\|\theta_w\|} d_i.$$

Plugging the above into (S.1) yields

$$W_i'\theta_w - Y_i \|\theta_w\| d_i = \bar{\theta}_0.$$

Solving for d_i, we have

$$d_i = \frac{W_i'\theta_w - \bar{\theta}_0}{Y_i \|\theta_w\|} = \frac{Y_i \left(W_i'\theta_w - \bar{\theta}_0\right)}{Y_i^2 \|\theta_w\|} = \frac{Y_i \left(W_i'\theta_w - \bar{\theta}_0\right)}{\|\theta_w\|}.$$

For a point W_i on the positive hyperplane, we have $W_i'\theta_w = \theta_0^+$ and $Y_i = +1$. So, for this point,

$$d_i = \frac{Y_i \left(\theta_0^+ - \bar{\theta}_0\right)}{\|\theta_w\|} = \frac{\theta_0^+ - \rho\theta_0^+ - (1-\rho)\,\theta_0^-}{\|\theta_w\|} = \frac{(1-\rho)\left(\theta_0^+ - \theta_0^-\right)}{\|\theta_w\|}.$$

Similarly, for a point W_j on the negative hyperplane, we have

$$d_j = \frac{Y_j \left(\theta_0^- - \bar{\theta}_0\right)}{\|\theta_w\|} = -\frac{\theta_0^- - \rho\theta_0^+ - (1-\rho)\,\theta_0^-}{\|\theta_w\|} = \frac{\rho\left(\theta_0^+ - \theta_0^-\right)}{\|\theta_w\|}.$$

The geometric distance between the two hyperplanes is then equal to

$$d_i + d_j = \frac{\theta_0^+ - \theta_0^-}{\|\theta_w\|}.$$

■

S.2 MU-SVM in the Matrix Form

S.2.1 The primal problem in the linear case

The primal problem in the linear MU-SVM case is

$$\min_{\kappa_0,\kappa_w,\{\xi_i\}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\psi(Y_i,X_i)}{\bar{\psi}}\xi_i + \frac{\mu}{2}\left\|\kappa_w\right\|^2 \right\} \text{ subject to}$$
$$Y_i\left(W_i'\kappa_w+\kappa_0\right)+\xi_i \geq q_i,\ \xi_i \geq 0,\ \text{for } i\in[n] \text{ and } \kappa_c \geq 0. \tag{S.3}$$

To write this quadratic programming problem in a matrix form, we let

$$H_{11} = \begin{pmatrix} O_{1\times 1} & O_{d_w\times 1} \\ O_{1\times d_w} & I_{d_w\times d_w} \end{pmatrix}_{d_{\tilde{w}}\times d_{\tilde{w}}}, H = \begin{pmatrix} H_{11} & O_{d_{\tilde{w}}\times n} \\ O_{n\times d_{\tilde{w}}} & O_{n\times n} \end{pmatrix},$$
$$\Psi^{\circ} = \frac{1}{n\bar{\psi}}[O_{1\times d_{\tilde{w}}}, \psi\left(Y_1,X_1\right),\ldots,\psi\left(Y_n,X_n\right)]',$$
$$\boldsymbol{q} = (q_1,\ldots,q_n)',$$
$$G_{11} = \begin{pmatrix} Y_1(1,W_1'), \\ \ldots \\ Y_i(1,W_i'), \\ \ldots \\ Y_n(1,W_n'), \end{pmatrix}_{n\times d_{\tilde{w}}}, \ G = -\begin{pmatrix} G_{11}, & I_n \\ O_{n\times d_{\tilde{w}}}, & I_n \\ [O_{1\times d_w},1], & O_{1\times n} \end{pmatrix}, g = -\begin{pmatrix} \boldsymbol{q}_{n\times 1} \\ O_{n\times 1} \\ O_{1\times 1} \end{pmatrix},$$
$$\text{and } u = \left(\tilde{\kappa}_w', \xi_1,\ldots,\xi_n\right)',$$

where $d_{\tilde{w}} = d_w + 1$, $\tilde{\kappa}_w = (\kappa_0, \kappa_w')'$, $\kappa_w = (\kappa_x', \kappa_c')'$ and $O_{\text{row,col}}$ stands for a row $\times$ col matrix of zeros. The primal minimization problem can then be written as the quadratic programming problem:

$$\min_{u\in\mathbb{R}^{d_{\tilde{w}}+n}} (\Psi^{\circ})'u + \frac{\mu}{2}u'Hu \text{ subject to } Gu \leq g.$$

This is the required form for using quadratic programming in software packages such as Matlab.

S.2.2 The dual problem in the kernel case

Recall that the dual problem in the kernel case is

$$\max_{\lambda_1,\ldots,\lambda_n,\lambda_c} L_D\left(\lambda,\lambda_c\right) \text{ subject to}$$
$$\sum_{i=1}^{n}\lambda_i Y_i = 0,\ \lambda_c \geq 0, \text{ and}$$
$$0 \leq \lambda_i \leq \frac{1}{n}\frac{\psi(Y_i,X_i)}{\bar{\psi}},\ \text{for all } i\in[n], \tag{S.4}$$

where

$$L_D\left(\lambda,\lambda_c\right) = \sum_{i=1}^{n} q_i\lambda_i - \frac{1}{2\mu}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \mathcal{K}_c\left(X_i,X_j\right) + 2\lambda_c\sum_{i=1}^{n}\lambda_i Y_i c_{-}(X_i) + \lambda_c^2\right].$$

Let $\lambda = (\lambda_1, \ldots, \lambda_n)'$ and $u = \left(\lambda', \lambda_c'\right)'$ be the choice variable. Denote

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{S.5}$$

where

$$\begin{aligned} H_{11} &= -\begin{pmatrix} Y_1 Y_1 \mathcal{K}_c(X_1, X_1) & \ldots & Y_1 Y_n \mathcal{K}_c(X_1, X_n) \\ Y_2 Y_1 \mathcal{K}_c(X_2, X_1) & \ldots & Y_2 Y_n \mathcal{K}_c(X_2, X_n) \\ \ldots & \ldots & \ldots \\ Y_n Y_1 \mathcal{K}_c(X_n, X_1) & \ldots & Y_n Y_n \mathcal{K}_c(X_n, X_n) \end{pmatrix}, \; H_{12} = -\begin{pmatrix} Y_1 c_-(X_1) \\ \ldots \\ Y_i c_-(X_i) \\ \ldots \\ Y_n c_-(X_n) \end{pmatrix}, \\ H_{21} &= H_{12}', \; H_{22} = -1. \end{aligned}$$

Also, let

$$\boldsymbol{q}_\circ = [\underbrace{q_1, q_2, \ldots, q_n}_{1\times n}, 0]'.$$

Then the objective function becomes $L_D(\lambda, \lambda_c) = \boldsymbol{q}_\circ' u + \frac{1}{2\mu} u' H u$. Next, we let

$$G = \begin{pmatrix} -I_{n\times n} & O_{n\times 1} \\ O_{1\times n} & -1 \\ I_{n\times n} & O_{n\times 1} \end{pmatrix}, g = \begin{pmatrix} O_{(n+1)\times 1} \\ \frac{1}{n\psi}\psi(Y_1, X_1) \\ \ldots. \\ \frac{1}{n\psi}\psi(Y_n, X_n) \end{pmatrix},$$

and

$$G_{eq} = \begin{pmatrix} Y_1, & \ldots & Y_n, & 0 \end{pmatrix}, g_{eq} = \begin{pmatrix} 0 \end{pmatrix}.$$

Then the inequality and equality constraints can be written as $Gu \leq g$ and $G_{eq} u = g_{eq}$, respectively. With the above definitions of $H, \boldsymbol{q}_\circ, G, g, G_{eq}$, and g_{eq}, the dual problem becomes

$$\max_u \left(\boldsymbol{q}_\circ' u + \frac{1}{2\mu} u' H u \right) \text{ subject to } Gu \leq g \text{ and } G_{eq} u = g_{eq}.$$

This can be solved using standard quadratic programming packages. It is important to note that when dealing with the dual problem, the objective function must be maximized. Therefore, if a program is designed to minimize an objective function, we need to flip the sign and minimize $-\boldsymbol{q}_\circ' u - u' H u / (2\mu)$.

Proposition S.1 *If the matrix $(\mathcal{K}(X_i, X_j))_{n\times n}$ is positive definite, then H given in (S.5) is negative definite, and hence (S.4) has a unique solution.*

Proof of Proposition S.1 . For any $u = \left(\lambda', \lambda_c'\right)'$, we have

$$u' H u = u' \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} u = H_{22} \lambda_c^2 + 2\lambda_c \lambda' H_{12} + \lambda' H_{11} \lambda.$$

This is a quadratic equation in λ_c with a negative quadratic coefficient ($H_{22} = -1$). The discriminator is

$$\begin{aligned}\Delta &= 4\lambda' H_{12} H_{21} \lambda - 4H_{22}\left(\lambda' H_{11} \lambda\right) \\ &= 4\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j c_-(X_i)\, c_-(X_j) - 4\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \mathcal{K}_c(X_i, X_j) \\ &= -4\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \mathcal{K}(X_i, X_j).\end{aligned}$$

Under the assumption that $(\mathcal{K}(X_i, X_j))_{n\times n}$ is positive definite, we have $\Delta < 0$ for all λ such that $\lambda \odot Y := (\lambda_1 Y_1, \ldots, \lambda_n Y_n)' \neq 0$.

Now, for any $u \neq 0$, we have either ($\lambda \neq 0$) or ($\lambda = 0$ and $\lambda_c \neq 0$).

- When $\lambda \neq 0$, we have $\lambda \odot Y \neq 0$, and so $u'Hu < 0$.
- When $\lambda = 0$ but $\lambda_c \neq 0$, we have $u'Hu = H_{22}\lambda_c^2 = -\lambda_c^2 < 0$.

We have shown that $u'Hu < 0$ for any $u \neq 0$. That is, H is negative definite. This implies that the quadratic programming in (S.4) has a unique solution. ■

S.3 Characterizing the Solution to the Linear MU-SVM

Recall that the linear MU-SVM problem is

$$\begin{aligned}&\min_{\kappa_0, \kappa_w, \xi}\left\{\frac{1}{n}\sum_{i=1}^{n}\xi_i \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\kappa_w\|^2\right\} \text{ subject to} \\ &Y_i\left(\kappa_0 + W_i'\kappa_w\right) + \xi_i \geq q_i,\ \xi_i \geq 0 \text{ for all } i \in [n],\ \text{and } \kappa_c \geq 0. \qquad \text{(S.6)}\end{aligned}$$

Proposition S.2 *The solution of the coefficient vector κ_w to the MU-SVM is unique.*

Proof of Proposition S.2 . We prove the uniqueness by contradiction. Suppose $\hat{u} = (\hat{\kappa}_0, \hat{\kappa}_w', \hat{\xi}')'$ and $\check{u} = (\check{\kappa}_0, \check{\kappa}_w', \check{\xi}')'$ are two minimizers with $\hat{\kappa}_w' \neq \check{\kappa}_w$, both of which achieve the minimum of the objective function, say C^*. Given that $\hat{u}$ and $\check{u}$ are both feasible, it is clear that $u_\delta := \left(\kappa_{0,\delta}, \kappa_{w,\delta}', \xi_\delta'\right)' = \delta\hat{u} + (1-\delta)\check{u}$ for any $\delta \in (0,1)$ is also feasible. But,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\xi_{i,\delta}\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\kappa_{w,\delta}\|^2 \\ &= \delta\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + (1-\delta)\frac{1}{n}\sum_{i=1}^{n}\check{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\kappa_{w,\delta}\|^2 \\ &< \delta\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + (1-\delta)\frac{1}{n}\sum_{i=1}^{n}\check{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\delta\|\hat{\kappa}_w\|^2 + \frac{\mu}{2}(1-\delta)\|\check{\kappa}_w\|^2 \\ &= \delta\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\hat{\kappa}_w\|^2\right) + (1-\delta)\left(\frac{1}{n}\sum_{i=1}^{n}\check{\xi}_i\frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2}\|\check{\kappa}_w\|^2\right) \\ &= \delta(C^*) + (1-\delta)(C^*) = C^*,\end{aligned}$$

where we have used the strict convexity of $\|\cdot\|^2$. Hence, we have found a feasible u_δ that delivers a smaller objective function. This contradicts the assumption that $\hat{u}$ and $\check{u}$ are both solutions. Therefore, all solutions to the MU-SVM problem have the same $\hat{\kappa}_w$. ■

While Proposition S.2 establishes the uniqueness of $\hat{\kappa}_w$, it does not say anything about $\hat{\kappa}_0$. To characterize $\hat{\kappa}_0$, we define

$$I_+(\hat{\kappa}_0|\hat{\kappa}_w) = \left\{i : Y_i = +1 \text{ and } Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) \leq q_i\right\},$$
$$I_-(\hat{\kappa}_0|\hat{\kappa}_w) = \left\{i : Y_i = -1 \text{ and } Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) \leq q_i\right\}.$$

Given $\hat{\kappa}_0$ and $\hat{\kappa}_w$, the index sets $I_+(\hat{\kappa}_0|\hat{\kappa}_w)$ and $I_-(\hat{\kappa}_0|\hat{\kappa}_w)$ consist of the indices for the positive and negative (geometric) support vectors, respectively.

Proposition S.3 *If* $\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i) \neq \sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i)$*, then there exists an* i^* *such that* $Y_{i^*}\left(\hat{\kappa}_0 + W_{i^*}'\hat{\kappa}_w\right) = q_i$.

As a direct implication of Proposition S.3 , at the solution $(\hat{\kappa}_0, \hat{\kappa}_w)$, either there is a support vector on the margin boundary (i.e., the positive and negative hyperplanes) such that $Y_{i^*}\left(\hat{\kappa}_0 + W_{i^*}'\hat{\kappa}_w\right) = q_{i^*}$ for some i^*, or the total loss from the positive support vectors is equal to that from the negative support vectors. At least one of these two equations holds and can be used to identify the unique $\hat{\kappa}_0$ with the smallest absolute value.

Proof of Proposition S.3 . Without loss of generality, we assume that $\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i) > \sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i)$. Given $\hat{\kappa}_w$, the solution $\hat{\kappa}_0$ for κ_0 satisfies

$$\hat{\kappa}_0 \in \arg\min_{\kappa_0}\tilde{Q}_n(\kappa_0) = \frac{1}{n}\sum_{i=1}^{n}\left[q_i - Y_i\left(\kappa_0 + W_i'\hat{\kappa}_w\right)\right]_+\psi(Y_i,X_i).$$

We prove the stated result by contradiction. Suppose no observation satisfies $Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) = q_i$. Then, $I_+(\hat{\kappa}_0|\hat{\kappa}_w)$ and $I_-(\hat{\kappa}_0|\hat{\kappa}_w)$ reduce to

$$I_+(\hat{\kappa}_0|\hat{\kappa}_w) = \left\{i : Y_i = +1 \text{ and } Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) < q_i\right\},$$
$$I_-(\hat{\kappa}_0|\hat{\kappa}_w) = \left\{i : Y_i = -1 \text{ and } Y_i\left(\hat{\kappa}_0 + W_i'\hat{\kappa}_w\right) < q_i\right\}.$$

As a result, we can increase $\hat{\kappa}_0$ by an infinitesimal amount $\epsilon > 0$ without changing $I_+(\hat{\kappa}_0|\hat{\kappa}_w)$ or $I_-(\hat{\kappa}_0|\hat{\kappa}_w)$ so that

$$I_+(\hat{\kappa}_0 + \epsilon|\hat{\kappa}_w) = I_+(\hat{\kappa}_0|\hat{\kappa}_w),$$
$$I_-(\hat{\kappa}_0 + \epsilon|\hat{\kappa}_w) = I_-(\hat{\kappa}_0|\hat{\kappa}_w),$$

for a small enough $\epsilon > 0$. Now,

$$
\begin{aligned}
&\tilde{Q}_n\left(\hat{\kappa}_0+\epsilon\right)-\tilde{Q}_n\left(\hat{\kappa}_0\right) \\
&=\frac{1}{n}\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}\left[q_i-Y_i\left(\hat{\kappa}_0+\epsilon+W_i'\hat{\kappa}_w\right)\right]_+\psi(Y_i,X_i)-\frac{1}{n}\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}\left[q_i-Y_i\left(\hat{\kappa}_0+W_i'\hat{\kappa}_w\right)\right]_+\psi(Y_i,X_i) \\
&+\frac{1}{n}\sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}\left[q_i-Y_i\left(\hat{\kappa}_0+\epsilon+W_i'\hat{\kappa}_w\right)\right]_+\psi(Y_i,X_i)-\frac{1}{n}\sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}\left[q_i-Y_i\left(\hat{\kappa}_0+W_i'\hat{\kappa}_w\right)\right]_+\psi(Y_i,X_i) \\
&=\frac{1}{n}\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}-Y_i\epsilon\psi(Y_i,X_i)+\frac{1}{n}\sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}-Y_i\epsilon\psi(Y_i,X_i) \\
&=\epsilon\frac{1}{n}\left(\sum_{i\in I_-(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i)-\sum_{i\in I_+(\hat{\kappa}_0|\hat{\kappa}_w)}\psi(Y_i,X_i)\right)<0.
\end{aligned}
$$

This contradicts the optimality of $\hat{\kappa}_0$. ■

S.4 List of Methods Used in Simulations

Before listing the methods, a few clarifications are necessary. Since $c\left(x\right)$ is linear in the simulation study and is absorbed into the polynomial approximation of $P\left(x\right)$, we do not include $c(x)$ in constructing the decision rule for the polynomial-based MU method or the Poly-MU-SVM in the simulation study. Specifically, for the latter, we do not include $c(x)$ in the separation constraints (including it has virtually no effect on the simulation results). More generally, when $c\left(x\right)$ is highly nonlinear, unreported simulation results clearly demonstrate the advantage of including $c\left(x\right)$ as a special covariate in implementing the Poly-MU-SVM.

S.4.1 The ML method

- The specification:
$$m(x,\theta)=\Lambda\left(\mathcal{P}_j\left(x;\theta\right)\right).$$
where $\mathcal{P}_j$ is a polynomial of order j with constant term included.

- The estimator:
$$\hat{\theta}_{\text{MLE}}\in\arg\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i+1}{2}\log m(X_i,\theta)+\left(1-\frac{Y_i+1}{2}\right)\log\left[1-m\left(X_i,\theta\right)\right]\right\}.$$

- The estimated action rule:
$$\hat{a}_{\text{MLE}}\left(x\right)=\text{sign}\left\{\Lambda(\mathcal{P}_j(x;\hat{\theta}_{\text{MLE}}))-c(x)\right\}.$$

S.4.2 The MU method

- The proposed action rule:
$$a\left(x;\theta\right)=\text{sign}\left(\mathcal{P}_j\left(x;\theta\right)\right).$$

- The estimator:

$$\hat{\theta}_{\text{MU}} \in \arg\min_{\theta\in\times} \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, X_i) 1\left\{Y_i \neq a_{\text{MU}}\left(X_i, \theta\right)\right\}$$
$$\in \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} U(a_{\text{MU}}\left(X_i, \theta\right), Y_i, X_i).$$

- The estimated action rule:

$$\hat{a}_{\text{MU}}\left(x\right) = \text{sign}\{\mathcal{P}_j(x; \hat{\theta}_{\text{MU}})\}.$$

S.4.3 The MU-SMD method

- The proposed action rules:

$$a_j\left(x; \theta^{(j)}\right) = \text{sign}\{\mathcal{P}_j(x; \theta^{(j)})\},\ j = 1, 2, \ldots, J.$$

- The estimators:

$$\hat{\theta}^{(j)}_{\text{MU-SMD}} \in \arg\max_{\theta^{(j)}\in\Theta_j} \frac{1}{n}\sum_{i=1}^{n} U(a_j\left(X_i, \theta^{(j)}\right), Y_i, X_i)$$
$$= \arg\min_{\theta^{(j)}\in\times_j} \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, X_i) 1\left\{Y \neq a_j\left(X_i, \theta^{(j)}\right)\right\},$$

and

$$\hat{\theta}_{\text{MU-SMD}} = \hat{\theta}^{(\hat{j})}_{\text{MU-SMD}} \text{ for}$$
$$\hat{j} = \arg\max_{j=1,\ldots,J_{\max}} \left\{\frac{1}{n}\sum_{i=1}^{n} U(a_j(X_i, \hat{\theta}^{(j)}_{\text{MU-SMD}}), Y_i, X_i) - \text{SMD}_j\right\},$$

where SMD_j is the simulated maximal discrepancy for the model class $\left\{\mathcal{P}_j\left(x; \theta^{(j)}\right) : \theta^{(j)} \in \Theta_j\right\}$.

- The estimated action rule:

$$\hat{a}_{\text{MU-SMD}}\left(x\right) = a_{\hat{j}}(x, \hat{\theta}^{(\hat{j})}_{\text{MU-SMD}}) = \text{sign}(\mathcal{P}_{\hat{j}}(x; \hat{\theta}^{(\hat{j})}_{\text{MU-SMD}})).$$

S.4.4 The L_p-SVM for $p = 1, 2$

- The estimator:

$$(\hat{\kappa}_0, \hat{\kappa}_\phi, \hat{\xi}) = \arg\min_{(\kappa_0, \kappa_\phi, \xi)} \left\{\frac{1}{n}\sum_{i=1}^{n} \xi_i + \frac{\mu}{2}\left(\|\kappa_\phi\|_p^p + \|\kappa_c\|_p^p\right)\right\} \text{ subject to}$$
$$Y_i\left(\kappa_0 + \phi(X_i)'\kappa_\phi\right) + \xi_i \geq 1,\ \xi_i \geq 0 \text{ for all } i \in [n].$$

- The estimated action rule:

$$\hat{a}_{\text{Lp-SVM}}\left(x\right) = \text{sign}(\hat{\kappa}_0 + \phi\left(x\right)'\hat{\kappa}_\phi).$$

S.4.5 The CS-SVM

- The estimator:

$$(\hat{\kappa}_0, \hat{\kappa}_\phi, \hat{\xi}) = \arg \min_{(\kappa_0, \kappa_\phi, \xi)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i \frac{\bar{\psi}(Y_i)}{\bar{\psi}} + \frac{\mu}{2} \left(\|\kappa_\phi\|_p^p + \|\kappa_c\|_p^p \right) \right\} \text{ subject to}$$
$$Y_i \left(\kappa_0 + \phi(X_i)' \kappa_\phi \right) + \xi_i \geq 1, \ \xi_i \geq 0 \text{ for all } i \in [n],$$

  where

$$\bar{\psi}(1) = \frac{1}{n + n_{out}} \sum_{j=1}^{n+n_{out}} \psi(1, X_j), \ \bar{\psi}(-1) = \frac{1}{n + n_{out}} \sum_{j=1}^{n+n_{out}} \psi(-1, X_j).$$

- The estimated action rule:

$$\hat{a}_{\text{CS-SVM}}(x) = \text{sign}(\hat{\kappa}_0 + \phi(x)' \hat{\kappa}_\phi).$$

S.4.6 Poly-MU-SVM

- The estimator:

$$(\hat{\kappa}_0, \hat{\kappa}_\phi, \hat{\xi}) = \arg \min_{(\kappa_0, \kappa_\phi, \xi)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i \frac{\psi(Y_i, X_i)}{\bar{\psi}} + \frac{\mu}{2} \left(\|\kappa_\phi\|^2 + \|\kappa_c\|^2 \right) \right\} \text{ subject to}$$
$$Y_i \left(\kappa_0 + \phi(X_i)' \kappa_\phi \right) + \xi_i \geq q_i, \ \xi_i \geq 0 \text{ for all } i \in [n],$$

  where

$$q^+ = \frac{1 - \rho}{\min(\rho, 1 - \rho)}, \ q^- = \frac{\rho}{\min(\rho, 1 - \rho)},$$

  and

$$q_i = q^+ \cdot 1\{Y_i = +1\} + q^- \cdot 1\{Y_i = -1\}.$$

- The estimated action rule:

$$\hat{a}_{\text{Poly-MU-SVM}}(x) = \text{sign}\left(\hat{\kappa}_0 + \phi(x)' \hat{\kappa}_\phi \right).$$

S.4.7 RBF-MU-SVM

- The specification:

$$m(x, \theta) = \kappa_0 + \phi(x)' \kappa_\phi,$$

  where $\phi(x) = (\sqrt{\alpha_1^*} \phi_1^*(x), \ldots \sqrt{\alpha_j^*} \phi_j^*(x), \ldots)'$ and $\{(\alpha_j^*, \phi_j^*(\cdot))\}_{j=1}^{\infty}$ are the eigenvalues and eigenfunctions of the RBF kernel $\mathcal{K}(x, \tilde{x}) = \exp(-\tau \|x - \tilde{x}\|^2)$.

- The estimator:

$$\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\lambda}_c \right) = \arg \max_{\lambda_1, \ldots, \lambda_n, \lambda_c} L_D(\lambda, \lambda_c) \text{ subject to}$$
$$\sum_{i=1}^{n} \lambda_i Y_i = 0, \ \lambda_c \geq 0, \text{ and}$$
$$0 \leq \lambda_i \leq \frac{1}{n} \frac{\psi(Y_i, X_i)}{\bar{\psi}} \text{ for all } i \in [n], \tag{S.7}$$

where

$$L_D(\lambda, \lambda_c) = \sum_{i=1}^{n} q_i \lambda_i - \frac{1}{2\mu} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \mathcal{K}_c(X_i, X_j) + 2\lambda_c \sum_{i=1}^{n} \lambda_i Y_i c_-(X_i) + \lambda_c^2 \right]$$

and $\mathcal{K}_c(X_i, X_j) = \mathcal{K}(X_i, X_j) + c_-(X_i) c_-(X_j)$.

- The estimated action rule:

$$\hat{a}_{\text{RBF-MU-SVM}}(x) = \text{sign}\left(\frac{1}{\mu} \left[\sum_{i=1}^{n} \hat{\lambda}_i Y_i \mathcal{K}_c(x, X_i) + \hat{\lambda}_c c_-(x) \right] + \hat{\kappa}_0 \right),$$

where

$$\hat{\kappa}_0 = \frac{1}{\left| S_+^{\psi} \right|} \sum_{i \in S_+^{\psi}} \left(Y_i - \frac{1}{\mu} \left[\sum_{j=1}^{n} \hat{\lambda}_j Y_j \mathcal{K}_c(X_i, X_j) + \hat{\lambda}_c c_-(X_i) \right] \right)$$

for $S_+^{\psi} = \{ i : 0 < \hat{\lambda}_i < \psi(Y_i, X_i) / \left(n \bar{\psi} \right) \}$.